A General ANOVA Method for Robust Tests of Additive Models for Variances

Ralph G. O'Brien


Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28197912%2974%3A368%3C877%3AAGAMFR%3E2.0.CO%3B2-1

A General ANOVA Method for Robust Tests of Additive Models for Variances

RALPH G. O'BRIEN*

Linearly combining Levene's $z^2$ variable with the jackknife pseudo-variables of $s^2$ produces a family of variables that allows for analysis of variance (ANOVA) tests of additive models for the variances in fixed effects designs. Some distributional theory is developed, and a new robust homogeneity of variance test is advocated.

KEY WORDS: Homogeneity of variance tests; Jackknifing variances; Dispersion; Spread; Analysis of variance; Testing variances using ANOVA.

1. INTRODUCTION

There are now many techniques that test homogeneity of variance (HOV) hypotheses by applying the analysis of variance (ANOVA) to dependent variables that are constructed to measure the spread (a more general term than variance) of each group's distribution. For example, Levene (1960) suggested the spread variable, $z^2_{ij} = (y_{ij} - \bar{y}_j)^2$, where $y_{ij}$ is the $i$th observation in the $j$th group. Other spread variables have been investigated by Bartlett and Kendall (1946), Box (1953), Brown and Forsythe (1974a), Games, Winkler, and Probert (1972), Gartside (1972), Layard (1973), Levy (1975), Martin (1976), Martin and Games (1977), Miller (1968), Mosteller and Tukey (1968), and O'Brien (1978). Although there is a legitimate controversy concerning the relative merits among these ANOVA-based tests, the consensus is that they are much more robust to distributional form than the traditional normal theory procedures, such as Bartlett's test, Hartley's $F$-max test, and Cochran's test.

The success of the ANOVA-based tests is due largely to the fact that the estimates of the variabilities of the average spreads are obtained directly from the data and consequently are sensitive to the kurtosis ($\gamma_2$) of the parent distribution. The traditional procedures base such variability estimates on theoretical properties that are tied directly to the normality assumption (specifically that $\gamma_2 = 0$) and are valid only when such an assumption is satisfied. In fact, such tests are not even asymptotically distribution free.

This article describes an ANOVA spread variable that allows HOV tests to be conducted by using common additive fixed effects models for the variances, $\sigma_j^2$, rather than the means, $\mu_j$. The sample variances, $s_j^2$, replace the sample means, $\bar{y}_j$, as the focus of attention. If some type of factorial design defines the relationships of the $J$ groups, the concepts of main effects and interactions among the variances conform to the traditional definitions commonly applied to the means.

2. THE $r_{ij}(w)$ VARIABLE AND ITS PROPERTIES

The spread variable examined here is

$$r_{ij}(w) = \left[ \frac{((w + n_j - 2)n_i(y_{ij} - \bar{y}_j))^2 - wn_i^2(n_j - 1)}{(n_i - 1)(n_j - 2)} \right].$$

Now

$$r_{ij}(0) = n_i(y_{ij} - \bar{y}_j)^2/(n_j - 1) = z_{ij}^2,$$

a slight modification of Levene's $z^2$ variable. Of course, for balanced designs ($n_j = n$), $z^2$ and $z^2$ produce identical $F$ tests. At the other extreme,

$$r_{ij}(1) = \left[ n_i(y_{ij} - \bar{y}_j)^2 - s_{ij}^2 \right]/[n_j - 2],$$

where $s_{ij}^2$ is the sample variance of group $j$ if the $i$th observation is deleted; that is, $r_{ij}(1)$ is a jackknife pseudovalue of $s_j^2$ (Miller 1968).

Several investigators have considered $z^2$, $z^2$, and/or $q$ (Levene 1960; Miller 1968; Games, Winkler, and Probert 1972; O'Brien 1978). When used in a regular ANOVA, $z^2$ and $z^2$ produce moderately inflated empirical Type I error rates in most situations. They also produce relatively low power in designs with very small total sample sizes ($N$), but are competitive to other robust methods when the designs have moderate $N$. The $q$ variable produces conservative rejection rates and has less power than $z^2$. The $r(w)$ variable is simply a weighted average of $z^2$ and $q$ and provides a way to balance the inflated test sizes of $z^2$ and the conservative test sizes of $q$.

An argument will be made that a modification to the degrees of freedom for the ANOVA $F$ test will increase the power.

Regardless of the choice of $w$,

$$\bar{r}_j(w) = \frac{\sum_{i=1}^{n_j} r_{ij}(w)}{n_j} = s_j^2.$$

Thus, ANOVA tests using $r_{ij}(w)$ are readily interpretable because they conform to tests of additive models for...
Equating (2.10) and (2.11) and solving for \( w \) produces the weighting factor that yields unbiased estimates for \( \text{var}[\bar{r}_j(w)] \):

\[
w^* = \left[ \frac{2n_j(n_j - 1)(n_j - 2) + \gamma_2(n_j - 1)^2(n_j - 2)}{2n_j + (n_j - 2)\gamma_2} \right]^{-1}
\]

The table contains values of \( w^* \) for various parent distributions. The value of \( w^* \) increases as \( \gamma_2 \) increases, but is not significantly affected by \( n_j \). It can be shown that \( w^* < 1 \); thus, the jackknife variable, \( r(1) \), is always conservative.

If the design is balanced and \( \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_J^2 \), it can be shown that

\[
E[M_{SH}] = E[M_{SWG}(w^*)] = \text{var}[\bar{s}_j^2]n
\]

for any testable null hypothesis.

In order to examine the relationships of \( E[M_{SH}] \) and \( E[M_{SWG}(w^*)] \) for unbalanced designs, consider the single degree of freedom contrast

\[
H_0 = \sum_{j=1}^J c_j \sigma_j^2 = 0 \quad \text{where} \quad \sum_{j=1}^J c_j = 0
\]

It can be shown that

\[
E[M_{SH}] = \sum_{j=1}^J c_j \text{var}[\bar{s}_j^2]/(\sum_{j=1}^J c_j^2/n_j)
\]

\[
E[M_{SWG}(w^*)] = \sum_{j=1}^J n_j(n_j - 1)\text{var}[\bar{s}_j^2]/(N - J)
\]

These formulas can be used to infer several properties that are true regardless of the value of \( w \).

1. If \( |c_j| = 1 \) and \( \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_J^2 \), then \( E[M_{SH}] \) increases relative to \( E[M_{SWG}(w)] \) when the design is unbalanced, because \( n_j \) \text{var}[\bar{s}_j^2] \) decreases slightly as \( n_j \) increases. If the design is nearly balanced, this heterogeneity of the variances of \( r(w) \) should have little effect on empirical rejection rates.

2. If \( \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_J^2 \) and the cells with smaller \( n_j \) have the larger \( c_j \), then \( E[M_{SH}] \) increases relative to \( E[M_{SWG}(w)] \).

3. If some cells are not involved in the contrast \( (c_j = 0) \) yet have relatively low \( \sigma_j^2 \), their low \text{var}[\bar{s}_j^2] \) reduces \( E[M_{SWG}(w)] \) relative to \( E[M_{SH}] \), even for balanced designs.

Several converses to the second and third properties also are immediately evident, but will not be stated here. These results extend to tests with \( df_H > 1 \), because any such test can be formulated in terms of a set of \( df_H \) single degree of freedom orthogonal contrasts. When the structure of the analysis indicates that these properties may cause problems, it may be prudent to conduct a Welch-type ANOVA that does not assume homogeneity of variance (Brown and Forsythe 1974b; Kohr and Games 1977).
Unlike the familiar normal theory result, \( \hat{r}_j(w) \) and \( \hat{P}[\hat{r}_j(w)] \) are not independent; in fact, they are highly correlated. Because an analytical expression for this correlation, \( \hat{\rho}_j \), was not obtainable, Monte Carlo estimates of \( \hat{\rho}_j \) were computed for several parent distributions with \( \gamma_2 \) varying from \(-1.2\) to \(3\) and \( n = 8, 16, \) and \( 32 \). Estimates of \( \hat{\rho}_j \) ranged from .55 (uniform parent, \( n = 32 \)) to .79 (Laplace parent, \( n = 8 \)). This correlation results naturally from (2.11) and therefore must be present to some degree in every spread variable for \( \sigma^2 \).

To assess the effect of this relationship, consider testing the contrast (2.14) by using

\[
t = \sum_{j=1}^{J} c_j \hat{\rho}_j^2 / [\left( \sum_j c_j^2 / n_j \right) \text{MSWG}(w)]^{1/2} \tag{2.17}
\]
as a \( t \) random variable with \( N - J \) degrees of freedom. It can be shown that

\[
\text{corr}\left\{ \sum_{j=1}^{J} c_j \hat{\rho}_j^2, \text{MSWG}(w) \right\} = \frac{\sum_{j=1}^{J} c_j n_j (n_j - 1) \hat{\rho}_j}{\left[ \left( \sum_j c_j^2 \right) \sum_j n_j^2 (n_j - 1)^2 \right]^{1/2}}. \tag{2.18}
\]

If the design is balanced (and all groups have the same distribution), then \( \hat{\rho}_j = \hat{\rho} \) and the correlation (2.18) is zero. It deviates somewhat from zero for unbalanced designs, although its magnitude is usually small. When the correlation between the numerator and denominator of a \( t \) is positive, its distribution tends to have larger lower tails and smaller upper tails than otherwise. If this correlation is negative, the opposite pattern occurs. Because the rejection rates for the two directional alternatives to \( H_0 \) are unequal, one-tailed tests should be used with caution.

Because the correlation between \( \hat{r}_j(w) \) and \( \hat{P}[\hat{r}_j(w)] \) is so strong, the single group test of \( H_0: \sigma^2 = \sigma^2 \) using

\[
t_{n-1} = \left| \hat{r}(w) - \sigma^2 \right| / \left[ \hat{P}[\hat{r}(w)] / n \right]^{1/4} \tag{2.19}
\]

and the associated confidence intervals for \( \sigma^2 \) should not be used. This explains why Lemmer (1978) obtained extremely low rejection rates when he used \( z^2 \) to test \( H_0: \sigma^2 = \sigma^2 \) versus \( H_a: \sigma^2 > \sigma^2 \).

It follows from (2.1) that

\[
\lim_{n_j \to \infty} r_{ij}(w) = (y_{ij} - E[\hat{p}_j])^2. \tag{2.20}
\]

The limiting value of the kurtosis of \( r_{ij}(w) \) is

\[
\gamma[r(w)] = \frac{\mu_4 - 4\sigma^4 \mu_4 + 6\sigma^2 \mu_4 - 3\sigma^4}{(\mu_4 - \sigma^4)^2} - 3, \tag{2.21}
\]

where \( \mu_4 \) is the kth central moment of the parent distribution. The table contains values of \( \gamma[r(w)] \) for several parent distributions. If \( y \) is normally distributed, then the limiting distribution of \( r(w) \) is \( \chi^2(n) \), which has a kurtosis of 12. For the exponential parent (\( \gamma_2 = 6 \)), \( \gamma[r(w)] = 213 \). Monte Carlo estimates of \( \gamma[r(w)] \) for finite \( n \) paralleled these limiting values.

Box and Andersen (1955) showed that when all other fixed effects ANOVA assumptions are met, except parent normality, then \( F = \text{MSH}/\text{MSWG} \) is approximately distributed as an \( F \) with \( \delta \cdot \text{df}_H \) and \( \delta \cdot \text{df}_W \) degrees of freedom, where \( \delta = 1 + \gamma_2 / N \). If this result applies to \( r(w) \), the high values of \( \gamma[r(w)] \) reduce the empirical rejection rates for the customary \( F \) test, especially in cases with low \( \text{df}_H \) and low \( N \). Of course, these results suggest possible adjustments to the \( F \) test.

It should be noted that these properties are not invariant to the usual data transformations, such as log or square root. For example, log \( z^2 \) produces extremely inflated regular ANOVA tests (O’Brien 1975), and ANOVA’s using \( |z| \) are not asymptotically distribution free (Miller 1968).

3. A Single “Utility” Test for Most Situations

Most interval level data encountered by researchers are not characterized easily by one of the standard parent distributions. Thus, precise calculations of \( w^* \) and \( \gamma[r(w)] \) will usually be impractical. Nevertheless, most researchers will be satisfied with a single “utility” test that works satisfactorily in a majority of situations. The \( r_{ij}(5) \) variable might assume this role, because \( E[\hat{P}[\hat{r}_{ij}(5)] \) is nearly unbiased under the normal parent. Using similar logic, one would use \( \delta = 1 + 12 / N \) and adopt the \( F(\delta \cdot \text{df}_H, \delta \cdot \text{df}_W) \) distribution as the sampling distribution of \( F(5) = \text{MSH}/\text{MSWG}(5) \).

These choices are based on the philosophy that empirical and nominal Type I error rates should be synchronized for the normal parent. Readers disagreeing with this conventional view should have little trouble adjusting \( w \) and \( \delta \) to conform with their own criteria for robustness.

In the event that \( w^* \) and \( \delta^* = 1 + \gamma[r(w)] / N \) can be easily determined, then of course they should be used. The consequences of using estimates of \( \gamma_2 \) and \( \gamma[r(w)] \) to specify values for \( w \) and \( \delta \) have not been determined. This strategy may prove to be useful, however, because Bartlett’s HOV test is improved (but not salvaged) by modifications based on estimates of \( \gamma_2 \) (Box and Andersen 1955; Miller 1968; Games, Winkler, and Probert 1972).

4. Empirical Investigation

Levene (1960) studied \( z^2 \) for uniform, normal, and Laplace parents and reported the empirical .05 critical values, \( F_E(z^2) = F_E(r(0)) \), for the usual balanced one-way ANOVA tests with \( J = 2, n = 10 \); \( J = 2, n = 20 \); \( J = 4, n = 20 \); and \( J = 10, n = 20 \). O’Brien (1978) studied \( r(0) \) for \( 2 \times 2 \) and \( 4 \times 3 \) designs, with \( n = 12 \) and \( 24 \), and uniform, normal, and exponential parents and saved the \( F_E(r(0)) \) values. For the present study, these \( F_E(r(0)) \) values were converted by (2.7) to \( F_E(r(w)) \) and then compared with nominal critical values, \( F_N(\delta) = F(\delta \cdot \text{df}_H, \delta \cdot \text{df}_W, .95) \). When various values
for $w$ and $\delta$ were selected, the general theory and the proposed utility test were examined empirically for these balanced designs.

These analyses will not be detailed here, because their results do closely parallel the theoretical conclusions. The $F_N[r(w^*)]$ values were reasonably close to $F_N(\delta = 1 + \frac{r(w)}{N})$. These results also supported the conjecture that the utility test is essentially robust. Comparing $F_N[r(0.1)]$ with $F_N(\delta = 1 + 12/N)$ showed that this test is mildly conservative for platykurtic parents and mildly inflated for leptokurtic parents, although the high $\hat{r}(w)$ produces conservative rejection rates if $N$ is small and the parent distribution is leptokurtic.

At least for balanced designs, all characterizations of $r(w)$ produce the same power if they are held to the same Type I error rate. Accordingly, this theory focuses on the problem of obtaining proper rejection rates. The $r(0.1)$ with $F_N(\delta = 1 + 12/N)$ usually provides such rates and is uniformly more powerful than the use of $q$ with $F_N(\delta = 1)$, which was previously recommended for common use.

A review of the literature suggests that no single spread variable produces the most power in all situations. The only alternative to $r(w)$ that also effectively tests linear contrasts among the $\sigma_i$ is the unlogged version of Box-Scheffé subgrouping method. This method is basically less efficient because, if $v_{ij}$ is the sample variance of subgroup $i$ in group $j$ and the subgroup size, $m$, is an even divisor of $n_i$, then

$$\text{var}[r_{ij}] = \sigma_i^2/n_i + \frac{2m}{m - 1}v_{ij} \geq \text{var}[s_i^2] = \text{var}[\hat{r}_j(w)]. \quad (4.1)$$

### 5. CONCLUSION

A general spread variable, useful for testing HOV using ANOVA methodology, has been introduced, and its distributional theory has been developed and related to the properties of ANOVA. One characteristic of the method has been suggested for common use. Usually the method is well behaved, although some applications do produce rejection rates that deviate in predictable directions from the nominal rate. The $r(w)$ variable should not be further transformed and is poorly suited for one-group tests and confidence intervals for $\sigma^2$.

[Received March 1978. Revised April 1979.]

**REFERENCES**


