BOOTSTRAP METHODS FOR TESTING HOMOGENEITY OF VARIANCES

Dennis D. Boos and Cavell Brownie

Institute of Statistics Mimeo Series No. ~~1969~~ 1696

December, 1986

Bootstrap Methods for Testing Homogeneity of Variances

Dennis D. Boos and Cavell Brownie
Department of Statistics
Box 8203
North Carolina State University
Raleigh, NC 27695-8203, USA

Summary

This paper describes the use of bootstrap and permutation methods for the problem of testing homogeneity of variances when means are not assumed equal or known. The methods are new in this context, and nontrivial, since the composite null hypothesis involves nuisance mean parameters. They allow the use of normal-theory test statistics such as $F = s_1^2/s_2^2$ without the normality assumption which is crucial for validity of critical values obtained from the F distribution. In a Monte Carlo study the new resampling methods are seen to compare favorably with older methods, except in the case of heavily skewed distributions.

1. Introduction

The main point of agreement in the numerous articles on testing homogeneity of variances is the non-robustness of test procedures derived from the likelihood ratio statistic assuming normal distributions. For example, Box and Andersen (1955), Miller (1968), Gartside (1972), Hall (1972), Layard (1973), Brown and Forsythe (1974), Keselman et al. (1979) and Conover et al. (1981) all demonstrate the sensitivity to departures from normality of Type I error rates for either the two-sample F-test or the k-sample analog due to Bartlett (1937). There is considerably less agreement among these authors however, as to which alternative procedures are best in terms of both robustness to variations in the underlying distribution and ability to detect departures from the equal variance hypothesis.

In motivating their study, Conover et al. (1981) noted the lack of consensus in the literature concerning alternative procedures, and commented that because of this confusion "many users default to Bartlett's" procedure. In fact examples of such confusion can still be found in the biological literature, as in Schiffelbein and Hills (1984, 1985), where interval estimates for variances were obtained using a jackknife procedure because of possible nonnormality, but variances were apparently compared using the normal-theory F-test.

To resolve the problem of conflicting recommendations concerning alternatives to the normal-theory tests, Conover et al. (1981) undertook a large Monte Carlo study. Their stated objectives were to provide a list of tests with stable Type I error rates under nonnormality and small and/or unequal sample sizes, and to compare power for tests found to be robust in this respect. They compared 56 tests under 91 situations (corresponding to combinations of distribution types, sample sizes, means and variances), and

thus provided considerable information concerning the relative performance of a large number of tests. Absent from the list of 56 tests, however, were resampling procedures other than the jackknife.

The objective of this study was therefore to investigate the use of bootstrap and permutation methods in testing equality of variances. Classical permutation procedures can be used for testing homogeneity of variances for the situation where means are assumed known or equal, but not for the more realistic situation where means are unknown and possibly unequal. As described in Section 3, this is because the usual permutation argument is destroyed when the null hypothesis is not one of identical populations. For this more interesting variance testing problem, we show in Section 3 how to construct permutation and bootstrap procedures which have approximately the correct level in small samples and exact level asymptotically.

Additional motivation for our study related to the use by Conover et al. (1981) of extremely skewed and leptokurtic distributions to evaluate robustness of test size. Our concern was that this could have led to the selection of unnecessarily conservative tests, with less than optimal power in realistic situations.

Requiring robustness of test size for distributions more skewed than the extreme value (obtained by log-transforming the exponential) does not seem important from a practical viewpoint. This is because most researchers know when they are dealing with highly skewed variables and will apply a suitable transformation. In fact, Conover et al. (1981) in their own real data example analyze log-transformed values. In many applications power should not be sacrificed for unnecessary robustness of test size. Examples include testing for variance homogeneity to justify pooling samples as in Zammuto (1984), and testing for treatment effects on variance, as when increased uniformity is a

desirable result. The latter situation arises in quality control of manufacturing processes (Nair, 1986), in the study of management practices in commercial poultry operations (Fairfull, Crober and Gowe, 1985), and in the study of educational methods (Games, Winkler, and Probert, 1972). We therefore concentrated on situations less extreme than the skewed distributions in Conover et al. (1981), to determine whether bootstrap or permutation methods could provide robustness of Type I error rates and increased power.

The new resampling procedures, described in Section 3, were compared by Monte Carlo simulation with three of the procedures studied by Conover et al. (1981). These latter procedures are Miller's jackknife (Miller 1968), the ANOVA F-test on absolute deviations from the median (Levene 1960, Brown and Forsythe, 1974) and a variation of Box and Andersen's (1955) M_1'. We define and discuss these three tests in Section 2.

Results of the Monte Carlo study are given in Section 4. They show that bootstrapping of $F = s_1^2/s_2^2$ and Bartlett's statistic results in a valid and powerful test procedure except for heavily skewed data. Moreover, the results show that bootstrapping works well with two other statistics and thus can be recommended as a general technique in the variance testing problem.

Real data examples are presented in Section 5 and concluding remarks are given in Section 6.

2. Some Previously Studied Tests for Equality of Variances

Let $X_{i1},\ldots,X_{in}$, $i = 1,\ldots,k$ represent k independent samples, where X_{ij}, $j = 1,\ldots,n_i$ are independent and identically distributed with cumulative distribution function $G_o((x-\mu_i)/\sigma_i)$ having finite fourth moment. The problem of interest is to test H_o: $\sigma_1^2 = \cdots = \sigma_k^2$.

Of the 56 tests for equality of variances studied by Conover et al. (1981), many are more properly called tests for equal spread or dispersion, including the three procedures said to be best on the basis of robustness and power. These three tests, denoted Levl:med, F-K:medX^2 and F-K:medF, are each based on absolute deviations from the median $|X_{ij} - \tilde{X}_i|$. Levl:med is the usual one-way ANOVA F-statistic for comparing k means, applied to the $|X_{ij} - \tilde{X}_i|$, and the other two are normal scores linear rank statistics (see Conover et al. 1981).

Of the procedures ruled out by Conover et al. as being too liberal with the skewed distributions, the best, denoted Bar2, is a modification of Bartlett's statistic that allows for non-zero coefficient of excess, $\gamma_2 = \beta_2 - 3$, of the cdf G_0. Conover et al. justified dismissal of Bar2 by the statement that when only symmetric distributions were considered, power for Bar2 was about the same as for Levl:med, F-K:medX^2 and F-K:medF. This is not completely borne out however, by empirical results in Conover et al. (1981), and elsewhere, that indicate there are situations where Levl:med and the F-K:med statistics lack power.

One such situation where Bar2 has better power is discussed by Conover et al. in their conclusions, and is evident in results for $(n_1,n_2,n_3,n_4) = (5,5,5,5)$ in their Tables 5 and 6. This is the case where n_i are small and odd. With respect to Levl:med, O'Brien (1978) explained that the extremely conservative performance for n_i small and odd was due to zero values of $|X_{ij} - \tilde{X}_i|$ inflating the estimate of within-group variance in the denominator of the F-statistic. Suggestions for deleting a random observation in each group (O'Brien, 1978) or the middle observation in each group (Conover et al. 1981) do not seem entirely satisfactory because they can result in a too-liberal test. This rather bizarre property is illustrated by

results for Levl:med for sample sizes (4,4,4,4) and (5,5,5,5) and 1000 Monte Carlo replicates using normal and exponential distributions. In the null case, for a test at nominal level .05, observed size at the normal was .074 and .003 for (4,4,4,4) and (5,5,5,5) respectively, and at the exponential was .100 and .015 respectively.

Another situation where statistics based on $|X_{ij} - \tilde{X}_i|$ can be expected to lack power is with data arising from short-tailed distributions. Empirical results show the jackknife statistic to have considerably better power than Levl:med at the uniform (O'Brien, 1978) and somewhat better power at the normal for k = 2 (Brown and Forsythe, 1974).

Thus, while it is evident that Levl:med is remarkably robust, unqualified recommendation of its use, as in Conover et al. (1981), does not seem appropriate, and ignores the possible advantages of moderately robust tests, such as Bar2 or the jackknife, for data that are not markedly nonnormal (see for example, Gad and Weil (1986) for lists of such variables). We therefore included Bar2 and Levl:med (which is more familiar and easier to interpret when the groups correspond to structured treatments than the linear rank statistics) in a Monte Carlo study aimed primarily at investigating resampling methods. Layard's (1973) generalization of Miller's (1968) jackknife was also included, partly to determine whether bootstrapping could be used to improve its robustness.

For clarity, these three statistics are defined below, with a few comments concerning their properties. Considerably more detail can be found in O'Brien (1978) concerning properties of Levl:med and the jackknife.

Levl:med. The test statistic is

$$F(Z_{ij}) = \frac{\sum_i n_i(\bar{Z}_i - \bar{Z})^2/(k-1)}{\sum\sum_{ij}(Z_{ij} - \bar{Z}_i)^2/(N-k)} \quad , \qquad (2.1)$$

where $Z_{ij} = |X_{ij} - \tilde{X}_i|$, $\tilde{X}_i$ is the median of sample i,

$$\bar{Z}_i = \frac{1}{n_i} \sum_j Z_{ij} \quad , \quad \bar{Z} = \frac{1}{N} \sum_i \sum_j Z_{ij} \quad \text{and} \quad N = \sum_i n_i \quad .$$

$F(Z_{ij})$ is compared to quantiles from the F distribution with $k - 1$ and $N - k$ degrees of freedom.

Levl:med is a modification of Levene's (1960) Z-test, obtained by taking absolute deviations about the median rather than the mean (see Miller, 1968; Brown and Forsythe, 1974). Empirical results in O'Brien (1978) concerning the expectation, variance and within-group correlations of the Z_{ij} suggest that ANOVA assumptions will not be seriously violated, hence, null performance of $F(Z_{ij})$ will generally be good, for $n_i > 8$ or so. However the conservative nature of the test for n_i small and odd should not be ignored.

Jackknife or Mill. The test statistic is $F(U_{ij})$ with $F(\cdot)$ as in (2.1), with

$$U_{ij} = n_i \log s_i^2 - (n_i - 1)\log s_{ij}^2 \quad ,$$

$$s_i^2 = \sum_j (X_{ij} - \bar{X}_i)^2/(n_i - 1) \qquad (2.2)$$

and

$$s_{ij}^2 = [(n_i - 1)s_i^2 - n_i(X_{ij} - \bar{X}_i)^2/(n_i - 1)]/(n_i - 2) \quad .$$

Critical values are obtained as for Levl:med.

$F(U_{ij})$ is Layard's (1973) k-sample generalization of Miller's (1968) two-sample jackknife procedure. O'Brien (1978) notes that the null behaviour

of this statistic is adversely affected, for unequal sample sizes, by the dependence on n_i of both the mean and variance of the pseudovalues U_{ij}. Empirical results in O'Brien (1978) showing positive within-group correlations between the U_{ij} for the exponential distribution explain the liberal nature of this test at the exponential.

Bar2. The test statistic is

$$T/[C(1 + \hat{\gamma}_2/2)]$$

where T/C is Bartlett's statistic,

$$T = (N - k)\log\{\sum_i (n_i - 1)s_i^2/(N - k)\} - \sum_i (n_i - 1)\log s_i^2, \qquad (2.3)$$

$$C = 1 + \frac{1}{3(k-1)}\left[\sum_i \frac{1}{n_i - 1} - \frac{1}{N-k}\right],$$

and

$$\hat{\gamma}_2 = \frac{\sum_i \sum_j (X_{ij} - \bar{X}_i)^4}{[\sum_i (X_{ij} - \bar{X}_i)^2]^2} - 3 .$$

Critical values are obtained from the chi-square distribution with k - 1 degrees of freedom.

Bar2 is similar to Box and Andersen's (1955) M_1', the difference being that Box and Andersen omit the correction factor C and estimate γ_2 using k-statistics. This is not at all clear in Conover et al. (1981), who refer to the Box-Andersen test as being a permutation test which they excluded from their study because of poor null performance reported by Hall (1972). We generated empirical reuslts which convinced us that results for M_1' in Table 3 of Hall (1972), including null levels cited by Conover et al. (1981), are incorrect. Our results also indicated that estimating γ_2 as in Bar2 using

Layard's (1973) pooled estimate of kurtosis, instead of using k-statistics, improved null performance.

3. Bootstrap and Permutation Procedures

Standard permutation methods can be used with any test statistic in the k-sample problem provided the null hypothesis is that all populations are identically distributed, i.e., H_I: $G_1(x) = G_2(x) = \cdots = G_k(x)$. This null hypothesis is appropriate in the location problem with equal variance, $G_i(x) = G_o((x - \mu_i)/\sigma)$, $i = 1,\ldots,k$. It also applies to the variance testing problem with equal location, $G_i(x) = G_o((x - \mu)/\sigma_i)$, $i = 1,\ldots,k$. H_I does not apply, however, to the more common situation of testing for variance differences without assuming equal location.

Before considering this latter null hypothesis, we briefly review results for testing H_I. The usual permutation approach is to obtain all $M = N!/(n_1!n_2!\cdots n_k!)$ sets of k samples of size $(n_1,\ldots,n_k)$ taken without replacement from the pooled data $S = \{X_{ij}, j = 1,\ldots,n_i, i = 1,\ldots,k\}$ and compute the test statistic T_i for each set. Let T_o be the statistic calculated from the original unpermuted data in S. If H_I is to be rejected for large values of T, then we would reject H_I at level α if T_o is greater than or equal to $(1 - \alpha)M$ of the T_i values. Simple arguments show that this procedure has exactly $P(\text{Type I error} | H_I) = \alpha$. See Bell and Sen (1984) for a recent survey of permutation procedures. Bootstrap procedures are similar but based on sampling from S *with replacement.* In general, bootstrap tests do not have exact level α but the level converges to α in large samples. See Efron (1979) for an introduction to bootstrap procedures.

The focus of this paper is on tests for equality of variance or dispersion in the presence of unknown and possibly unequal location. The distribution functions are $G_i(x) = G_o((x - \mu_i)/\sigma_i)$ where G_o is also unknown,

and the null hypothesis H_0: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$ is *not* equivalent to the null hypothesis H_I of identical populations. This lack of equality under H_0 destroys the usual permutation argument, and in fact both permutation and bootstrap procedures would have incorrect α levels even in large samples.

A straightforward remedy is to adjust the samples so that they have equal locations. Our approach is to resample from the aligned set

$$\bar{S} = \{X_{ij} - \bar{X}_i,\ j = 1,\ldots,n_i,\ i = 1,\ldots,k\} \quad . \tag{3.1}$$

In skewed samples it appears useful to replace $\bar{X}_i$ by $\tilde{X}_i$, the ith sample median. Neither permutation or bootstrap tests based on $\bar{S}$ have exact level α under H_0, but new theory in Boos, Janssen, and Veraverbeke (1986) shows that the levels converge to α in large samples. Since neither approach has exact level α, we prefer to emphasize the bootstrap approach because it is more intuitive and more closely associated with random sampling.

The basic idea behind bootstrap test procedures is to mimic the H_0 sampling situation as closely as possible. Bootstrapping from $\bar{S}$ of (3.1) can be viewed as drawing sets of iid samples $\{X_{i1}^*,\ldots,X_{in_i}^*,\ i = 1,\ldots,k\}$ from the "pseudo-population" whose distribution function is

$$G_N(x) = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} I(X_{ij} - \bar{X}_i \le x) \ .$$

Since $G_N(x) \approx G_0(x/\sigma)$ under H_0: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2$, the distribution of a test statistic T based on iid samples from G_N should be similar to that based on $G_0(x/\sigma)$. The latter is the null distribution of interest for variance type test statistics which are invariant to location shifts.

In theory one can evaluate T at all the $M = (N)^N$ equally likely sets of samples from G_N, $T_1^*,\ldots,T_M^*$, construct the exact distribution of the test statistic under G_N, and compute the bootstrap p value

$$p_N = (\#\text{ of } \{T_1^*,\ldots,T_M^*\} \geq T_0)/M \ .$$

In practice, $B << M$ sets of random samples are drawn from G_N, $T_1^*,\ldots,T_B^*$ are computed, and

$$\hat{p}_B = (\#\text{ of } \{T_1^*,\ldots,T_B^*\} \geq T_0)/B$$

is used as an estimator of p_N. Since this is binomial sampling, $\text{Var } \hat{p}_N = p_N(1-p_N)/B$, and the range $1000 \leq B \leq 10000$ works well in practice.

For the Monte Carlo study of Section 4 we have used the bootstrap approach to get critical values in the two-sample case for the usual F statistic $F = s_1^2/s_2^2$, for a similar ratio of robust dispersion estimators, and for Miller's jackknife t. In addition we have used the permutation approach from $\bar{S}$ with $F = s_1^2/s_2^2$. For the case of four independent samples, we have bootstrapped Bartlett's test statistic and the jackknife F statistic, both described in Section 2.

4. Monte Carlo Results

Test statistics described in Sections 2 and 3 were compared in the $k = 2$ and $k = 4$ sample situations for a variety of sample sizes, population distribution types, and null and alternative hypotheses. We shall discuss the $k = 2$ and $k = 4$ situations separately, after mentioning the following common features.

1. In every situation $N = 1000$ independent sets of Monte Carlo replications were generated. Thus, empirical test rejection rates follow the binomial ($N = 1000$, $p =$ probability of rejection) distribution.

2. P values were computed for each test statistic and rejection of H_o at $\alpha = .05$ means p value $\leq .05$.

3. Recall that B denotes the number of bootstrap replications within each of the N = 1000 Monte Carlo replications. It was too costly to let B = 1000 as suggested in Section 3. Therefore, we used a two-stage sequential procedure for bootstrap and permutation tests: a) start with B = 100; b) if $\hat{p}_B > .20$, stop; c) if $\hat{p}_B \leq .20$, take 400 more replications and use all B = 500 replications to compute $\hat{p}_B$.

4. Whenever $n_i < 10$ for at least one sample size, the smooth bootstrap (Efron, 1979, p. 7) was used for all bootstrap resampling. Here this smoothing is purely a computational device to avoid getting sample variances with value zero.

5. Since p values were obtained, a more comprehensive check on test statistic distribution under H_o was possible. Recall that under H_o a p value should have the uniform (0,1) distribution. For each statistic we counted the number of p values falling in the intervals (0,.01), (.01,.02),...,(.09,.10), (.10,1.0) and computed a chi-squared goodness-of-fit test of uniformity based on the 11 intervals. This approach conveys more information that just reporting empirical rejection rates for a level .05 test. (Box and Andersen, 1955, show histograms of p values).

6. In non-null situations it can be useless to compare empirical rejection rates ("observed power") if the null levels are much higher than the nominal levels. Therefore, when reporting estimates of power, we also include "adjusted power" estimates using the cell counts described above. These are obtained by simply adding the counts (or appropriate fraction thereof) for those cells for which

counts sum to α under H_0. For example, if the first 5 cells had counts (8, 13, 14, 11, 15) under H_0, and (164, 122, 98, 76, 62) under an alternative H_a, then the estimated true level under H_0 for nominal $\alpha = .05$ is .061, the observed power under H_a is .522, and the adjusted power is [164 + 122 + 98 + 76 + (62)(4/15)]/1000 = .477. These latter adjusted rates appear in parentheses in Tables 3 and 6. They attempt to estimate the power that would have been obtained if the correct critical values had been used.

Two-Sample Results

Tables 1 and 2 give null hypothesis rejection rates for nominal level $\alpha = .05$ and chi-squared goodness-of-fit tests on p values for sample sizes ($n_1 = 10$, $n_2 = 10$) and ($n_1 = 5$, $n_2 = 15$) for a variety of tests in the null case of $\sigma_1^2 = \sigma_2^2$. Table 1 is for three symmetric distributions, the uniform, the normal, and a t distribution with 5 degrees of freedom (t_5). Table 2 is for two skewed distributions, the extreme value with distribution function $F(x) = \exp(-\exp(-x))$, and the standard exponential.

---Insert Tables 1 and 2 Here---

In both tables the usual F statistic s_1^2/s_2^2 is the basis for the first four rows. They differ only in the way p values were obtained. Row 1 corresponds to the standard normal-theory test based on the F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom. Row 2 uses the F distribution with $d(n_1 - 1)$ and $d(n_2 - 1)$ degrees of freedom, where $d = [1 + \hat{\gamma}_2/2]^{-1}$ and $\hat{\gamma}_2$ is given in (2.3). This test was first proposed by Box and Andersen (1955). Rows 3 and 4 use p values obtained by bootstrapping and permutation sampling respectively, from $\bar{S}$ of (3.1).

Rows 5 and 6 are based on Miller's (1968) jackknife t statistic which is just the usual two-sample pooled t statistic for the log s^2 pseudovalues U_{ij}

in (2.2). In row 5 the p values are taken from a t distribution with $n_1 + n_2 - 2$ degrees of freedom. In row 6 the critical values are obtained by bootstrapping from $\tilde{S}$ of (3.1).

Row 7 is the Levl:med t statistic based on the Z_{ij} of (2.1) and using a t distribution with $n_1 + n_2 - 2$ degrees of freedom. Row 8 is a ratio of robust dispersion estimators $\hat{\sigma}_1/\hat{\sigma}_2$ with bootstrap critical values, where $\hat{\sigma}_i$ is the average of the first 50% of the ordered values of $|X_{ij} - X_{ik}|$, $1 \leq j < k \leq n_i$. This "Generalized L-statistic" on Gini type absolute differences (hence GLG) was found in Boos, et al. (1987) to perform well over a wide range of distributions. Note that for this statistic, we chose to base the bootstrap on samples centered with 20% trimmed means in place of the sample means in (3.1).

Since low χ^2_{1o} values are desired, we see that the bootstrap and permutation tests do well except at the exponential. The best test overall in terms of χ^2_{1o} values is actually the GLG bootstrap largely because of its performance at the exponential. The other bootstrap procedures would have performed better at the exponential if they had also used trimmed means to center. The jackknife t does poorly at unequal sample sizes and for skewed distributions. As mentioned in Section 2, O'Brien (1978) explains that this is due to unequal variances of the U_{ij} in different size samples, and to correlations in the U_{ij}. Bootstrapping produces a dramatic improvement in performance of the jackknife statistic.

Performance of the Box-Andersen test in Tables 1 and 2 is only mediocre, and the normal-theory F test is of course a disaster. The Levl:med t does reasonably well but its generally conservative levels result in higher χ^2_{1o} values. Note that Tables 1 and 2 are for one-sided tests. Results for two-sided tests were similar except that the GLG bootstrap did not do as

well, especially at the uniform where it was too conservative, and the Levl:med t had slightly lower χ^2_{To} values.

---Insert Table 3 here---

Table 3 gives estimates of the power of each one-sided test at (n_1 = 10, n_2 = 10) when the variance of the second population is four times that of the first. In parentheses is an estimate of the power which would have been obtained if correct .05 critical values had been used. The last columns of Table 3 are averages over the five distributions. As expected, the s^2 based tests are considerably more powerful at the uniform than the robust tests. It is interesting, however, that the robust tests do not dominate at the longer-tailed distributions. On average the Levl:med t comes out worst in terms of observed power and the F bootstrap and F permutation tests do quite well. Looking at adjusted power, Table 3 suggests that bootstrapping costs in terms of power (row 1 versus row 3, row 5 versus row 6) and so does studentization of statistics (row 1 versus row 5). Of course such techniques are essential to obtain valid tests based on s_1^2/s_2^2 at distributions other than the normal.

Four-Sample Results

Table 4 gives estimated levels for nominal level α = .05 and chi-squared goodness-of-fit statistics for p values for six tests, four distributions, and three sets of sample sizes. The normal and Laplace distributions and the sample sizes were chosen in order to make comparisons with Conover, et al. (1981).

---Insert Table 4 Here---

The first three tests relate to Bartlett's (1937) statistic T/C given in (2.3). Bar-χ^2 means that a χ^2_3 distribution was used for critical values and Bar-Boot means that the bootstrap approach was used. Bar2-χ^2 is the

Bartlett statistic with kurtosis adjustment given in (2.3) and using χ^2_3 critical values.

The fourth and fifth rows of Table 4, Jack-F and Jack-Boot, refer to the jackknife F statistic defined in (2.2) using either an F distribution or the bootstrap for critical values, respectively. Finally, Levl:med is the F test on the Z_{ij} in (2.1). Note that no permutation sampling from (3.1) was tried because we expect it to perform similarly to the bootstrap. Also, we did not run a GLG type test because of the computation costs.

In terms of χ^2_{10} values, these tests seem to perform worse than the two-sample versions in Tables 1 and 2. Bar-Boot and Jack-Boot do noticeably worse at the exponential compared to rows 3 and 6 of Tables 1 and 2. As before, however, these bootstrap results are dramatic improvements over using a χ^2 or F distribution. Jack-Boot seems to be the best performer in Table 4, but recall that odd sample sizes make Levl:med-F conservative. If one restricts attention to sample sizes (10,10,10,10), Levl:med-F is only mildly conservative and has very good χ^2_{10} values.

---Insert Table 5 Here---

In Table 5 we show the effect of using different centering estimators for bootstrapping the Bartlett and jackknife statistics. Rows 1 and 4 repeat some of the results for Table 4 where resampling is from $\bar{S}$ of (3.1). Rows 2 and 5 correspond to resampling from

$$\tilde{S} = \{X_{ij} - \tilde{X}_i,\ j = 1,\ldots,n_i,\ i = 1,\ldots,4\}\ , \qquad (4.1)$$

where $\tilde{X}_i$ is the ith sample 20% trimmed mean. Rows 3 and 6 of Table 5 correspond to $\tilde{X}_i$ = ith sample median. Note that the test statistic is the same for Rows 1, 2, 3 and for 4, 5, 6, only the critical values change. If medians are used for centering, the bootstrap procedures do quite well at the exponential but are then too conservative at the normal. The 20%

trimmed mean appears to be a reasonable compromise between centering at means or median. Recall that the GLG statistic had good performance in Tables 1 and 2 using a 20% trimmed mean for centering.

---Insert Table 6 Here---

Table 6 summarizes estimates of the power of the tests at the particular alternative H_a: $(\sigma_1^2,\sigma_2^2,\sigma_3^2,\sigma_4^2) = (1,2,4,8)$ averaged over the three sets of sample sizes found in Table 4. Adjusted power estimates are in parentheses. Note that these adjusted estimates are more variable than the observed powers and biased downward for Bar-χ^2 (a consequence of grouping p values into intervals of width .01). Excluding Bar-χ^2, Bar-Boot has the best power overall for either observed or adjusted power. Levl:med-F is second in average adjusted power. Note though, that at the Laplace Levl:med-F is still behind Bar-Boot in adjusted power, even though the mean absolute deviations from the median used in the numerator of Levl:med are maximum likelihood scale estimates for the Laplace. Perhaps Bar-Boot appears higher than it should in this case due to sampling variation. It is interesting that the studentized statistics Bar2 and Jack-F have quite low adjusted power relative to Bar-Boot.

5. Toxicity of Calcium Edetate

Evidence of toxicity in dosed animals is often reflected by an increase in variance of a response, because of differences among individuals in their ability to tolerate a given dose level. Data from a study on the toxicity (specifically teratogenicity) of calcium edetate (Brownie, et al. 1986) are used here to illustrate the tests for homogeneity of variances for k = 5 groups. In this study, several responses were measured on all animals (30 in the control group, 20 in all other groups), but food consumption was monitored on a smaller number of randomly chosen animals. Values for food

consumption for groups 1 to 5 (the control group and four increasing dose levels of calcium edetate) are given in Table 7.

---Insert Table 7 Here---

The six tests in Table 4 were applied to these data. In addition to bootstrapping from $\bar{S}$ to obtain p values for the Bartlett and jackknife statistics (as in Bar-Boot and Jack-Boot), we also bootstrapped from the median-centered residuals in $\tilde{S}$. The resulting tests are denoted Bar-Boot Med and Jack-Boot Med. For the original sample values, and for the residuals in $\bar{S}$, the pooled kurtosis was 4.78 ($\hat{\gamma}_2$ = 1.78). Kurtosis of the residuals in $\tilde{S}$ was 5.01.

P values (smallest to largest) for the eight tests were .008 for Bar-χ^2, .050 for Jack-F, .053 for Bar-Boot, .065 for Bar-Boot Med, .067 for Jack-Boot Med, .122 for Bar2-χ^2 and .157 for Levl:med-F. The small p value for Bar-χ^2 (Bartlett's test) is partly due to leptokurtosis in sample 5 (resulting in pooled $\hat{\gamma}_2$ = 1.78), but Jack-F and the bootstrap procedures all give p values near .05. The comparatively large p value for Levl:med is not due solely to $n_2 = n_3 = 7$ (small and odd) because deleting the fourth largest value in groups 2 and 3 (cf. Conover et al. 1981) gives p = .191. Instead, the conservative p value for Levl:med seems to be due to an extreme value (15.65) in group 5. Deleting this observation gives s_5^2 = 5.26, $\hat{\gamma}_2$ = 0.61, and p values .009 for Jack, .013 for Bar-χ^2, .017 for Jack-Boot, .018 for Jack-Boot Med, .037 for Bar-Boot, .042 for Bar-Boot Med, .046 for Bar2-χ^2 and .060 for Levl:med. There is no biological reason to delete this observation however, because it represents an animal that tolerated the high dose, as did at least two other animals in group 5 that were not monitored for food consumption.

To illustrate the use of the two-sample tests, suppose that we decide to compare group 3 with the control group 1. The one-sided p values for the

tests studied in Tables 1 - 3 (from smallest to largest) are .031 for the permutation F, .035 for Box-Andersen, .052 for the bootstrap F, .056 for the GLG bootstrap, .074 for the normal-theory F test, .085 for the jackknife bootstrap, and .089 for the Levl:med t. For the bootstrap and permutation procedures, p values were based on B = 10,000 resamples. It is interesting that all the resampling tests except for the jackknife have lower p values than the normal-theory F test. Apparently, these two samples together suggest shorter tails than the normal ($\hat{\gamma}_2 = -0.68$). Even so, the robust GLG bootstrap has good power.

6. Conclusions

Can the bootstrap method be trusted to determine critical values or p values when testing homogeneity of variance? Tables 1, 2, and 3 give strong support for its use in the two-sample problem. It performed well for the F ratio s_1^2/s_2^2, for the log s^2 jackknife t statistic, and for a ratio of the robust GLG dispersion estimators. It should work well for other statistics and distributions.

For more than two samples the picture is not as clear. The results suggest that skewed distributions are relatively more harmful to α-levels of the bootstrap procedures for $k > 2$ compared to $k = 2$. With skewed distributions it is wise to center with trimmed means before resampling. The power advantages of Bar-Boot should continue to hold for this centering. However, our results reinforce conclusions in Conover, et al. (1981) concerning the remarkable robustness of the Levl:med α levels. Thus when validity is more important than power and skewed data are expected (and especially with even sample sizes $n_i \geq 6$), use of Levl:med is very appropriate.

Finally, our results do not represent an exhaustive study of the bootstrap method in testing variance homogeneity. Generalizations to

problems where groups correspond to a factorial treatment structure are possible. The robust Levl:med also generalizes nicely to such problems, but as noted in O'Brien (1978) its use supposes a model with additive treatment effects on average absolute deviations. An advantage of the bootstrap approach is that it can be applied to the statistic of interest for a given model, and use of an arbitrary dispersion measure is avoided.

REFERENCES

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Journal of the Royal Statistical Society, Series A* 160, 268-282.

Bell, C. B. and Sen, P. K.(1984). Randomization procedures. In *Handbook of Statistics Volume 4*, P. R. Krishnaiah and P. K. Sen (eds.), 1-29, New York: North Holland.

Boos, D. D., Janssen, P., and Veraverbeke, N. (1986). Resampling from centered data in the two-sample problem. In preparation.

Boos, D. D., Janssen, P., Serfling, R. J., and Veraverbeke, N. (1987). Comparison of robust spread estimators. In preparation.

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

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

Brownie, C. F., Brownie, C., Noden, D., Krook, L., Haluska, M., and Aronson, A. L. (1986). Teratogenic effect of calcium edetate (Ca EDTA) in rats and the protective effect of zinc. *Toxicology and Applied Pharmacology* 82, 426-443.

Conover, W. J., Johnson, M. E., and Johnson, M. M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics* 23, 351-361.

Fairfull, R. W., Crober, D. C. and Gowe, R. S. (1985). Effects of comb dubbing on the performance of laying stocks. *Poultry Science* 64, 434-439.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics* 7, 1-26.

Gad, S. and Weil, C. S. (1986). *Statistics and experimental design for Toxicologists.* New Jersey: Telford Press.

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

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

Hall, I. J. (1972). Some comparisons of tests for equality of variances. *Journal of Statistical Computing and Simulation* 1, 183-194.

Keselman, H. J., Games, P. A., and Clinch, J. J. (1979). Tests for homogeneity of variance. *Communications in Statistics B-Simulation and Computation* 8, 113-129.

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

Levene, H. (1960). Robust tests for equality of variances. In *Contributions to Probability and Statistics,* I Olkin (ed.), 278-292. Palo Alto, California: Stanford University Press.

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

Nair, V. N. (1986). Testing in industrial experiments with ordered categorical data. *Technometrics* 28.

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

Schiffelbein, P. and Hills, S. (1984). Direct assessment of stable isotope variability in planktonic foraminifera populations. *Palaeogeography, Palaeoclimatology, Palaeoecology* 48, 197-213.

Schiffelbein, P. and Hills, S. (1985) Erratum. *Palaeogeography, Palaeoclimatology, Palaeoecology* 49, 361-362.

Zammuto, R. M. (1984). Relative abilities of three test to detect variance heterogeneity among mammalian litter sizes. *Canadian Journal of Zoology* 62, 2287-2289.

Table 1. Estimated Levels of One-Sided $\alpha = .05$ Tests under H_0: $\sigma_1^2 = \sigma_2^2$ for Symmetric Distributions

	Uniform				Normal				t_5			
(n_1,n_2) =	(10,10)		(5,15)		(10,10)		(5,15)		(10,10)		(5,15)	
	.05	χ^2_{10}	.05	χ^2_{10}	.05	χ^2_{10}	.05	χ^2_{10}	.05	χ^2_{10}	.05	χ^2_{10}
$F = s_1^2/s_2^2$												
F table	.014	50.2	.026	20.4	.055	6.0	.052	10.3	.104	101.7	.071	21.4
Box-Andersen	.048	9.4	.084	60.0	.069	12.3	.079	40.4	.050	16.4	.048	10.2
Bootstrap	.035	15.3	.032	19.0	.056	5.9	.045	7.9	.050	13.1	.035	6.7
Permutation	.041	13.4	.048	6.1	.058	3.4	.058	10.8	.048	11.7	.042	15.2
Jackknife												
t table	.029	20.6	.063	58.6	.054	9.6	.078	57.2	.053	6.0	.065	11.2
Bootstrap	.050	10.3	.052	9.7	.058	14.5	.058	6.0	.047	9.5	.043	9.4
Robust												
Levl:med t	.035	10.8	.070	25.4	.047	8.9	.063	42.4	.042	10.8	.039	15.4
GLG Bootstrap	.024	21.1	.033	11.9	.042	9.9	.044	13.7	.041	16.2	.035	11.4

Note: Estimated levels have standard deviation near $[(.95)(.05)/1000]^{-\frac{1}{2}} = .007$. χ^2_{10} 90% critical value is 16.0. Bootstrap and permutation are from pooled samples after subtracting sample means except for GLG which uses 20% trimmed means.

Table 2. Estimated Levels of One-Sided $\alpha = .05$ Tests under H_0: $\sigma_1^2 = \sigma_2^2$ for Extreme Value and Exponential Distributions.

	Extreme Value				Exponential				
(n_1,n_2) =	(10,10)		(5,15)		(10,10)		(5,15)		
$F = s_1^2/s_2^2$	.05	χ^2_{10}	.05	χ^2_{10}	.05	χ^2_{10}	.05	χ^2_{10}	Average χ^2_{10} for Tables 1 and 2
F table	.095	75.8	.078	38.6	.148	417.9	.152	270.7	101.3
Box-Andersen	.078	45.6	.050	12.8	.091	72.8	.076	34.0	31.4
Bootstrap	.058	25.6	.047	3.0	.078	27.0	.070	22.9	14.6
Permutation	.069	22.0	.055	5.7	.091	113.3	.082	44.8	24.6
Jackknife									
t table	.066	19.0	.063	44.8	.086	95.2	.098	129.5	45.2
Bootstrap	.057	19.2	.042	13.8	.071	27.2	.069	18.9	13.9
Robust									
Levl:med t	.046	3.9	.046	23.1	.046	6.5	.038	28.0	17.5
GLG Bootstrap	.046	14.4	.039	12.9	.052	10.8	.045	13.7	13.6

Note: Estimated levels have standard deviation near $[(.95)(.05)/1000]^{-\frac{1}{2}} = .007$. χ^2_{10} 90% critical value is 16.0. Bootstrap and permutation are from pooled samples after subtracting sample means except for GLG which uses 20% trimmed means.

Table 3. Observed and Adjusted Power of One-Sided $\alpha = .05$ Tests under H_a: $\sigma_2^2 = 4\sigma_1^2$, $n_1 = 10$, $n_2 = 10$

$F = s_1^2/s_2^2$	Uniform	Normal	t_5	Ext.-Value	Exponential	Average
F table	.70 (.88)	.60 (.58)	.61 (.43)	.61 (.47)	.58 (.29)	.62 (.53)
Box-Andersen	.77 (.78)	.57 (.48)	.49 (.49)	.53 (.42)	.40 (.24)	.55 (.48)
Bootstrap	.72 (.81)	.52 (.49)	.46 (.46)	.49 (.46)	.42 (.31)	.52 (.51)
Permutation	.74 (.78)	.52 (.49)	.46 (.47)	.50 (.42)	.43 (.29)	.53 (.49)
Jackknife						
t table	.78 (.85)	.52 (.50)	.45 (.44)	.46 (.41)	.37 (.25)	.52 (.49)
Bootstrap	.79 (.79)	.50 (.46)	.40 (.41)	.42 (.40)	.29 (.21)	.48 (.45)
Robust						
Levl:med t	.57 (.64)	.46 (.47)	.38 (.42)	.40 (.43)	.29 (.31)	.42 (.45)
GLG Bootstrap	.59 (.72)	.44 (.48)	.40 (.43)	.44 (.46)	.37 (.35)	.45 (.49)

Note: Basic entries have standard error bounded by $(4000)^{-\frac{1}{2}} = .016$. In parentheses are estimates of power for the tests with correct .05 levels.

Table 4. Estimated Levels of $\alpha = .05$ 4-sample Tests under H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$

(n_1,n_2,n_3,n_4) =	(5,5,5,5)		(10,10,10,10)		(5,5,20,20)		
	.05	χ^2_{10}	.05	χ^2_{10}	.05	χ^2_{10}	Av. χ^2_{10}
Normal							
Bar-χ^2	.054	12.5	.045	17.7	.047	6.3	12.2
Bar2-χ^2	.048	9.2	.053	10.7	.052	5.2	8.4
Bar-Boot	.027	15.0	.032	10.5	.039	13.3	12.9
Jack-F	.034	17.7	.054	12.9	.086	42.0	24.2
Jack-Boot	.044	11.5	.053	14.5	.053	15.9	14.0
Levl:med-F	.003	82.7	.037	17.2	.022	25.6	41.8
Laplace							
Bar-χ^2	.184	620.6	.261	1537.5	.266	1441.9	1200.0
Bar2-χ^2	.061	30.7	.047	21.3	.058	4.9	19.0
Bar-Boot	.068	33.7	.048	18.3	.093	70.7	40.9
Jack-F	.064	21.2	.082	50.6	.123	160.3	77.4
Jack-Boot	.062	12.8	.048	13.1	.068	18.9	14.9
Levl:med-F	.007	67.4	.033	13.5	.031	11.6	30.8
Extreme Value							
Bar-χ^2	.122	158.3	.188	623.6	.170	440.6	407.5
Bar2-χ^2	.058	21.0	.049	20.4	.056	12.4	17.9
Bar-Boot	.042	24.1	.050	16.2	.073	31.5	23.9
Jack-F	.058	4.2	.074	31.6	.118	121.4	52.4
Jack-Boot	.067	19.2	.061	14.2	.058	20.8	18.1
Levl:med-F	.003	77.5	.032	14.1	.022	28.2	39.9
Exponential							
Bar-χ^2	.303	2090.6	.407	5501.8	.387	4750.6	4114.3
Bar2-χ^2	.109	133.0	.100	74.9	.062	18.0	75.3
Bar-Boot	.130	261.0	.102	137.4	.127	179.9	192.8
Jack-F	.083	35.2	.133	243.2	.163	395.4	224.6
Jack-Boot	.082	45.1	.088	61.7	.073	28.2	45.0
Levl:med-F	.015	52.5	.042	9.4	.036	11.0	24.3

Table 5. Estimated Levels and χ^2 Results for Bootstrap Tests with Different Resample Sets

	Normal				Extreme Value				Exponential			
(n_1,n_2,n_3,n_4) =	(5,5,5,5)		(10,10,10,10)		(5,5,5,5)		(10,10,10,10)		(5,5,5,5)		(10,10,10,10)	
	.05	χ^2_{10}	.05	χ^2_{10}	.05	χ^2_{10}	.05	χ^2_{10}	.05	χ^2_{10}	.05	χ^2_{10}
Bar-Boot												
Mean	.027	15.0	.032	10.5	.042	24.1	.050	16.2	.130	261.0	.102	137.4
20% Trim	.017	38.4	.024	21.2	.023	23.8	.039	12.5	.076	35.2	.057	12.7
Median	.007	60.4	.015	35.6	.015	43.1	.024	26.0	.044	7.8	.042	10.5
Jack-Boot												
Mean	.044	11.5	.053	14.5	.067	19.2	.061	14.2	.082	45.1	.088	61.7
20% Trim	.039	15.4	.056	5.2	.064	18.3	.055	11.7	.072	20.3	.072	24.9
Median	.030	17.6	.054	4.1	.058	21.7	.050	14.0	.063	15.7	.072	24.7

Note: Mean, 20% trim, and median refer to resampling from (4.1) with $\tilde{X}_i = \bar{X}_i$, $\tilde{X}_i$ = ith sample 20% trimmed mean, and $\tilde{X}_i$ = ith sample median, respectively.

Table 6. Observed and Adjusted Power of 4-sample $\alpha = .05$ Tests under H_a: $(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2) = (1,2,4,8)$. Power is Averaged over Sample Sizes $(n_1, n_2, n_3, n_4) = (5,5,5,5)$, $(10,10,10,10)$, $(5,5,20,20)$.

	Normal	Laplace	Extreme Value	Exponential	Average
Bar-χ^2	.56 (.57)	.64 (.22)	.60 (.32)	.69 (.13)	.62 (.31)
Bar2-χ^2	.41 (.41)	.24 (.23)	.30 (.29)	.28 (.18)	.31 (.28)
Bar-Boot	.44 (.50)	.37 (.31)	.39 (.38)	.39 (.21)	.40 (.35)
Jack-F	.42 (.40)	.33 (.23)	.36 (.28)	.31 (.18)	.36 (.27)
Jack-Boot	.36 (.35)	.24 (.22)	.29 (.26)	.23 (.16)	.28 (.25)
Levl:med-F	.32 (.48)	.18 (.26)	.22 (.36)	.16 (.19)	.32 (.32)

Note: Entries have standard deviation bounded by .01. Adjusted entries in parentheses are less accurate.

Table 7. Data from a study on toxicity of calcium edetate (Brownie et al. 1986)

Group	n_i	Average daily food intake (g)								$\bar{X}_i$	s_i^2
1	6	17.98	18.25	21.08	18.50	18.26	19.95			19.0	1.53
2	7	16.42	16.45	16.58	18.08	19.14	17.40	18.53		17.5	1.20
3	7	14.27	17.15	13.67	17.72	11.57	18.33	14.42		15.3	6.14
4	8	11.48	9.45	8.17	12.68	10.25	5.08	17.50	16.33	11.4	16.95
5	12	7.78	5.88	6.55	4.88	4.95	8.77	5.17	4.10		
		9.25	1.92	3.03	15.65					6.5	13.09