This paper was presented at the TECHNOMETRICS Session of the 25th Annual Fall Technical Conference of the Chemical Division of the American Society for Quality Control and the Section on Physical and Engineering Sciences of the American Statistical Association in Gatlinburg, Tennessee, October 29–30, 1981.

A Comparative Study of Tests for Homogeneity of Variances, with Applications to the Outer Continental Shelf Bidding Data

W. J. Conover

College of Business Administration
Texas Tech University
Lubbock, TX 79409

Mark E. Johnson and Myrle M. Johnson

Statistics Group, S-1
Los Alamos National Laboratory
Los Alamos, NM 87545

Many of the existing parametric and nonparametric tests for homogeneity of variances, and some variations of these tests, are examined in this paper. Comparisons are made under the null hypothesis (for robustness) and under the alternative (for power). Monte Carlo simulations of various symmetric and asymmetric distributions, for various sample sizes, reveal a few tests that are robust and have good power. These tests are further compared using data from outer continental shelf bidding on oil and gas leases.

KEY WORDS: Test for homogeneity of variances; Bartlett's test; Robustness; Power; Nonparametric tests; Monte Carlo.

1. INTRODUCTION

Tests for homogeneity of variances are often of interest as a preliminary to other analyses such as analysis of variance or a pooling of data from different sources to yield an improved estimated variance. For example, in the data base described in Section 4, if the variance of the logs of the bids on each offshore lease is homogeneous within a sale, then the scale parameter of the lognormal distribution can be estimated using all the bids in the sale. In quality control work, tests for homogeneity of variances are often a useful endpoint in an analysis.

The classical approach to hypothesis testing usually begins with the likelihood ratio test under the assumption of normal distributions. However, the distribution of the statistic in the likelihood ratio test for equality of variances in normal populations depends on the kurtosis of the distribution (Box 1953), which helps to explain why that test is so sensitive to departures from normality. This nonrobust (sometimes called "puny") property of the likelihood ratio test has prompted the invention of many alternative tests for variances. Some of these are modifications of the likelihood ratio test. Others are adaptations of the F test to test variances rather than means. Many are based on nonparametric methods, although their modification for the case in which the means are unknown often makes these tests distributionally dependent.

Among the many possible tests for equality of variances, one would hope that at least one is robust to variations in the underlying distribution and yet sensitive to departures from the equal variance hypothesis. However, recent comparative studies are not reassuring in this regard. For example, Gartside (1972) studied eight tests and concluded that the only robust procedure was a *log-anova* test that not only has poor power, but also depends on the unpleasant process of dividing each sample *at random* into smaller subsamples. Layard (1973) reached a similar conclusion regarding the log-anova test, but indicated that two other tests in his study of four tests, Miller's jackknife procedure and Scheffé's chi squared test, did not suffer greatly from lack of robustness and had considerably more power, at least when sample sizes were equal. These tests are included in our study as Mill and Sch2. Layard indicated a reluctance to use these tests when sample sizes are less than 10, and yet this is the case of interest to us, as we explain later. The jackknife pro-

cedure appeared to be the best of the six procedures investigated by Hall (1972) in an extensive simulation study, while Keselman, Games, and Clinch (1979) conclude that the jackknife procedure (Mill) has unstable error rates (Type I error) when the sample sizes are unequal. They conclude from their study of 10 tests that "the current tests for variance heterogeneity are either sensitive to nonnormality or, if robust, lacking in power. Therefore these tests cannot be recommended for the purpose of testing the validity of the ANOVA homogeneity assumption." The four tests studied by Levy (1978) all "were grossly affected by violations of the underlying assumption of normality."

The potential user of a test for equality of variances is thus presented with a confusing array of information concerning which test to use. As a result, many users default to Bartlett's (1937) modification of the likelihood ratio test, a modification that is well known to be nonrobust and that none of the comparative studies recommends except when the populations are known to be normal. The purpose of our study is to provide a list of tests that have a stable Type I error rate when the normality assumption may not be true, when sample sizes may be small and/or unequal, and when distributions may be skewed and heavy-tailed. The tests that show the desired robustness are compared on the basis of power. Further, we hope that our method of comparing tests may be useful in future studies for evaluating additional tests of variance.

The tests examined in this study are described briefly in Section 2. Fifty-six tests for equality of variances are compared, most of which are variations of the most popular and most useful parametric and nonparametric tests available for testing the equality of k variances ($k \geq 2$) in the presence of unknown means. Some tests not studied in detail are also mentioned in Section 2, along with the reason for their exclusion. This coverage is by far the most extensive that we are aware of and should provide valuable comparative information regarding tests for variances.

The simulation study is described in Section 3. Each test statistic is computed 1,000 times in each of 91 situations, representing various distributions, sample sizes, means, and variances. Nineteen of these sample situations have equal variances and are therefore studies of the Type I error rate, while the remaining 72 situations represent studies of the power.

The basic motivation for this study is described in Section 4. The lease production, and revenue (LPR) data base includes, among other data, the actual amount of each sealed bid submitted by oil and gas companies on individual tracts offered by the federal government in all of the sales of offshore oil and gas leases in the United States since 1954. The results of several tests for variances applied to those sales are described. A final section presents the summary and conclusions of this study.

2. A SURVEY OF k-SAMPLE TESTS FOR EQUALITY OF VARIANCES

For $i = 1, \ldots, k$, let $\{X_{ij}\}$ be random samples of size n_i from populations with means μ_i and variances σ_i^2. To test the hypothesis of equal variances, one additional assumption is necessary (Moses 1963). One possible assumption is that the X_{ij}'s are normally distributed. This leads to a large number of tests, some with exact tables available and some with only asymptotic approximations available, for the distributions of the test statistics. Another possible assumption is that the X_{ij}'s are identically distributed when the null hypothesis is true. This assumption enables various nonparametric tests to be formulated. In practice, neither assumption is entirely true, so that all of these tests for variances are only approximate. It is appropriate to examine all of the available tests for their robustness to violations of the assumptions. In this section we present a (nearly) chronological listing of tests for equal variances and a summary of these tests in Tables 1 through 4. Most of the tests in Tables 1 through 3 are based on some modification of the likelihood ratio test statistic derived under the assumption of normality. Tests that are essentially modifications of the likelihood ratio test or that otherwise rely on the assumption of normality are given in Table 1. Modifications to those tests, employing an estimate of the kurtosis, appear in Table 2. They are asymptotically distribution free for all parent populations, with only minor restrictions. Tests based on a modification of the F test for means are given in Table 3, along with the jackknife test, which does not seem to fit anywhere else. Finally, Table 4 presents modifications of nonparametric tests. The modification consists of using the sample mean or sample median instead of the population mean when computing the test statistic. Only nonparametric tests in the class of linear rank tests are included here, because this class of tests includes all locally most powerful rank tests (Hajek and Sidak 1967). Therefore, in Table 4, only the scores, $a_{n,i}$, for these tests are presented. From these scores, chi squared tests may be formulated based on the statistic

$$X^2 = \sum_{i=1}^{k} n_i(\bar{A}_i - \bar{a})^2/V^2, \qquad (2.1)$$

where $\bar{A}_i$ = mean score in the ith sample, $\bar{a}$ = overall mean score = $1/N \sum_{i=1}^{N} a_{N,i}$, and $V^2 = (1/N - 1) \sum_{i=1}^{N} (a_{N,i} - \bar{a})^2$, which is compared with quantiles from a chi squared distribution with $k - 1$ degrees of freedom. Alternatively, the statistic

Table 1. *Tests That Are Classically Based on an Estimate of Sampling Fluctuation Assuming Normality*

Abbreviation of Test	Test Statistic and Distribution
N-P	$\chi^2_{k-1} = T_1 = N \ln(\frac{N-k}{N} s^2) - \sum_{i=1}^{k} n_i \ln(\frac{n_i-1}{n_i} s_i^2)$
Bar	$\chi^2_{k-1} = \frac{T_2}{C}$ where $T_2 = (N-k)\ln s^2 - \sum_{i=1}^{k} (n_i-1)\ln s_i^2$ and $C = 1 + \frac{1}{3(k-1)}[\sum_i \frac{1}{n_i-1} - \frac{1}{N-k}]$
Coch	$\frac{\max s_i^2}{\sum_i s_i^2}$ See Pearson and Hartley (1970), p. 203 for special tables.
B-K	$\frac{\ln(\max s_i^2) - \ln(\min s_i^2)}{(n/2)^{\frac{1}{2}}}$ See Pearson and Hartley (1970), p. 177 for special tables. (n=average sample size)
Hart	$\frac{\max s_i^2}{\min s_i^2}$ See Pearson and Hartley (1970), p. 202 for special tables.
Cad	$\frac{\max r_i}{\min r_i}$ See Pearson and Hartley (1970), p. 264 for special tables.
Bar3	$F_{k-1,w} = \frac{w T_2}{(k-1)(b-T_2)}$ where $w = (k+1)/(C-1)^2$ and $b = \frac{w}{C+2/w}$ (See Bar for C and T_2)
Sam	$\chi^2_{k-1} = \sum_{i=1}^{k} \frac{(m_i-m)^2}{a_i^2}$ where $m_i = (1 - \frac{2}{9(n_i-1)}) s_i^{-2/3}$, $a_i^2 = 2/[9(n_i-1)s_i^{4/3}]$, and $m = \frac{\sum_i (m_i/a_i^2)}{\sum_j (1/a_j^2)}$
Bar:range	$[(N-k)\ln(\frac{1}{N-k}\sum_i (n_i-1)(\frac{r_i}{d_i})^2) - \sum_i (n_i-1)\ln(\frac{r_i}{d_i})^2]/C$ (See Bar for C) See Pearson and Hartley (1970), p. 201 for special tables.
Leh1	$\chi^2_{k-1} = T_3/2$ where $T_3 = \sum_i (n_i-1)(P_i - \frac{1}{N-k}\sum_{j=1}^{k} (n_j-1)P_j)^2$ and $P_j = \ln s_j^2$
Leh2	$\chi^2_{k-1} = (N-k)T_3/(2N-4k)$ (See Leh1 for T_3)

$$F = \frac{X^2/(k-1)}{(N-1-X^2)/(N-k)} \tag{2.2}$$

may be compared with quantiles from the F distribution with $k-1, N-k$ degrees of freedom.

In the following descriptions of the tests, we let $\bar{X}_i$, $\tilde{X}_i$, and r_i denote the ith sample mean, median, and range, respectively, while $\bar{\bar{X}}$ denotes the overall mean. The ith sample variance, with divisor $n_i - 1$, is s_i. In addition,

$$N = \sum n_i, \quad s^2 = \sum (n_i - 1)s_i/(N-k),$$

and

$$F(X_{ij}) = \frac{\sum_i n_i(\bar{X}_i - \bar{\bar{X}})^2/(k-1)}{\sum_i \sum_j (X_{ij} - \bar{X}_i)^2/(N-k)} \tag{2.3}$$

is the usual one-way analysis of variance test statistic.

In tests for equal variances, F is computed on some transformation of the X_{ij}'s rather than on the X_{ij}'s themselves.

Comments on the various tests are now presented. The notation *med* refers to the replacement of $\bar{X}_i$ with $\tilde{X}_i$ in the test statistic in an attempt to improve the robustness of the test.

N-P. The test proposed by Neyman and Pearson (1931) is the likelihood ratio test under normality. We also examine the modification N-P:med.

Bar. Bartlett (1937) modified N-P to "correct for bias." The resulting test is probably the most common used for equality of variances. It is well known to be sensitive to departures from normality. Recent papers by Glaser (1976), Chao and Glaser (1978), and Dyer and Keating (1980) give methods for finding the exact distribution of the test statistic. We also examine Bar:med.

Coch. The test introduced by Cochran (1941) was considerably easier to compute than the tests up to that time. With today's computers the difference in computation time is slight, however. We also look at Coch:med.

B-K. Another attempt to simplify calculations resulted in this test by Bartlett and Kendall (1946), which relies on the fact that $\ln s^2$ is approximately normal and uses tables for the normalized range in normal samples. We do not examine this test because of its equivalence to the following test.

Hart. Four years after B-K this test by Hartley (1950) was presented. Well known as the "F-max" test, it is merely an exponential transformation of B-K. An advantage of this test is the exact tables available for equal sample sizes (David 1952). We also examine Hart:med.

Table 2. *Tests That Attempt To Estimate Kurtosis*

Abbreviation of Test	Test Statistic and Distribution
Bar1	$\chi^2_{k-1} = \frac{T_2}{C(1+\hat{\gamma}/2)}$ where $\hat{\gamma} = \frac{1}{N}\sum_i n_i^2 \frac{\sum_{j=1}^{n_i} (X_{ij}-\bar{X}_i)^4}{[(n_i-1)s_i^2]^2} - 3$ (See Bar for T_2 and C)
Bar2	$\chi^2_{k-1} = \frac{T_2}{C(1+\tilde{\gamma}/2)}$ where $\tilde{\gamma} = \frac{N\sum\sum (X_{ij}-\bar{X}_i)^4}{[\sum_i (n_i-1)s_i^2]^2} - 3$
Sch1	$\chi^2_{k-1} = \frac{T_3}{2+(1-\frac{k}{N})\hat{\gamma}}$ (See Leh1 for T_3, Bar1 for $\hat{\gamma}$)
Sch2	$\chi^2_{k-1} = \frac{T_3}{2+(1-\frac{k}{N})\tilde{\gamma}}$ (See Leh1 for T_3, Bar2 for $\tilde{\gamma}$)

Table 3. *Tests Based on a Modification of the F Test for Means (see equation (2.3) for $F(\cdot)$)*

Lev1	$F_{k-1,\,N-k} = F(\lvert X_{ij}-\bar{X}_i\rvert)$
Lev2	$F_{k-1,\,N-k} = F((X_{ij}-\bar{X}_i)^2)$
Lev3	$F_{k-1,\,N-k} = F(\ln(X_{ij}-\bar{X}_i)^2)$
Lev4	$F_{k-1,\,N-k} = F(\lvert X_{ij}-\bar{X}_i\rvert^{1/2})$
Mill	$F_{k-1,\,N-k} = F(U_{ij})$ where $U_{ij} = n_i \ln s_i^2 - (n_i-1)\ln s_{ij}^2$ and $s_{ij}^2 = \frac{1}{n_i-2}[(n_i-1)s_i^2 - n_i(X_{ij}-\bar{X}_i)^2/(n_i-1)]$

Cad. A desire for simplification led to replacing the variance in Hart with the sample range in a paper by Cadwell (1953). Exact tables for equal sample sizes are given by Harter (1963) for $k = 2$ and Leslie and Brown (1966) for $k \leq 12$. We do not examine this test because we feel that the computational advantages are no longer real with present-day software.

Bar1. Box (1953) showed that the asymptotic distribution of Bar was dependent on the common kurtosis of the sampled distributions and that by dividing Bar by $(1 + \gamma/2)$, where $\gamma = E(X_{ij} - \mu_i)^4/\sigma_i^4 - 3$, the test would be asymptotically distribution free, provided the assumption of common kurtosis was met. Our form for this modification of Bar involves estimating γ with the sample moments, a suggestion that Layard (1973) attributes to Scheffé (1959). We also examine Barl:med. Bar2 and Bar2:med result from a different estimator for γ as given by Layard.

Box. An interesting approach to obtaining a more robust test for variance involves using the one-way layout F statistic, which is known to be quite robust. A concept suggested by Bartlett and Kendall (1946) was developed by Box (1953) into a test known as the log-anova test. For a preselected, arbitrary integer $m \geq 2$, each sample is divided into subsamples of size m in some random manner. (See Martin and Games 1975, 1977 and Martin 1976 for suggestions on the size of m.) Remaining observations either are not used or are included in the final subsample. The sample variance s_{ij} is computed for each subsample, $i = 1, \ldots, k$, $j = 1, \ldots, [n_i/m] = J_i$. A log transformation $Y_{ij} = \ln s_{ij}$ then makes the variables more nearly normal, and $F(Y_{ij})$ is used as a test statistic. Subsequent studies by Gartside (1972), Layard (1973), and Levy (1975) confirmed the robustness of this method, but also revealed a lack of power as compared with other tests that have the same robustness.

A modification that leads to a more nearly normal sample is attributed to Bargmann by Gartside (1972). It uses $W_{ij} = w_i(\ln s_{ij} + c_i)$, where w_i and c_i are normalizing constants. However, the random method of subdividing samples and the possibility of not using all of the observations make these procedures unattractive to the practitioner. For this reason we do not include these tests in our study. A Monte Carlo comparison of these methods with the jackknife methods (see Mill) is presented by Martin and Games (1977).

Mood. The first nonparametric test for the variance problem was presented by Mood (1954). It, like all of the nonparametric tests, assumes identical distributions under the null hypothesis. In particular, this requires equal means, or a known transformation to achieve equal means, which is often not met in applications. Therefore, we adapt the Mood test and all of the nonparametric tests as follows. Instead of letting R_{ij} be the rank of X_{ij} when the means are equal or of $(X_{ij} - \mu_i)$ when the means are unequal but known, we let R_{ij} be the rank of $(X_{ij} - \bar{X}_i)$. Each X_{ij} is then replaced by the score a_N, R_{ij} based on this rank. The result is a test that is not nonparametric but may be as robust and powerful as some of its parametric competitors. The use of $\tilde{X}_i$ instead of $\bar{X}_i$ results in Mood:med, which we also examine. The chi squared approximation and the F approximation for each test lead to four variations, which are studied.

F-A-B. Although the Mood test is a quadratic function of R_{ij}, this test introduced by Freund and Ansari (1957) and further developed by Ansari and Bradley (1960) is a linear function of R_{ij}. Again, we let R_{ij} be the rank of $(X_{ij} - \bar{X}_i)$. We examine four variations of F-A-B (see Mood). The B-D test was introduced by Barton and David (1958) shortly after the F-A-B test and is similar to the F-A-B test in principle. Whereas the F-A-B scores are triangular in shape, the B-D scores follow a V shape with the large scores at the extremes and the small scores at the grand median. The result is a test with the same robustness and power as F-A-B. The same can be said for the S-T test,

Table 4. *Linear Rank Tests (scores may be used in equations (2.1), (2.2), or (2.3))*

Abbreviation of Test	Score Function $a_{N,i}$	Score $a_{N,R_{ij}}$ is a function of R_{ij}, where R_{ij} is the rank of:
Mood	$(i - \frac{N+1}{2})^2$	$(X_{ij}-\bar{X}_i)$
F-A-B	$\frac{N+1}{2} - \lvert i - \frac{N+1}{2}\rvert$ = 1, 2, 3, ..., 3, 2, 1	$(X_{ij}-\bar{X}_i)$
B-D	..., 3, 2, 1, 1, 2, 3, ...	$(X_{ij}-\bar{X}_i)$
S-T	1, 4, 5, ..., 6, 3, 2	$(X_{ij}-\bar{X}_i)$
Capon	$[E(Z_{N,i})]^2$ where $Z_{N,i}$ is the i th order statistic from a standard normal random sample of size N	$(X_{ij}-\bar{X}_i)$
Klotz	$[\Phi^{-1}(\frac{i}{N+1})]^2$ where $\Phi(x)$ is the standard normal distribution function	$(X_{ij}-\bar{X}_i)$
T-C	i	$\lvert X_{ij}-\bar{X}_i\rvert$
S-R	i^2	$\lvert X_{ij}-\bar{X}_i\rvert$
F-K	$\Phi^{-1}(\frac{1}{2} + \frac{i}{2(N+1)})$ (See Klotz for Φ)	$\lvert X_{ij}-\bar{X}_i\rvert$

introduced by Siegel and Tukey (1960) at about the same time. The only advantage of the S-T test is that tables for the Mann-Whitney test may be used; no special exact tables are required. We do not examine the B-D and S-T tests here because the results would be essentially the same as those found for F-A-B.

Sch1. The test statistic of this parametric procedure, attributed by Layard (1973) to Scheffé (1959), resembles in some respects the numerator of an F statistic computed on s_i, weighted by the degrees of freedom $n_i - 1$. The denominator is a function of the (assumed) common kurtosis, which in practice must be estimated. We use the sample kurtosis for γ, and also examine Schl:med. The variations Sch2 and Sch2:med arise when Layard's estimator for γ is used.

Leh1. Lehmann's (1959) suggested procedure is the same as Sch1, but with $\gamma = 0$ as in normal distributions. Ghosh (1972) shows that multiplication by $(N - k)/(N - 2k)$ gives a distribution closer to the chi square. We call this variation Leh2 and examine Leh1:med and Leh2:med also.

Lev1. Levene (1960) suggested using the one-way analysis of variance on the variables $Z_{ij} = |X_{ij} - \bar{X}_i|$ as a method of incorporating the robustness of that test into a test for variance. Further variations suggested by Levene involve $Z_{ij}^{1/2}$ (Lev2), ln Z_{ij} (Lev3), and Z_{ij}^2 (Lev4). We also consider Lev1:med, recommended by Brown and Forsythe (1974), and Lev4:med, but do not examine Lev3:med because ln $0 = -\infty$ occurs with odd sample sizes. We also do not consider use of the trimmed mean as Brown and Forsythe did, largely because their results indicated no advantages in using this variation.

Capon. Instead of using scores that are a quadratic function of the ranks as Mood had done, Capon (1961) suggested choosing scores that give optimum power in some sense. The result is this normal scores test, which is locally most powerful among rank tests against the normal-type alternatives, and asymptotically locally most powerful among *all* tests for this alternative.

Klotz. Shortly thereafter, Klotz (1962) introduced another normal scores test that used the more convenient normal quantiles. The result has possibly less power locally for small sample sizes, but has the same asymptotic properties as Capon. Because of its convenience, we examine the Klotz test, but not the very similar Capon. As in Mood, four variations of Klotz are considered.

Bar:range. Implicit in the literature since Patnaik's (1950) paper on the use of the range instead of the variance, but not explicitly mentioned until Gartside (1972), is this variation of Bar that uses the standardized range instead of the variance. The standardizing constants d_i are available from Pearson and Hartley (1970, p. 201). The number of degrees of freedom of the resulting chi squared test is adjusted from $(k - 1)$ to ν_i, where ν_i is available in the same reference. We do not examine this test because in general the range is less efficient than the sample variance.

Mill. The innovative jackknife procedure was applied to variance testing by Miller (1968). The jackknife procedure relies on partitioning the samples into subsamples of some predetermined size m. We take $m = 1$, to remove the chance variation involved with $m > 1$. We do not examine Mill:med.

Bar3. Dixon and Massey (1969) reported a variation of Bar that uses the F distribution. We also examine Bar3:med.

Sam. The cube root of s^2 is more nearly normal than s^2, which leads to this test by Samuiddin (1976). We also examined Sam:med.

F-K. Fligner and Killeen (1976) suggest ranking $|X_{ij}|$ and assigning increasing scores $a_{N,i} = i$, $a_{N,i} = i^2$, and $a_{N,i} = \Phi^{-1}(1/2 + (i/2(N + 1)))$ based on those ranks. We suggest using the ranks of $|X_{ij} - \bar{X}_i|$ and call the first test T-G after Talwar and Gentle (1977), who used a trimmed mean instead of $\bar{X}_i$. The second test, called the squared ranks test S-R, was discussed by Conover and Iman (1978), but has roots in earlier papers by Shorack (1965), Duran and Mielke (1968), and others. We denote the third test by F-K, even though we have taken liberties with their suggestion. We also examine, as with Mood, the four variations associated with each test. We do not examine Fligner and Killeen's suggestion of using the grand median in place of $\bar{X}_i$.

This list of tests does not include others such as one by Moses (1963) that relies on a random pairing within samples or one by Sukhatme (1958) that is closely related to some of the linear rank tests already included. Also, the Box-Anderson (1955) permutation test for two samples, which Shorack (1965) highly recommends, was found by Hall (1972) to have Type I error rates as high as 27 percent in the multisample case with normal populations at $\alpha = .05$, so it is not included in our study. However, the list is extensive enough for our purposes, namely, to obtain a listing of tests for variances that appear to have well-controlled Type I error rates, and to compare the power of the tests. This is accomplished in the next section.

3. THE RESULTS OF A SIMULATION STUDY

In the search for one or more tests that are robust as well as powerful, it became necessary to obtain pseudorandom samples from several distributions, using several sample sizes and various combinations of variances. The simulation study is described in this section. The results in terms of percent of times the null hypothesis was rejected are summarized in Tables 5 and 6.

For symmetric distributions we chose the uniform,

normal, and double exponential distributions. Uniform random numbers were simulated using CDC's uniform generator RANNUM, which is a multiplicative congruential generator type. The normal and double exponential variates were obtained from the respective inverse cumulative distribution functions. Four samples were drawn with respective sample sizes $(n_1, n_2, n_3, n_4) = (5, 5, 5, 5), (10, 10, 10, 10), (20, 20, 20, 20)$, and $(5, 5, 20, 20)$. The null hypothesis of equal variances (all equal to 1) was examined along with the four alternatives $(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2) = (1, 1, 1, 2), (1, 1, 1, 4), (1, 1, 1, 8)$, and $(1, 2, 4, 8)$. The mean was set equal to the standard deviation in each population under the alternative hypothesis. Zero means were used for H_0. Each of these 60 combinations of distribution type, sample size, and variances was repeated 1,000 times, so that the 56 test statistics mentioned in Section 2 were computed and compared with their 5 percent and 1 percent nominal critical values 60,000 times each. The observed frequency of rejection of the null hypothesis is reported in Table 5 for normal distributions and in Table 6 for double exponential distributions. The figures in parentheses in those tables represent the averages over the four variance combinations under the alternative hypothesis. The standard errors of all entries in Tables 5 and 6 are less than .016. The results for the uniform distribution are not reported here to save space. A table with the results for the uniform distribution is available from the authors on request.

The corresponding figures for the asymmetric case were obtained by squaring the random variables obtained in the symmetric case to obtain highly skewed and extremely leptokurtic distributions. To be more specific, we used $\sigma X_i^2 + \mu$ rather than $(\sigma X_i + \mu)^2$, where X_i represents the null distributed random variable, because the latter transformation does not allow as much control over means and variances as does the former. The three distributions (uniform)^2, (normal)^2, and (double exponential)^2, in combination with two sample sizes (10, 10, 10, 10) and (5, 5, 20, 20) and the five variance combinations (the null case and four alternatives, as before) gave a total of 30 combinations. For each combination, 1,000 repetitions were run for each of the 56 test statistics. The average frequency of rejection, averaged over the four variance combinations under the alternative, is presented in Tables 5 and 6 also.

The columns in Tables 5 and 6 represent the various sample sizes under symmetric and asymmetric distributions. For convenience, the nonsymmetric distributions are simply called asymmetric, although this is not meant to imply that the simulation results are attributable to the skewness of those distributions rather than to the extreme leptokurtic nature of those same asymmetric distributions. The seventh column in Table 5 represents a special study chosen to resemble the application situation described in Section 4. In brief, 13 samples in which the sample sizes were 2 (7 samples), 3 (2 samples), 4, 7 (2 samples), and 13, were drawn from standard normal distributions. This was repeated 1,000 times and 55 test statistics (Mill cannot be computed for $n_i = 2$) were computed each time. This case was investigated to see how the tests might behave under conditions typically encountered in oil-lease-bidding data.

There are many different ways of interpreting the results of Tables 5 and 6, just as there are many ways of defining what is a "good" test as opposed to a "bad" test. We will define a test to be robust if the maximum Type I error rate is less than .10 for a 5 percent test. The four tests that qualify under this criterion, and their maximum estimated test size in parentheses, are Bar2:med (.071), Lev1:med (.060), Lev2:med (.078), and F-K:med X^2 (.099). We include F-K:med F (.112) in this group of robust tests also, because in 18 of the 19 null cases examined the estimated test size was less than .084, which is well under control. Of these five tests the second, fourth, and fifth tests appear to have slightly more power than the other two. It is interesting to note that if the qualifications for robustness are loosened somewhat to max test size $\leq .15$, only one new test is included, Lev4:med (.145). Two additional tests have max test size $\leq .20$. These are Lev2 (.163) and Bar2 (.172). The increase in the Type I error rates of Lev2 and Bar2 over Lev2:med and Bar2:med is accompanied by only a 40 percent relative increase in power. The other test has less power. Therefore, a reasonable conclusion seems to be that the five tests with max test size $< .112$ qualify as robust tests for variances, with the tests Lev1:med, F-K:med X^2, and its sister test F-K:med F having slightly more power than the other two. Notice the resemblance among these three tests. The first uses an analysis of variance on $|X_{ij} - \tilde{X}_i|$, while the second and third convert $|X_{ij} - \tilde{X}_i|$ to ranks and then to normal type scores, where they are then subjected to either a chi squared test or an analysis of variance F test.

Similar conclusions were drawn using $\alpha = .01$. The only tests with a reasonably well-controlled test size are the same five tests that were selected using $\alpha = .05$. On the basis of demonstrated power at $\alpha = .01$, the same three tests mentioned for $\alpha = .05$ again appear to be the best. Therefore, the number of rejections for each test at $\alpha = .01$ is not reported.

If we consider only those five cases that have symmetric distributions, there are many additional tests that qualify as robust under the above definition. The five that show the most power, in order of decreasing power, are Bar2, Klotz:med F, Klotz:med X^2, Lev 4:med, and S-R:med F. However, the power of these five tests for symmetric distributions is about the same

Table 5. *For Normal and (Normal)^2 Distributions, Proportion of Times the Null Hypothesis of Equal Variance Was Rejected by the Various Tests, Under the Null Hypothesis (test size) and (in parentheses) Under the Alternative Hypothesis (power), at $\alpha = .05$*

	Normal Distribution: Symmetric					(Normal)^2: Asymmetric	
	n=(5,5,5,5)	(10,10,10,10)	(20,20,20,20)	(5,5,20,20)	Special Study	(10,10,10,10)	(5,5,20,20)
TABLE 1 TESTS							
N-P	.103 (.455)	.069 (.662)	.071 (.812)	.104 (.759)	.625	.674 (.826)	.663 (.865)
N-P:med	.115 (.454)	.081 (.662)	.077 (.814)	.098 (.750)	.639	.687 (.830)	.669 (.864)
Bar	.033 (.298)	.051 (.600)	.060 (.796)	.049 (.646)	.032	.614 (.788)	.567 (.797)
Bar:med	.034 (.299)	.052 (.596)	.064 (.798)	.049 (.630)	.034	.629 (.795)	.577 (.798)
Coch	.040 (.356)	.045 (.602)	.043 (.791)	.138 (.706)	.234	.480 (.663)	.576 (.768)
Coch:med	.041 (.353)	.042 (.604)	.045 (.792)	.151 (.684)	.223	.493 (.669)	.592 (.762)
Hart	.028 (.231)	.055 (.554)	.052 (.774)	.218 (.739)	.625	.604 (.772)	.720 (.882)
Hart:med	.029 (.235)	.058 (.552)	.056 (.776)	.213 (.720)	.627	.613 (.777)	.725 (.879)
Bar3	.034 (.303)	.051 (.600)	.060 (.796)	.049 (.648)	.040	.614 (.788)	.570 (.799)
Bar3:med	.037 (.306)	.053 (.597)	.064 (.798)	.049 (.633)	.046	.629 (.796)	.579 (.799)
Sam	.022 (.269)	.046 (.587)	.058 (.794)	.045 (.607)	.008	.606 (.781)	.538 (.764)
Sam:med	.019 (.274)	.048 (.582)	.064 (.795)	.054 (.594)	.006	.616 (.790)	.547 (.766)
Leh1	.094 (.377)	.082 (.618)	.069 (.792)	.102 (.731)	.498	.664 (.814)	.634 (.858)
Leh1:med	.104 (.381)	.085 (.615)	.078 (.794)	.099 (.722)	.511	.676 (.819)	.648 (.854)
Leh2	.179 (.514)	.108 (.665)	.079 (.806)	.119 (.761)	.745	.697 (.837)	.673 (.873)
Leh2:med	.198 (.515)	.106 (.665)	.087 (.807)	.112 (.750)	.748	.717 (.844)	.680 (.870)
TABLE 2 TESTS							
Bar1	.273 (.612)	.154 (.709)	.105 (.822)	.123 (.729)	.648	.487 (.689)	.301 (.545)
Bar1:med	.121 (.435)	.087 (.638)	.082 (.807)	.092 (.667)	.397	.365 (.549)	.182 (.414)
Bar2	.047 (.132)	.053 (.383)	.051 (.734)	.048 (.505)	.050	.143 (.249)	.083 (.206)
Bar2:med	.007 (.039)	.024 (.281)	.033 (.696)	.029 (.431)	.014	.043 (.100)	.021 (.090)
Sch1	.272 (.603)	.163 (.710)	.119 (.819)	.176 (.790)	.808	.558 (.742)	.421 (.706)
Sch1:med	.170 (.477)	.114 (.649)	.090 (.802)	.140 (.737)	.722	.443 (.630)	.321 (.605)
Sch2	.112 (.242)	.079 (.419)	.063 (.720)	.103 (.645)	.510	.247 (.380)	.206 (.447)
Sch2:med	.056 (.136)	.048 (.322)	.049 (.682)	.072 (.577)	.402	.137 (.228)	.122 (.312)
TABLE 3 TESTS							
Lev1	.083 (.303)	.064 (.543)	.058 (.768)	.060 (.583)	.263	.349 (.561)	.293 (.489)
Lev1:med	.002 (.065)	.025 (.437)	.039 (.732)	.032 (.521)	.057	.054 (.184)	.043 (.142)
Lev2	.057 (.235)	.047 (.489)	.048 (.774)	.055 (.456)	.163	.097 (.208)	.116 (.107)
Lev2:med	.011 (.080)	.015 (.388)	.033 (.749)	.035 (.383)	.048	.014 (.061)	.044 (.029)
Lev3	.069 (.192)	.062 (.337)	.057 (.554)	.069 (.461)	.403	.461 (.637)	.471 (.699)
Lev4	.091 (.283)	.069 (.493)	.060 (.716)	.070 (.571)	.372	.491 (.688)	.477 (.710)
Lev4:med	.000 (.004)	.037 (.383)	.034 (.659)	.049 (.552)	.020	.144 (.297)	.104 (.337)
Mill	.030 (.134)	.040 (.435)	.054 (.752)	.077 (.550)	-	.153 (.254)	.172 (.324)
TABLE 4 TESTS							
Mood X^2	.070 (.247)	.069 (.472)	.060 (.711)	.066 (.562)	.215	.752 (.862)	.684 (.827)
Mood F	.091 (.296)	.077 (.494)	.063 (.716)	.076 (.578)	.317	.768 (.874)	.702 (.837)
Mood:med X^2	.002 (.033)	.036 (.342)	.038 (.657)	.032 (.491)	.059	.410 (.577)	.370 (.623)
Mood:med F	.009 (.063)	.041 (.367)	.038 (.663)	.036 (.506)	.094	.433 (.595)	.381 (.636)
F-A-B X^2	.070 (.193)	.058 (.395)	.056 (.634)	.060 (.516)	.269	.728 (.838)	.638 (.803)
F-A-B F	.094 (.240)	.068 (.415)	.058 (.643)	.065 (.532)	.380	.741 (.852)	.648 (.811)
F-A-B:med X^2	.000 (.000)	.034 (.276)	.029 (.566)	.040 (.486)	.033	.395 (.550)	.340 (.643)
F-A-B:med F	.000 (.000)	.037 (.294)	.030 (.575)	.043 (.504)	.065	.418 (.572)	.357 (.660)
Klotz X^2	.057 (.265)	.053 (.526)	.058 (.772)	.062 (.538)	.152	.713 (.841)	.678 (.802)
Klotz F	.078 (.311)	.064 (.547)	.062 (.777)	.072 (.551)	.222	.741 (.855)	.704 (.815)
Klotz:med X^2	.011 (.078)	.031 (.407)	.034 (.734)	.033 (.472)	.050	.352 (.554)	.328 (.570)
Klotz:med F	.015 (.104)	.039 (.424)	.036 (.740)	.033 (.490)	.084	.387 (.574)	.348 (.588)
S-R X^2	.060 (.248)	.062 (.474)	.057 (.709)	.060 (.566)	.228	.613 (.770)	.589 (.789)
S-R F	.093 (.296)	.074 (.494)	.059 (.714)	.068 (.586)	.322	.630 (.783)	.611 (.802)
S-R:med X^2	.000 (.015)	.023 (.332)	.032 (.658)	.026 (.491)	.054	.171 (.322)	.105 (.347)
S-R:med F	.003 (.038)	.029 (.353)	.035 (.664)	.032 (.509)	.092	.183 (.340)	.119 (.364)
F-K X^2	.044 (.248)	.043 (.521)	.051 (.776)	.050 (.528)	.127	.422 (.623)	.361 (.576)
F-K F	.061 (.296)	.052 (.540)	.053 (.782)	.054 (.544)	.174	.442 (.646)	.383 (.596)
F-K:med X^2	.004 (.058)	.018 (.413)	.033 (.746)	.030 (.470)	.034	.066 (.218)	.052 (.197)
F-K:med F	.009 (.081)	.020 (.436)	.033 (.751)	.032 (.489)	.054	.084 (.235)	.057 (.211)
T-G X^2	.068 (.203)	.058 (.397)	.056 (.636)	.058 (.525)	.305	.610 (.753)	.608 (.809)
T-G F	.089 (.247)	.067 (.420)	.059 (.643)	.065 (.540)	.418	.621 (.770)	.623 (.818)
T-G:med X^2	.000 (.000)	.027 (.268)	.025 (.564)	.035 (.472)	.027	.256 (.390)	.172 (.444)
T-G:med F	.000 (.000)	.033 (.288)	.026 (.573)	.038 (.491)	.053	.272 (.413)	.189 (.458)

Table 6. *For Double Exponential and (Dbl. Exp.)^2 Distributions, Proportion of Times the Null Hypothesis of Equal Variances Was Rejected by the Various Tests, Under the Null Hypothesis (test size) and (in parentheses) Under the Alternative Hypothesis (power), at* $\alpha = .05$

	Dbl. Exp. Distribution: Symmetric			(Dbl. Exp.)^2: Asymmetric		
	n=(5,5,5,5)	(10,10,10,10)	(20,20,20,20)	(5,5,20,20)	(10,10,10,10)	(5,5,20,20)
TABLE 1 TESTS						
N-P	.316 (.553)	.339 (.713)	.316 (.836)	.333 (.801)	.876 (.912)	.883 (.933)
N-P:med	.322 (.556)	.340 (.713)	.317 (.835)	.333 (.795)	.881 (.914)	.886 (.934)
Bar	.157 (.395)	.273 (.661)	.288 (.821)	.233 (.710)	.856 (.892)	.832 (.897)
Bar:med	.164 (.397)	.275 (.659)	.292 (.822)	.240 (.697)	.860 (.891)	.830 (.898)
Coch	.154 (.386)	.232 (.592)	.214 (.762)	.324 (.721)	.723 (.773)	.798 (.856)
Coch:med	.164 (.381)	.236 (.593)	.212 (.764)	.340 (.712)	.724 (.777)	.799 (.855)
Hart	.134 (.345)	.248 (.624)	.264 (.806)	.460 (.815)	.845 (.888)	.908 (.950)
Hart:med	.139 (.348)	.252 (.629)	.267 (.806)	.457 (.807)	.849 (.888)	.906 (.950)
Bar3	.161 (.402)	.275 (.661)	.288 (.821)	.236 (.711)	.856 (.892)	.833 (.897)
Bar3:med	.170 (.401)	.275 (.659)	.293 (.822)	.243 (.699)	.862 (.892)	.834 (.898)
Sam	.135 (.364)	.261 (.653)	.284 (.819)	.213 (.670)	.853 (.888)	.808 (.884)
Sam:med	.145 (.365)	.265 (.650)	.285 (.820)	.231 (.663)	.853 (.887)	.805 (.885)
Lehl	.275 (.504)	.315 (.687)	.314 (.831)	.317 (.790)	.868 (.907)	.866 (.929)
Lehl:med	.278 (.507)	.313 (.689)	.315 (.828)	.314 (.779)	.874 (.910)	.872 (.930)
Leh2	.404 (.620)	.361 (.728)	.334 (.841)	.357 (.809)	.888 (.921)	.883 (.939)
Leh2:med	.401 (.627)	.366 (.727)	.341 (.840)	.356 (.803)	.889 (.924)	.886 (.939)
TABLE 2 TESTS						
Barl	.450 (.553)	.273 (.641)	.169 (.727)	.179 (.605)	.696 (.766)	.439 (.557)
Barl:med	.238 (.470)	.199 (.563)	.144 (.701)	.133 (.546)	.551 (.654)	.332 (.454)
Bar2	.047 (.129)	.054 (.232)	.050 (.492)	.046 (.289)	.172 (.230)	.100 (.153)
Bar2:med	.010 (.041)	.016 (.165)	.033 (.440)	.020 (.229)	.071 (.099)	.024 (.073)
Schl	.470 (.671)	.313 (.668)	.190 (.739)	.254 (.717)	.758 (.819)	.598 (.759)
Schl:med	.325 (.548)	.250 (.603)	.170 (.718)	.210 (.663)	.652 (.741)	.514 (.693)
Sch2	.167 (.298)	.101 (.322)	.079 (.511)	.116 (.486)	.361 (.453)	.284 (.468)
Sch2:med	.087 (.176)	.069 (.243)	.058 (.460)	.082 (.416)	.249 (.310)	.188 (.355)
TABLE 3 TESTS						
Levl	.097 (.268)	.077 (.415)	.068 (.645)	.087 (.396)	.473 (.579)	.384 (.420)
Levl:med	.008 (.051)	.033 (.291)	.039 (.591)	.035 (.325)	.048 (.092)	.060 (.057)
Lev2	.057 (.155)	.048 (.266)	.040 (.524)	.079 (.194)	.074 (.115)	.149 (.077)
Lev2:med	.010 (.051)	.024 (.184)	.027 (.473)	.048 (.143)	.012 (.024)	.078 (.027)
Lev3	.098 (.229)	.077 (.326)	.078 (.498)	.087 (.404)	.741 (.805)	.729 (.836)
Lev4	.121 (.290)	.093 (.419)	.082 (.630)	.092 (.458)	.715 (.803)	.688 (.797)
Lev4:med	.000 (.008)	.045 (.306)	.041 (.562)	.047 (.413)	.145 (.226)	.099 (.199)
Mill	.046 (.136)	.067 (.319)	.087 (.537)	.107 (.419)	.195 (.240)	.214 (.291)

as the power of the three tests mentioned previously for those same symmetric distributions. Therefore, the three tests, *Levl:med*, *F-K:med* X^2, and *F-K:med F*, again appear to be the best tests to use on the basis of robustness and power.

4. APPLICATION TO THE LPR DATA BASE

Since 1954 the United States government has periodically held sales in which offshore leases have been offered to the highest bidder for the production of oil and gas. The lease, production, and revenue (LPR) data base includes detailed information on the bids submitted, as well as the yearly production and revenue data on each lease. Our interest is in the bids submitted on the various leases within each sale. Often, the lognormal distribution is used to model these bids (Dougherty and Lohrenz 1976). If it is reasonable to assume that the variance of the log of the bids on each lease is constant within a sale, then the scale parameter of the lognormal distribution can be estimated using all the bids in the sale.

The bids in 40 sales were examined. These included all the sales held from October 13, 1954 to October 27, 1977, which is the date of the last sale recorded in the data base at the time of this study. We considered only leases within a sale receiving two or more bids on the lease. The 40 sales averaged about 50 leases per sale, with a range from 5 to 133. Although some of the leases have as many as 12 or 13 bids, small numbers of bids are the general rule, with about half of the leases examined having only two bids submitted on them.

For example, the sale held on July 21, 1970 was the

Table 6 (*Continued*)

	Dbl. Exp. Distribution: Symmetric				(Dbl. Exp.)^2: Asymmetric	
	n=(5,5,5,5)	(10,10,10,10)	(20,20,20,20)	(5,5,20,20)	(10,10,10,10)	(5,5,20,20)
TABLE 4 TESTS						
Mood X^2	.080 (.221)	.087 (.372)	.065 (.592)	.082 (.424)	.919 (.936)	.855 (.892)
Mood F	.121 (.275)	.095 (.388)	.069 (.598)	.090 (.440)	.928 (.945)	.863 (.899)
Mood:med X^2	.003 (.027)	.036 (.262)	.041 (.523)	.036 (.356)	.573 (.670)	.505 (.668)
Mood:med F	.009 (.048)	.041 (.275)	.045 (.533)	.039 (.373)	.597 (.687)	.528 (.682)
F-A-B X^2	.091 (.195)	.082 (.325)	.068 (.536)	.071 (.403)	.908 (.925)	.829 (.862)
F-A-B F	.112 (.240)	.094 (.349)	.077 (.546)	.082 (.420)	.913 (.935)	.834 (.869)
F-A-B:med X^2	.000 (.000)	.045 (.228)	.035 (.447)	.043 (.368)	.562 (.660)	.474 (.679)
F-A-B:med F	.000 (.000)	.053 (.245)	.036 (.454)	.048 (.385)	.575 (.680)	.494 (.694)
Klotz X^2	.072 (.223)	.077 (.388)	.070 (.629)	.079 (.390)	.907 (.923)	.838 (.865)
Klotz F	.105 (.273)	.082 (.410)	.075 (.637)	.085 (.408)	.923 (.934)	.849 (.877)
Klotz:med X^2	.012 (.061)	.039 (.286)	.037 (.575)	.045 (.330)	.516 (.615)	.483 (.595)
Klotz:med F	.016 (.081)	.044 (.303)	.039 (.584)	.050 (.345)	.537 (.641)	.512 (.616)
S-R X^2	.087 (.241)	.086 (.386)	.069 (.599)	.081 (.445)	.842 (.891)	.837 (.909)
S-R F	.115 (.289)	.097 (.408)	.071 (.607)	.087 (.460)	.851 (.902)	.846 (.915)
S-R:med X^2	.000 (.010)	.031 (.250)	.042 (.526)	.029 (.355)	.254 (.342)	.145 (.312)
S-R:med F	.003 (.029)	.034 (.269)	.042 (.536)	.032 (.371)	.262 (.365)	.153 (.326)
F-K X^2	.058 (.214)	.067 (.387)	.063 (.632)	.074 (.383)	.660 (.755)	.632 (.711)
F-K F	.086 (.263)	.076 (.405)	.067 (.639)	.077 (.401)	.677 (.768)	.651 (.729)
F-K:med X^2	.005 (.040)	.026 (.274)	.033 (.581)	.032 (.317)	.099 (.195)	.076 (.152)
F-K:med F	.011 (.063)	.030 (.293)	.036 (.588)	.037 (.331)	.112 (.210)	.080 (.160)
T-G X^2	.095 (.217)	.095 (.342)	.070 (.545)	.076 (.429)	.847 (.890)	.845 (.920)
T-G F	.122 (.264)	.099 (.364)	.072 (.554)	.082 (.446)	.863 (.899)	.858 (.924)
T-G:med X^2	.000 (.000)	.039 (.222)	.033 (.447)	.037 (.358)	.364 (.458)	.251 (.439)
T-G:med F	.000 (.000)	.047 (.237)	.034 (.457)	.043 (.376)	.382 (.480)	.266 (.451)

20th sale in chronological sequence. It had 13 leases that received two or more bids apiece. A special simulation study for this number of leases, with the same sample sizes, was reported in Table 5 and mentioned in Section 3. Some of the tests for variances rejected the null hypothesis over 70 percent of the time even though the normal distribution was used in the simulation and H_0 was true. It is useless to consider such tests for real data, since the results of such tests would be meaningless. Therefore, the results of only those tests that had well-controlled Type I error rates in the simulation study are examined in this section. This includes the five tests that had estimated test sizes less than .112 in all cases described in Section 3. For each of the five test statistics in each of the 40 sales, the P values were obtained by referring to the appropriate chi squared or F distribution. If H_0 is true these P values should be uniform on (0, 1), but if H_0 is false they should tend to be smaller. For each test, the 40 P values were summed and normalized by subtracting 20 and dividing by $\sqrt{40/12}$. The results appear in Table 7, column (2). Column (3) in Table 7 is simply the overall P value obtained by comparing the statistic in column (2) with the standard normal distribution.

For all five tests the overall P value is well above 5 percent, clearly indicating that the null hypothesis of equal variances should be accepted. In fact, for the two tests Bar2:med and Lev2:med, the overall P value is in the opposite tail of the distribution, suggesting that the asymptotic approximations used in those tests may be too conservative. This could also explain the well-controlled Type I error rate and the low power in the simulation study of Section 3 for those two tests. The three tests, *Lev1 :med*, *F-K :med* X^2, and *F-K:med F*, do not exhibit this weakness. They all have overall P values that do in fact resemble observations on a uniformly distributed random variable. Again, the same three tests show the same desirable properties.

It was mentioned previously that if H_0 is true, the p values should be uniform on (0, 1). A Kolmogorov goodness-of-fit test was used on the 40 P values to see how well they agreed with the uniform distribution. The test statistics for Lev1:med, F-K:med X^2, and

Table 7. *Summary of P Values for 5 Tests, 40 Applications Each*

(1) Test	(2) Standardized p-value Sum	(3) p-value of Col (2)
Bar2:med	6.109	1.000
Lev1:med	-0.530	.298
Lev2:med	2.998	.999
F-K:med X^2	0.766	.778
F-K:med F	1.034	.849

F-K:med F are .136, .112, and .132, respectively, all well below the $\alpha = .20$ critical value .165 for $n = 40$. Thus, these P values are consistent with a uniform distribution.

The correlation between pairs of these three tests is interesting to examine to see if the tests tend to agree in results. The sample correlation coefficient for these 40 P values is .846 between Lev1:med and F-K:med X^2, and .838 between Lev1:med and F-K:med F, which indicates a strong, but not perfect, linear association in both cases. Since the test statistics for F-K:med X^2 and F-K:med F are functionally related, the high correlation value of .997 is expected between those two.

5. SUMMARY AND CONCLUSIONS

Many of the tests for variances that receive widespread usage, such as Bar, Coch, and Hart, have uncontrolled risk of Type I errors when the populations are asymmetric and heavy-tailed. Even the more popular nonparametric tests, such as Mood, Klotz, and F-A-B, show unstable error rates when they are modified for the case in which the population means are unknown. Thus, it is important to find some tests for variances, when the population means are unknown, that show stable error rates and reasonable power.

After extensive simulation involving different distributions, sample sizes, means, and variances, three tests appear to be superior selections in terms of robustness and power. These are Lev1:med, F-K:med X^2, and F-K:med F, which are described in Section 2. These tests and two others that showed some good properties were applied to oil and gas lease bidding data to see if the logs of the bids exhibited homogeneity of variance from lease to lease within a sale. After combining test information over 40 different sales (40 applications of each multisample test for variance), the results were conclusive. All of the three selected tests indicated a good agreement with the null hypothesis. The other two tests appeared to be too conservative. Therefore, it seems reasonable to assume homogeneity of variance of the logs of the bids from lease to lease within a sale. Also, it seems reasonable to recommend Lev1:med, F-K:med X^2, and F-K:med F as robust and powerful tests for variances when the population means are unknown. Some more specific comments pertaining to the individual results are as follows.

1. Replacing the mean $\bar{X}$ by the median $\tilde{X}$ produced a dramatic decrease in the Type I error rate in some tests, but had almost no discernible effect on other tests. All five of the tests chosen as robust tests used the median rather than the mean. On the other hand, the tests of Table 1 gave essentially the same results with the median as with the mean. Use of the median affected all of the tests in Table 4 by bringing their Type I error rates closer to acceptable limits.

2. The X^2 and F approximations resulted in nearly identical tests when both approximations were tried. In all cases the Type I error rate and the power were slightly larger when the F approximation was used than when the X^2 approximation was used.

3. The kurtosis tests of Table 2 were the only ones that performed poorly with the normal distributions when the sample sizes were equal. Their performance improved with increasing sample sizes, however.

4. A striking result of this simulation study is the extremely poor performance of most of these tests when the distributions were asymmetric and heavy-tailed.

5. Some of the tests never rejected the null hypothesis when the sample sizes were (5, 5, 5, 5). These were the Talwar-Gentle test using the median and the Freund-Ansari-Bradley test with the median. The nominal critical values were larger than the maximum possible value of the test statistic in both cases. This peculiarity occurs with small, odd, sample sizes when the median is used. If the middle observation in each sample is deleted, the problem is eliminated. Additional simulations, not reported here, bear this out. However, the Type I error rate sometimes becomes inflated to an unsatisfactory level. This happened with the T-G:med and Lev1:med tests, but the F-K:med X^2 and F-K:med F tests were still under control in these additional studies. Therefore, we recommend that the median be deleted when $n_i \leq 19$ and odd with those two tests.

6. ACKNOWLEDGMENTS

The work performed in this paper was supported by the Applied Research and Analysis Section, Conservation Division, U. S. Geological Survey, Denver, CO under the auspices of John Lohrenz. The first author was serving as a Visiting Staff Member at the Los Alamos Scientific Laboratory. Professional assistance from Dr. Benjamin S. Duran is gratefully acknowledged.

[*Received November 1979. Revised March 1981.*]

REFERENCES

ANSARI, A. R., and BRADLEY, R. A. (1960), "Rank-Sum Tests for Dispersion," *Ann. Math. Statist.*, 31, 1174–1189.

BARTLETT, M. S. (1937), "Properties of Sufficiency and Statistical Tests," *Proc. Roy. Soc.*, Ser. A, 160, 268–282.

BARTLETT, M. S., and KENDALL, D. G. (1946), "The Statistical Analysis of Variance-Heterogeneity and the Logarithmic Transformation, *J. Roy. Statist. Soc.*, 8, 128–138.

BARTON, D. E., and DAVID, F. N. (1958), "A Test for Birth Order Effects," *Ann. Eugenics*, 22, 250–257.

BOX, G. E. P. (1953), "Non-normality and Tests on Variances," *Biometrika*, 40, 318–335.

BOX, G. E. P., and ANDERSON, S. L. (1955), "Permutation Theory in the Derivation of Robust Criteria and the Study of Departures From Assumption," *J. Roy. Statist. Soc.*, Ser. B, 17, 1–26.

BROWN, M. B., and FORSYTHE, A. B. (1974), "Robust Tests for the Equality of Variances," *J. Amer. Statist. Assoc.*, 69, 364–367.

CADWELL, J. H. (1953), "Approximating to the Distributions of Measures of Dispersion by a Power of X^2," *Biometrika*, 40, 336–346.

CAPON, J. (1961), "Asymptotic Efficiency of Certain Locally Most Powerful Ranks Tests," *Ann. Math. Statist.*, 32, 88–100.

CHAO, M., and GLASER, R. E. (1978), "The Exact Distribution of Bartlett's Test Statistic for Homogeneity of Variances With Unequal Sample Sizes," *J. Amer. Statist. Assoc.*, 73, 422–426.

COCHRAN, W. G. (1941), "The Distribution of the Largest of a Set of Estimated Variances as a Fraction of Their Total," *Ann. Eugenics, London*, 11, 47–52.

CONOVER, W. J., and IMAN, R. L. (1978), "Some Exact Tables for the Squared Ranks Test," *Comm. Statist. B—Simulation Comput.*, 7, 491–513.

DAVID, H. A. (1952), "Upper 5 and 1% Points of the Maximum F-Ratio," *Biometrika*, 39, 422–424.

DIXON, W. J., and MASSEY, F. J., JR. (1969), *Introduction to Statistical Analysis*, New York: McGraw-Hill.

DOUGHERTY, E. L., and LOHRENZ, J. (1976), "Statistical Analyses of Bids for Federal Offshore Leases," *J. Petroleum Techn.*, 28, 1377–1390.

DURAN, B. S., and MIELKE, P. W., JR. (1968), "Robustness of Sum of Squared Ranks Test," *J. Amer. Statist. Assoc.*, 63, 338–344.

DYER, D., and KEATING, P. (1980), "On the Determination of Critical Values for Bartlett's Test," *J. Amer. Statist. Assoc.*, 75, 313–319.

FLIGNER, M. A., and KILLEEN, T. J. (1976), "Distribution-Free Two-Sample Tests for Scale," *J. Amer. Statist. Assoc.*, 71, 210–213.

FREUND, J. E., and ANSARI, A. R. (1957), "Two-Way Rank Sum Test for Variances." Technical Report No. 34, Virginia Polytechnic Institute, Blacksburg, Virginia.

GARTSIDE, P. S. (1972), "A Study of Methods for Comparing Several Variances," *J. Amer. Statist. Assoc.*, 67, 342–346.

GHOSH, B. K. (1972), "On Lehmann's Test for Homogeneity of Variances," *J. Roy. Statist. Soc.*, Ser. B, 34, 221–234.

GLASER, R. E. (1976), "Exact Critical Values for Bartlett's Test for Homogeneity of Variances," *J. Amer. Statist. Assoc.*, 71, 488–490.

HAJEK, J., and SIDAK, Z. (1967), *Theory of Rank Tests*, New York: Academic Press.

HALL, I. J. (1972), "Some Comparisons of Tests for Equality of Variances," *J. Statist. Comput. Simulation*, 1, 183–194.

HARTER, H. L. (1963), "Percentage Points of the Ratio of Two Ranges and Power of the Associated Test," *Biometrika*, 50, 187–194.

HARTLEY, H. O. (1950), "The Maximum F-Ratio as a Short-Cut Test for Heterogeneity of Variance," *Biometrika*, 37, 308–312.

KESELMAN, H. J., GAMES, P. A., and CLINCH, J. J. (1979), "Tests for Homogeneity of Variance," *Comm. Statist. B—Simulation Comput.*, 8, 113–129.

KLOTZ, J. (1962), "Nonparametric Tests for Scale," *Ann. Math. Statist.*, 33, 498–512.

LAYARD, M. W. J. (1973), "Robust Large-Sample Tests for Homogeneity of Variance," *J. Amer. Statist. Assoc.*, 68, 195–198.

LEHMANN, E. L. (1959), *Testing Statistical Hypothesis*, New York: John Wiley.

LESLIE, R. T., and BROWN, B. M. (1966), " Use of Range in Testing Heterogeneity of Variance," *Biometrika*, 53, 221–227.

LEVENE, H. (1960), "Robust Tests for Equality of Variances," in *Contributions to Probability and Statistics*, ed. I. Olkin, Palo Alto, Calif.: Stanford University Press, 278–292.

LEVY, K. J. (1975), "An Empirical Comparison of Several Range Tests for Variances," *J. Amer. Statist. Assoc.*, 70, 180–183.

——— (1978), "An Empirical Study of the Cube-Root Test for Homogeneity of Variances With Respect to the Effects of Nonnormality and Power," *J. Statist. Comput. Simulation*, 7, 71–78.

MARTIN, C. G. (1976), "Comment on Levy's 'An Empirical Comparison of the Z-Variance and Box-Scheffé Tests for Homogeneity of Variance,' " *Psychometrika*, 41, 551–556.

MARTIN, C. G., and GAMES, P. A. (1975), "Selection of Subsample Sizes for the Bartlett and Kendall Test of Homogeneity of Variance," paper presented at the meeting of the American Educational Research Association, Washington, D.C., April 1975, (ERIC Document Reproduction Service No. ED 117 150).

——— (1977), "ANOVA Tests for Homogeneity of Variance: Nonnormality and Unequal Samples," *J. Educ. Statist.*, 2, 187–206.

MILLER, R. G. (1968), "Jackknifing Variances," *Ann. Math. Statist.*, 39, 567–582.

MOOD, A. M. (1954), "On the Asymptotic Efficiency of Certain Nonparametric Two-Sample Tests," *Ann. Math. Statist.*, 25, 514–533.

MOSES, L. E. (1963), "Rank Tests for Dispersion," *Ann. Math. Statist.*, 34, 973–983.

NEYMAN, J., and PEARSON, E. S. (1931), "On the Problem of k Samples," *Bull. Acad. Polon. Sci. et Lettres*, Sér. A, 460–481.

PATNAIK, P. B. (1950), "The Use of Mean Range as an Estimator of Variance in Statistical Tests," *Biometrika*, 37, 78–87.

PEARSON, E. S., and HARTLEY, H. O. (1970), *Biometrika Tables for Statisticians*, (Vol. 1, 3rd ed.), (reprinted with corrections, 1976), Cambridge University Press.

SAMUIDDIN, M. (1976), "Bayesian Test of Homogeneity of Variance," *J. Amer. Statist. Assoc.*, 71, 515–517.

SCHEFFÉ, H. (1959), *The Analysis of Variance*, New York: John Wiley.

SHORACK, G. R. (1965), "Nonparametric Tests and Estimation of Scale in the Two Sample Problem," Technical Report No. 10, Stanford University.

SIEGEL, S., and TUKEY, J. W. (1960), "A Nonparametic Sum of Ranks Procedure for Relative Spread in Unpaired Samples," *J. Amer. Statist. Assoc.*, 55, 429–444.

SUKHATME, B. V. (1958), "A Two Sample Distribution Free Test for Comparing Variances," *Biometrika*, 45, 544–548.

TALWAR, P. P., and GENTLE, J. E. (1977), "A Robust Test for the Homogeneity of Scales," *Comm. Statist. A—Theory Methods*, 6, 363–369.