Statistical Science
2004, Vol. 19, No. 4, 571–578
DOI 10.1214/088342304000000503
© Institute of Mathematical Statistics, 2004

Comparing Variances and Other Measures of Dispersion

Dennis D. Boos and Cavell Brownie

Abstract. Testing hypotheses about variance parameters arises in contexts where uniformity is important and also in relation to checking assumptions as a preliminary to analysis of variance (ANOVA), dose-response modeling, discriminant analysis and so forth. In contrast to procedures for tests on means, tests for variances derived assuming normality of the parent populations are highly nonrobust to nonnormality. Procedures that aim to achieve robustness follow three types of strategies: (1) adjusting a normal-theory test procedure using an estimate of kurtosis, (2) carrying out an ANOVA on a spread variable computed for each observation and (3) using resampling of residuals to determine p values for a given statistic. We review these three approaches, comparing properties of procedures both in terms of the theoretical basis and by presenting examples. Equality of variances is first considered in the two-sample problem followed by the k-sample problem (one-way design).

Key words and phrases: Comparing variances, measures of spread, permutation method, resampling, resamples, variability.

1. INTRODUCTION

Tests for equality of variances are of interest in a number of research areas. Increasing uniformity is an important objective in quality control of manufacturing processes (e.g., Nair and Pregibon, 1988; Carroll, 2003), in agricultural production systems (e.g., Fairfull, Crowber and Gowe, 1985) and in the development of educational methods (e.g., Games, Winkler and Probert, 1972). Biologists are interested in differences in the variability of populations for many reasons, for example, as an indicator of genetic diversity and in the study of mechanisms of adaptation.

Procedures for comparing variances are also used as a preliminary to standard analysis of variance, dose–response modeling or discriminant analysis. For example, SAS PROC TTEST presents the F test for equality of variances as a tool for choosing between the pooled variance t and the unequal variances Welch t. The POOL = TEST option in SAS PROC DISCRIM uses Bartlett's test of equal covariance matrices to decide between fitting linear and quadratic discriminant functions (SAS Institute, 1999). Fifty years ago, the use of such nonrobust variance procedures before relatively robust means procedures prompted Box (1953, page 333) to comment, "To make the preliminary test on variances is rather like putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port!" It is unfortunate that procedures condemned in 1953 are still in practice today, and Box's comment is still relevant. On the other hand, more robust procedures for testing equality of variances are now available, although many are not easily implemented with commercial software. Our plan is to describe the development and properties of these more robust procedures and so encourage their use.

Testing equality of variances, or other measures of scale, is a fundamentally harder problem than comparing means or measures of location. There are two reasons for this. First, standard test statistics for mean comparisons (derived assuming normality) are naturally standardized to be robust to nonnormality via the central limit theorem. (Here we refer to *robustness of*

Dennis D. Boos and Cavell Brownie are Professors, Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695, USA (e-mail: boos@stat.ncsu.edu; brownie@stat.ncsu.edu).

validity, that is, whether test procedures have approximately the correct level.) In contrast, normal-theory test statistics for comparing variances are not suitably standardized to be insensitive to nonnormality. Asymptotically, these statistics are not distribution-free, but depend on the kurtosis of the parent distributions. Second, for comparing means, a null hypothesis of identical populations is often appropriate, allowing the use of permutation methods that result in exact level-α test procedures for any type of distribution. For variance comparisons, a null hypothesis of identical populations rarely makes sense—at minimum, we usually want to allow mean differences in the populations. Given that it is necessary to adjust for unknown means or locations, permutation procedures do not provide exact, distribution-free tests for equality of variances, because after subtracting means, the residuals are not exchangeable.

There are three basic approaches that have been used to obtain procedures robust to nonnormality:

1. Adjust the normal theory test procedure using an estimate of kurtosis (Box and Andersen, 1955; Shoemaker, 2003).
2. Perform an analysis of variance (ANOVA) on a data set in which each observation is replaced by a scale variable such as the absolute deviation from the mean or median (Levene, 1960; Brown and Forsythe, 1974). A related procedure is to perform ANOVA on the jackknife pseudo-values of a scale quantity such as the log of the sample variance (Miller, 1968).
3. Use resampling to obtain p values for a given test statistic (Box and Andersen, 1955; Boos and Brownie, 1989).

Assuming normality leads naturally to tests about variance rather than to tests about other measures of dispersion. Approach 1 above is thus related to variance comparisons. The ANOVA and resampling methods, however, can be used with statistics that focus on other measures of dispersion. A reason for emphasizing variances is that variances (and standard deviations) are the most frequently used measures of dispersion and are building blocks in the formulation of many statistical models. On the other hand, test procedures that are based on alternative measures of scale, such as the mean absolute deviation from the median (MDM), can have superior Type I and Type II error properties. In fact, taking these properties into account, as well as simplicity of computation and interpretation, lead us to recommend approach 2, using ANOVA on the absolute deviation from the median as the best procedure. This was also the recommendation of the comprehensive review and Monte Carlo study of Conover, Johnson and Johnson (1981). In situations where power is more important than computational simplicity, we recommend a bootstrap version (approach 3) of this procedure (see Lim and Loh, 1996).

In Section 2 we briefly review the difference between asymptotic properties of normal-theory tests for variances and for means. Important ideas in the development of procedures robust to nonnormality using the three approaches above, as well as Monte Carlo estimates of performance, are reviewed in Section 3. Several specific methods for the two- and k-sample problems are described in more detail. Examples are given in Section 4, followed by concluding remarks in Section 5.

2. ASYMPTOTIC BEHAVIOR OF NORMAL-THEORY TESTS FOR VARIANCES

In the groundbreaking papers by Box (1953) and Box and Andersen (1955), a clear distinction was made between the Type I error robustness of tests on means and the nonrobustness of tests on variances. Box, in fact, coined the term "robustness" in his 1953 paper. We briefly outline the differences in robustness of the two types of tests in the two-sample and k-sample designs.

Our notation for the general k-sample problem is as follows. Let $X_{11}, \dots, X_{1n_1}, \dots, X_{k1}, \dots, X_{kn_k}$ be k independent samples, each sample being i.i.d. with distribution function $G_i(x)$, mean μ_i and variance σ_i^2, $i = 1, \dots, k$. The sample means and variances are $\overline{X}_i = n_i^{-1} \sum_{j=1}^{n_i} X_{ij}$ and $s_i^2 = (n_i - 1)^{-1} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_i)^2$, respectively, and the pooled sample mean and variance are $\overline{X} = N^{-1} \sum_{i=1}^{k} n_i \overline{X}_i$ and $s_p^2 = (N - k)^{-1} \sum_{i=1}^{k} (n_i - 1) s_i^2$, respectively, where $N = n_1 + \cdots + n_k$.

We first consider the case of two samples and the pooled t statistic:

$$t_p = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s_p^2 (1/n_1 + 1/n_2)}}.$$

Under the null hypothesis $H_0 : \mu_1 = \mu_2$ and assuming equal variances in the two populations, t_p is asymptotically a standard normal random variable *for any type of population distributions* G_1 and G_2. Thus, if the two populations have equal variances, t_p will be asymptotically distribution-free under H_0. Moreover, if variances

are not equal, t_p can be replaced by the asymptotically distribution-free Welch t,

$$t_W = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}.$$

In contrast, consider the logarithm of the normal-theory test statistic s_1^2/s_2^2 for $H_0: \sigma_1^2 = \sigma_2^2$, suitably standardized by sample sizes:

$$T = \left(\frac{n_1 n_2}{n_1 + n_2}\right)^{1/2} [\log s_1^2 - \log s_2^2]. \tag{1}$$

The usual test is to compare s_1^2/s_2^2 to an F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom. Asymptotically this test is equivalent to comparing T to a $\mathrm{N}(0, 2)$ distribution. However, if we assume a common underlying distribution function for both populations, $G_1(x) = F_0((x - \mu_1)/\sigma_1)$ and $G_2(x) = F_0((x - \mu_2)/\sigma_2)$, then under $H_0: \sigma_1^2 = \sigma_2^2$, T converges in distribution to a $\mathrm{N}(0, \beta_2(F_0) - 1)$ distribution, where $\beta_2(F_0) = \mu_4(F_0)/[\mu_2(F_0)]^2$ is the kurtosis of F_0 and $\mu_i(F_0) = \int [x - \int y\, dF_0(y)]^i\, dF_0(x)$ is the ith central moment of F_0. Some authors use the term "kurtosis" for $\gamma_2 = \beta_2 - 3$, but γ_2 is more properly called the *coefficient of excess* or *kurtosis excess*.

Location–scale population models such as $F_0((x - \mu)/\sigma)$ have the same kurtosis as the parent F_0. Normal distibutions have kurtosis $\beta_2 = 3$ and thus the $\mathrm{N}(0, \beta_2(F_0) - 1)$ distribution is the correct limiting $\mathrm{N}(0, 2)$ distribution of (1) when the populations are normal. However, if the populations have kurtosis greater than 3, commonly observed with real data, comparison of s_1^2/s_2^2 to an F distribution is asymptotically equivalent to comparing a $\mathrm{N}(0, \beta_2(F_0) - 1)$ random variable to a $\mathrm{N}(0, 2)$ distribution. For example, the true asymptotic level of a nominal $\alpha = 0.05$ one-sided test would be

$$P\left(Z > \sqrt{\frac{2}{\beta_2(F_0) - 1}}\, 1.645\right),$$

where Z is a standard normal random variable. Thus, if $\beta_2(F_0) = 5$, then the asymptotic level is 0.122. Table 1 of Box (1953) gave similar levels for two-sided tests for comparing two or more population variances. In fact, the problem is even worse when comparing more than two variances.

Assuming the populations are normally distributed, Bartlett's modified likelihood ratio statistic for H_0: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$ in the k-sample problem is given by B/C, where

$$B = \sum_{i=1}^{k} (n_i - 1) \log \frac{s_p^2}{s_i^2}, \tag{2}$$

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

and C is a correction factor to speed convergence to a χ_{k-1}^2 distribution.

Now let us relax the normal assumption and assume that each $G_i(x)$ is from the location–scale family generated by F_0, $G_i(x) = F_0((x - \mu_i)/\sigma_i)$. Then under H_0, B and B/C converge in distribution to $(1/2)[\beta_2(F_0) - 1]$ times a χ_{k-1}^2 random variable. Thus, for a test at nominal level α based on B/C with χ_{k-1}^2 critical values, the asymptotic level will be

$$P\left(\chi_{k-1}^2 > \left[\frac{2}{\beta_2(F_0) - 1}\right] \chi_{k-1}^2(1 - \alpha)\right),$$

where $\chi_{k-1}^2(1 - \alpha)$ refers to the upper $1 - \alpha$ quantile of the χ_{k-1}^2 distribution. For example, when comparing $k = 5$ population variances, if $\beta_2(F_0) = 5$, the true asymptotic level of a nominal $\alpha = 0.05$ test is 0.315.

In contrast, the one-way ANOVA F statistic for means,

$$F = \frac{1}{k-1} \sum_{i=1}^{k} n_i \frac{(\overline{X}_i - \overline{X})^2}{s_p^2}, \tag{3}$$

with $F(k - 1, N - k)$ critical values is asymptotically correct for all types of distributions G_i under the null hypothesis of equal means and assuming equal variances. Box (1953) and others have pointed out that (3) is fairly robust to unequal variances in the case of equal sample sizes.

3. DEVELOPMENT OF TYPE I ERROR ROBUST METHODS

In this section we explain three basic strategies for constructing Type I error correct methods for comparing variances or other measures of dispersion. All three approaches originate from ideas in Box (1953) and in Box and Andersen (1955). For specific procedures, we note whether asymptotic validity holds and summarize results for small sample properties obtained via Monte Carlo simulations, often citing the large survey study performed by Conover, Johnson and Johnson (1981) as well as the more recent studies by Boos and Brownie (1989), Lim and Loh (1996) and Shoemaker (2003).

3.1 Kurtosis Adjustment of Normal-Theory Tests on Variances

Box (1953, page 330) commented that an "internal" estimate of the variation of the sample variances

s_i^2 should be used to reduce dependency of a normal-theory test statistic on kurtosis. In that paper, Box proposed dividing each sample into subsets, computing the sample variance s_{ij}^2 for each subset and using (3) on the log s_{ij}^2. This ad hoc approach actually inspired the strategies of Section 3.2. The idea of an internal estimate of the variation also led to the proposal in Box and Andersen (1955) to divide the statistic B of (2) by a consistent estimate, based on sample cumulants, of $(1/2)[\beta_2(F_0) - 1]$. Layard (1973) proposed the less biased kurtosis estimator

$$\widehat{\beta}_2 = \frac{N \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_i)^4}{\left[\sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_i)^2\right]^2}. \tag{4}$$

Conover, Johnson and Johnson (1981) used Layard's $\widehat{\beta}_2$ to "correct" Bartlett's statistic. Thus they compared $B/[C(1/2)(\widehat{\beta}_2 - 1)]$ to percentiles of the χ_{k-1}^2 distribution and referred to the resulting procedure as Bar2. In their Monte Carlo study of 56 procedures, Conover, Johnson and Johnson (1981) found that Bar2 holds its level well for symmetric distributions, but is liberal for skewed distributions. For example, for the square of a double exponential distribution with $k = 4$ and $n_1 = n_2 = n_3 = n_4 = 10$, Conover, Johnson and Johnson (1981, Table 6) reported that the estimated level of Bar2 for nominal $\alpha = 0.05$ is 0.17. Boos and Brownie (1989) found that Bar2 lost power when compared to bootstrapping of Bartlett's B/C, especially for large k, say $k \geq 16$.

Shoemaker (2003) proposed several approximate F tests for the two-sample case by matching moments of $\log(s_1^2/s_2^2)$ with the moments of the logarithm of an F random variable. The estimated degrees of freedom use Layard's estimate of kurtosis (4). For $k > 2$, Shoemaker proposed using a sum of squares on $\log s_i^2$, standardized using (4), with χ_{k-1}^2 percentiles. Shoemaker's simulations suggest that, similar to Bar2, the procedures hold their level fairly well for symmetric distributions, but can be liberal for skewed distributions.

3.2 ANOVA on Scale Variables

Levene (1960) proposed using the one-way ANOVA F statistic (3) on new variables $Y_{ij} = |X_{ij} - \overline{X}_i|$ or, more generally, $Y_{ij} = g(|X_{ij} - \overline{X}|)$, where g is monotonically increasing on $(0, \infty)$. Miller (1968) showed that ANOVA on Levene's variables $|X_{ij} - \overline{X}_i|$ will be asymptotically incorrect if the population means are not equal to the population medians (essentially requiring symmetry) and that the problem can be corrected by using medians instead of means to center the variables. Brown and Forsythe (1974) and Conover, Johnson and Johnson (1981) studied the small sample properties of ANOVA on $Y_{ij} = |X_{ij} - M_i|$, where M_i is the sample median for the ith sample. Conover, Johnson and Johnson (1981) found that this procedure, referred to as Lev1:med, had satisfactory Type I and Type II error properties for a variety of distributions, although more recent simulation studies have shown that Lev1:med can be conservative with corresponding loss of power (Lim and Loh, 1996; Shoemaker, 2003). Conover, Johnson and Johnson (1981) also reported good properties for ANOVA on normal scores of the ranks of the $|X_{ij} - M_i|$. The asymptotic validity of this rank ANOVA has not been demonstrated, however, and O'Brien (1992) reported loss of power of the rank ANOVA compared to Lev1:med. Other spread variables have been proposed with the goal of producing F tests with good Type I error properties (e.g., O'Brien, 1978, 1979).

Miller (1968) provided the basis for an asymptotically correct approach to develop scale variables to which the ANOVA F test can be applied. The method is based on the jackknife and was prompted by Box's ANOVA on $\log s_{ij}^2$, the s_{ij}^2 being based on splitting each sample into two or more subsets. Miller applied his jackknife approach to $\log s_i^2$, which is not a scale estimator, but rather a monotone function of the scale estimator s_i. The F statistic (3) is calculated on the jackknife pseudovalues

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

where $s_{i(j)}^2$ is the sample variance in the ith sample with X_{ij} left out. Miller (1968) proved that the sample variances of the U_{ij} converge in probability to $\beta_2 - 1$ and that the means of the pseudovalues in the ith sample are asymptotically normal with mean $\log \sigma_i^2$ and variance $(\beta_2 - 1)/n_i$. Together these facts give the asymptotic correctness of (3) applied to the U_{ij}. In small samples, especially with unequal n_i, the procedure does not hold its level as well as ANOVA on the Levene-type variables $Y_{ij} = |X_{ij} - M_i|$ (e.g., Conover, Johnson and Johnson, 1981; Boos and Brownie, 1989).

The beauty of the jackknife approach, however, is that it can be applied to any chosen scale estimator for which the jackknife variance estimate is appropriate. For example, Gini's mean difference

$$\frac{1}{\binom{n_i}{2}} \sum_{j<k} |X_{ij} - X_{ik}| \tag{5}$$

is a measure of scale that could be used in Miller's jackknife-plus-ANOVA method. Although Gini's mean difference is not fully robust, it has similar robustness properties as the mean absolute deviation from the median. Moreover, as a U statistic it is unbiased for any sample size and, therefore, the mean of the pseudovalues for each sample is exactly equal to (5).

Procedures based on nonrobust scale quantities such as s_i^2, $\log s_i^2$ or MDM_i are surprisingly powerful for heavy-tailed distributions. However, in the face of extremely large outliers, say generated from the Cauchy distribution, one might want to consider more robust scale estimators such as those based on trimmed averages of the $|X_{ij} - X_{ik}|$ (e.g., GLG of Boos and Brownie, 1989, page 73) or single ordered values (Q_n of Rousseeuw and Croux, 1993). For these more robust scale estimators, the resampling methods of the next subsection are more appropriate because the corresponding jackknife variance estimates are deficient in small samples.

3.3 Resampling Methods

Box and Andersen (1955, page 16) introduced the permutation version of the two-sample test of $H_0: \sigma_1^2 = \sigma_2^2$ with known means based on $F' = n_2 \sum_{i=1}^{n_1} (X_{1i} - \mu_1)^2 / n_1 \sum_{i=1}^{n_2} (X_{2i} - \mu_2)^2$. They then transformed to a beta random variable and gave an F approximation to the permutation distribution involving the permutation moments that in turn are functions of the sample second and fourth moments. Estimating the means then led to an approximate F distribution (with estimated degrees of freedom) for the standard test statistic $F = s_1^2/s_2^2$. Thus, although motivated by the permutation approach, this method correctly belongs in Section 3.1.

Boos and Brownie (1989) implemented the permutation approach implied by Box and Andersen (1955). That is, they used the permutation distribution based on drawing samples without replacement from

$$(6) \quad \overline{S} = \{e_{ij} = X_{ij} - \widehat{\mu}_i, j = 1, \ldots, n_i, i = 1, \ldots, k\},$$

where the $\widehat{\mu}_i$ are location estimates such as the sample mean or trimmed mean. They implemented the approach in the two-sample case for $F = s_1^2/s_2^2$ and for a ratio of robust scale estimators. Note that because the residuals $X_{ij} - \widehat{\mu}_i$ from different samples are not exchangeable, such a permutation procedure is not exact, but Boos, Janssen and Veraverbeke (1989) proved that the procedure is asymptotically correct for a large class of U statistics.

Since the permutation method based on $\overline{S}$ is not exact, it makes sense to consider sampling with replacement from $\overline{S}$, in other words, to bootstrap resample from the residuals $X_{ij} - \widehat{\mu}_i$. Formally, the procedure is to generate B sets of resamples from $\overline{S}$ and compute the statistic of interest T for each set of resamples, resulting in $T_1^*, \ldots, T_B^*$. Then $\widehat{p}_B = \{\#\text{ of } T_i^* \geq T_0\}$ is an estimated bootstrap p value, where T_0 is the statistic calculated for the original sample. It might be more intuitive to think of generating residuals e_{ij}^* and then adding the location estimates $\widehat{\mu}_i$, resulting in a bootstrap set of resamples ($X_{ij}^* = \widehat{\mu}_i + e_{ij}^*$, $j = 1, \ldots, n_i, i = 1, \ldots, k$), as is the case for bootstrapping in regression settings. However, test statistics for comparing scale are invariant to location shifts, so that adding the $\widehat{\mu}_i$ back in is not necessary. Note also that drawing each sample from (6) produces a null situation of equal variablility (and equal kurtosis, etc.) that is crucial for the bootstrap to work correctly. We mention this because in applications of the bootstrap to obtain standard errors and confidence intervals, it is more usual to resample separately from the original individual samples.

Monte Carlo results in Tables 1–3 of Boos and Brownie (1989) show minor differences between the Box–Andersen approach with estimated degrees of freedom, the permutation approach based on $\overline{S}$ and the bootstrap approach based on $\overline{S}$. Generally, though, the bootstrap seemed to be the best procedure and so only bootstrap resampling was studied for the $k > 2$ cases. In those cases, the bootstrap approach applied to Bartlett's B/C statistic in (2) was comparable to the kurtosis adjusted version of B/C (called Bar2) in terms of Type I error, but the bootstrap had an advantage in terms of power. In fact, bootstrapping can be used to improve the Type I error (and possibly Type II error) properties of any reasonable procedure. Thus, Lim and Loh (1996) found that bootstrapping Lev1:med results in a procedure that is less conservative and has better power than using F percentiles to obtain critical values. It is also worth noting that bootstrapping a Studentized statistic such as Lev1:med produces better Type I error rates compared to bootstrapping an unstudentized statistic like B/C, due to improved theoretical convergence rates (see, e.g., Hall, 1992, page 150).

3.4 The Case Against Linear Rank Tests

Rank tests are noticeably absent from the robust methods reviewed. Tests for dispersion based on linear rank statistics do not compare favorably with other robust methods for the following reasons:

1. Linear rank statistics are not typically asymptotically distribution-free under the null of equal scales (i.e., do not hold the correct level in large samples) unless the medians are equal or are known and can be subtracted (rarely the case, but of course then permutation methods give exact tests), or the medians are estimated and subtracted before performing the rank tests and the underlying populations are symmetric (e.g., Boos, 1986, Theorem 1).
2. Even if the tests hold their level, Klotz (1962) found that small sample performance of linear rank tests did not match expectations based on their asymptotic relative efficiency (ARE) values. For example, the normal scores test has ARE $= 1$ compared to the F test for normal data. For two samples at $n_1 = n_2$ and $\alpha = 0.0635$, he obtained small sample efficiencies from 0.803 to 0.640 as the ratio of scale parameters ranged from 1.5 to 4. The most powerful rank test did not perform much better, having efficiencies from 0.810 to 0.688. Thus Klotz states, "It appears that the loss in efficiency is inherent in the use of ranks for small samples with a rather high price being paid for the insurance obtained with rank statistics."

4. EXAMPLE

To facilitate the study of pheromone blend composition in two interbreeding moth species (*Heliothis subflexa* and *Heliothis virescens*), Groot et al. (2005) investigated the use of pheromone biosynthesis activating neuropeptide (PBAN) to stimulate production of pheromone in mated females. Total production (and blend composition) was compared for virgin and mated PBAN-injected females in each of two seasons for both species. One of the interesting (and unexpected) findings was that variation in pheromone amounts among individuals appeared to be smaller for the mated PBAN-injected females than for virgin females. We use these data to illustrate a number of the tests for variance equality described in Section 3.

Measurements of the total amount of pheromone (nanograms) produced by females in each of eight groups are displayed graphically in Figure 1. Table 1 gives summary statistics for each group. The groups are arranged by species (Hs, Hv), season $(1, 2)$ and PBAN ($0 =$ virgin female; $1 =$ mated PBAN-injected female). The data tend to be right-skewed with evidence of variance heterogeneity. Before carrying out an ANOVA to compare means, common practice would be to log-transform or square-root-transform the data.

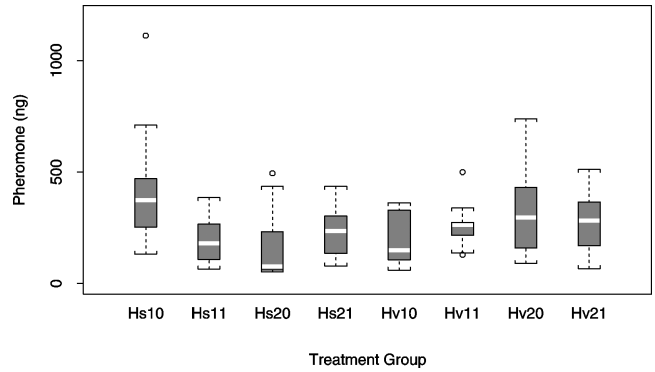

FIG. 1. *Box plots of total pheromone for eight treatment groups: Hsij refers to the Heliothis subflexa in season $i = 1$ or $i = 2$ and $j = 0$ (virgin) or $j = 1$ (injected). Notation for the four Heliothis virescens groups is similar.*

When interest is in variability, rather than mean response, there is less justification for employing a "variance stabilizing" transformation. We therefore carried out tests for differences in variation, both on the original scale and on square-root-transformed data (a log-transformation produced negative skewness).

Table 2 gives the p values for the following comparisons:

1. Bartlett's statistic B/C from (2) compared to χ^2_{k-1} critical values (Bartlett χ^2).
2. Statistic B/C divided by $(1/2)[\widehat{\beta}_2 - 1]$ compared to χ^2_{k-1} critical values (Bar2 χ^2).
3. Shoemaker's $X^2 = \sum_{i=1}^k (n_i - 1)(Z_i - \overline{Z})^2/(\widehat{\beta}_2 - (n_i - 3)/n_i)$, where $Z_i = \log s_i^2$, compared to χ^2_{k-1} critical values (Shoe X^2).
4. The ANOVA F (3) on the $|X_{ij} - M_i|$ compared to F critical values (Lev1:med F).
5. Statistic B/C compared to bootstrap critical values with $B = 9999$ resamples using fractional 20% trimmed means in (6) (Bar boot trim).
6. The ANOVA F (3) on the $|X_{ij} - M_i|$ compared to bootstrap critical values (Lev1:med boot trim).

TABLE 1
Summary statistics for pheromone production

Statistic	*Hs*10	*Hs*11	*Hs*20	*Hs*21	*Hv*10	*Hv*11	*Hv*20	*Hv*21
Sample size	13	23	13	14	7	9	18	14
Mean	414	201	166	231	196	260	317	279
Std. dev.	261	101	149	106	116	112	194	130
MDM	175	82	103	85	91	95	147	107

NOTE. MDM is the mean absolute deviation from the median.

TABLE 2
p values for variance equality among groups in the pheromone study

Test	1 vs. 2	All groups	PBAN
Bartlett χ^2	0.000	0.002	
Bar2 χ^2	0.056	0.306	
Shoe X^2	0.049	0.390	
Lev1:med F	0.039	0.192	0.058
Bar boot trim	0.028	0.144	
Lev1:med boot trim	0.021	0.143	

NOTE. PBAN main effect contrast.

To illustrate two-sample tests we consider only groups 1 and 2, and report the two-tailed p values in column 1 of Table 2. The Bartlett test is very significant, but Layard's $\widehat{\beta}_2$ in (4) is 8.7, suggesting that the Bartlett test cannot be trusted. Each of the five robust tests suggests that variation is lower in the PBAN-injected females, with the bootstrapped Bartlett and Lev1:med having lower p values than Shoemaker's X^2 and Bar2.

Equality of variances for the eight groups was tested, ignoring the 2^3 factorial structure, to illustrate the k-group analysis. Layard's estimate of β_2 for the eight groups was 6.5. Results for the tests appear in column 2 of Table 2, and ignoring Bartlett χ^2, only the bootstrap tests are close to significance, agreeing with Monte Carlo results that show greater power compared to procedures that utilize an explicit correction for kurtosis excess.

With the exception of the bootstrap procedures, test statistics and p values can be obtained, with a little effort, using standard software such as SAS (SAS Institute, 1999). (PROC MEANS followed by a data step can be used to compute absolute deviations from medians, and also second and fourth central sample moments.) The Lev1:med F is produced directly for $k \geq 2$ groups using the MEANS/HOVTEST = BF option in SAS PROC GLM. The bootstrap procedures are not available in any commercial software that we know of.

Finally, to address the question of whether PBAN increased uniformity of response, Lev1:med was used to test the main effect contrast for PBAN (easily computed by applying PROC GLM, followed by a CONTRAST statement, to the $|X_{ij} - M_i|$ values). The third column p value of 0.058 is consistent with the results for groups 1 and 2, and suggests that variation in the amount of pheromone produced is lower among mated PBAN-injected females compared to virgin females.

When the same analyses are carried out on square-root-transformed values (not shown), an indication of increased uniformity among mated PBAN-injected females is provided by the bootstrapped Bartlett test for groups 1 and 2 (p value $= 0.098$) and by the PBAN main effect contrast (Lev1:med, p value $= 0.074$).

5. CONCLUDING REMARKS

We have presented the three main approaches for comparing measures of scale or spread with proper Type I error control under a range of distributional types. The computationally simplest Type I error robust procedure is based on comparing means of the scale variable $Y_{ij} = |X_{ij} - M_i|$. This appealing approach (Lev1:med) is based on an efficient scale estimator—the mean absolute deviation from the median—and can be generalized to factorial designs and multivariate data.

A number of Monte Carlo studies have found, however, that in the k-sample problem Lev1:med has significance levels below the nominal level, and especially low if the sample sizes are small and odd. Lim and Loh (1996) demonstrated that this conservatism can be eliminated by using the bootstrap of Section 3.3, but then the computational simplicity is lost. Shoemaker (2003) argued persuasively for kurtosis-adjusted normal-theory methods in situations where the distributions are approximately symmetric. Simulations by Shoemaker (2003) and Lim and Loh (1996) showed good power properties for these methods when $k \leq 4$. However, our own simulations showed that for larger k ($k = 8$ not shown; $k = 16, 18$ in Boos and Brownie, 1989, Table 6), these methods lack power compared to Lev1:med (even with the more conservative F percentiles) and bootstrapping Bartlett's statistic.

If computational simplicity is not important, we agree with Lim and Loh's (1996) recommendation for the use of Lev1:med with bootstrap critical values. Moreover, once a decision is made to bootstrap, there is incentive to consider a statistic based on one of the more robust scale estimators mentioned at the end of Section 3.2.

ACKNOWLEDGMENT

We thank Astrid Groot, Department of Entomology, North Carolina State University, for permission to use the data in the example.

REFERENCES

BOOS, D. D. (1986). Comparing k populations with linear rank statistics. *J. Amer. Statist. Assoc.* **81** 1018–1025.

BOOS, D. D. and BROWNIE, C. (1989). Bootstrap methods for testing homogeneity of variances. *Technometrics* **31** 69–82.

BOOS, D. D., JANSSEN, P. and VERAVERBEKE, N. (1989). Resampling from centered data in the two-sample problem. *J. Statist. Plann. Inference* **21** 327–345.

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

BOX, G. E. P. and ANDERSEN, S. L. (1955). Permutation theory in the derivation of robust criteria and the study of departures from assumption (with discussion). *J. Roy. Statist. Soc. Ser. B* **17** 1–34.

BROWN, M. B. and FORSYTHE, A. B. (1974). Robust tests for the equality of variances. *J. Amer. Statist. Assoc.* **69** 364–367.

CARROLL, R. J. (2003). Variances are not always nuisance parameters. *Biometrics* **59** 211–220.

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.

GAMES, P. A., WINKLER, H. B. and PROBERT, D. A. (1972). Robust tests for homogeneity of variance. *Educational and Psychological Measurement* **32** 887–909.

GROOT, A., FAN, Y., BROWNIE, C., JURENKA, R. A., GOULD, F. and SCHAL, C. (2005). Effect of PBAN on pheromone production by mated *Heliothis virescens* and *Heliothis subflexa* females. *J. Chemical Ecology* **31** 15–28.

HALL, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer, New York.

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 variances. *J. Amer. Statist. Assoc.* **68** 195–198.

LEVENE, H. (1960). Robust tests for equality of variances. In *Contributions to Probability and Statistics* (I. Olkin, ed.) 278–292. Stanford Univ. Press, Stanford, CA.

LIM, T.-S. and LOH, W.-Y. (1996). A comparison of tests of equality of variances. *Comput. Statist. Data Anal.* **22** 287–301.

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

NAIR, V. and PREGIBON, D. (1988). Analyzing dispersion effects from replicated factorial experiments. *Technometrics* **30** 247–257.

O'BRIEN, P. C. (1992). Robust procedures for testing equality of covariance matrices. *Biometrics* **48** 819–827.

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

O'BRIEN, R. G. (1979). A general ANOVA method for robust tests of additive models for variances. *J. Amer. Statist. Assoc.* **74** 877–880.

ROUSSEEUW, P. J. and CROUX, C. (1993). Alternatives to the median absolute deviation. *J. Amer. Statist. Assoc.* **88** 1273–1283.

SAS INSTITUTE INC. (1999). SAS online doc, version 8. SAS Institute Inc., Cary, NC.

SHOEMAKER, L. H. (2003). Fixing the F test for equal variances. *Amer. Statist.* **57** 105–114.