PSYCHOMETRIKA—VOL. 43, NO. 3.
SEPTEMBER, 1978

ROBUST TECHNIQUES FOR TESTING HETEROGENEITY OF VARIANCE EFFECTS IN FACTORIAL DESIGNS

RALPH G. O'BRIEN

UNIVERSITY OF VIRGINIA

Several ways of using the traditional analysis of variance to test heterogeneity of spread in factorial designs with equal or unequal n are compared using both theoretical and Monte Carlo results. Two types of spread variables, (1) the jackknife pseudovalues of s^2 and (2) the absolute deviations from the cell median, are shown to be robust and relatively powerful. These variables seem to be generally superior to the Z-variance and Box-Scheffé procedures.

Key words: variance testing, homogeneity of variance, analysis of variance, jackknife, dispersion.

Recently several procedures have been suggested for testing homogeneity of variance hypotheses in factorial designs. Overall and Woodward [1974] introduced the Z-variance test; Levy [1975] suggested extending the Box-Scheffé method to factorial designs. Martin [1976] criticized Levy's use of subgroup sizes and suggested some other alternatives. Much earlier, Zelen [1959, 1960] developed a set of likelihood ratio tests which are two-way generalizations of Bartlett's test for homogeneity of variance in the one-way design. For normally distributed data, it appears that the Z-variance and Zelen procedures are more powerful than the Box-Scheffé test. However, the Box-Scheffé is quite robust to the distributional form of the data, while the Z-variance and Zelen procedures are overly sensitive to nonnormal data. In this paper, alternatives to these procedures are investigated and it is argued that two of these other methods are quite robust and yet more powerful than the Box-Scheffé. This argument is based on both theoretical and Monte Carlo results.

The basic procedure examined here uses estimates of each independent group's (cell's) "spread" (a more general term than "variance") as the dependent measure in a normal theory fixed effects least squares analysis of variance. The statistical designs for such analyses would be identical to those used for an ANOVA means analysis: Only the dependent variable is changed. The Box-Scheffé method is simply one way to generate those "spread variables." The alternative measures studied here are (1) a modification of Levene's [1960] z^2 (or s) variable, (2) the jackknife pseudovalues of $\log s^2$, (3) the jackknife pseudovalues of s^2, and (4) Brown and Forsythe's [1974] *W50* variable. Many other spread variables have appeared in the literature, but are not considered here. Those that are considered have been successful in Monte Carlo studies limited to the 2-group or k-group cases and/or they have been shown to have favorable distributional properties. Discussion of unincluded spread variables is necessarily limited.

The Properties of the Spread Variables

The discussion here is limited to the two-way ($I \times J$) design with population variances σ^2_{ij} and sample sizes n_{ij}. Because each of the spread variables is to be used as a dependent

This research was sponsored by Public Health Service Training Grant MH-08258 from the National Institute of Mental Health. The author thanks Mark I. Appelbaum, Elliot M. Cramer, and Scott E. Maxwell for their helpful criticisms of this paper. An earlier version of this work was presented at the Annual Meeting of the Psychometric Society, Murray Hill, New Jersey, April, 1976.

Requests for reprints should be sent to Ralph G. O'Brien, Department of Psychology, Gilmer Hall, University of Virginia, Charlottesville, Virginia, 22901.

variable in a traditional ANOVA, it is necessary to investigate the properties of these spread variables—especially their expected value, variance, intraclass correlation and distributional form. Once it is known how these properties are affected by changes in σ_{ij}^2, n_{ij}, and the distributional form of the raw data, then the existing theory of the least squares ANOVA can be used to infer how these variables might perform as ANOVA dependent measures. These theoretical inferences are limited, however, by the fact that these properties can only be investigated individually rather than jointly. For example, the effect of intraclass correlations among the spread observations is assessed as if the other properties conformed to the regular ANOVA assumptions of normality and homoscedasticity. The consequences of joint ANOVA assumption violations have received so little attention in the literature that such considerations are not feasible for the present problem. Therefore, the development below assumes that there are no "crossover interaction" type effects among the particular violations that are encountered. A large Monte Carlo study (described later) is used to provide some check on these inferences. The ANOVA theory is reviewed as needed.

The Log s^2 Variable

The log s^2 (Box-Scheffé) variable is formed by partitioning the raw observations into mutually exclusive subgroups within each cell and then taking the logarithm (any base) of the unbiased sample variance of each subgroup. The log s^2 observations are independent given that the raw observations are independent. Monte Carlo work with balanced one-way designs has demonstrated that this test produces reasonable Type I error rates, $P[EI]$, under a variety of parent distributions, but lacks power [Levy, 1975; Games, Winkler, & Probert, 1972; Layard, 1973; Miller, 1968].

Using a Taylor series expansion to approximate the mean and variance of a function of random variables [Kendall & Stuart, 1969, p. 231], it can be shown that

$$\mathcal{E}[\log s_{ijk}^2] = \log \sigma_{ij}^2 - (m_{ijk} - 1)^{-1} - 2\gamma m_{ijk}^{-1} + o(m_{ijk}^{-1}) \tag{1}$$

$$\mathrm{Var}[\log s_{ijk}^2] = 2(m_{ijk} - 1)^{-1} + \gamma m_{ijk}^{-1} + o(m_{ijk}^{-1}) \tag{2}$$

where m_{ijk} is the subset size, γ is the standardized kurtosis of the parent distribution, and $o(m_{ijk}^{-1})$ is a remainder term of order less than m_{ijk}^{-1}.

If all subset sizes are the same in all cells, artificial differences in $\mathcal{E}[\log s_{ijk}^2]$ and $\mathrm{Var}[\log s_{ijk}^2]$ are not introduced. Unequal subset sizes can produce artificial differences among the cell means for $\log s_{ijk}^2$ and also heterogeneity of variance: Both can affect $P[EI]$. While some authors have recognized the problems with $\mathrm{Var}[\log s_{ijk}^2]$, the bias problem has been largely ignored. Gartside [1972] studied a weighted type analysis for unequal m_{ijk}, but his results indicated that this test is even less powerful and may also produce larger $P[EI]$ with increases in the number of cells—a potential problem with even moderately sized factorial designs.

For equal subset sizes ($m_{ijk} = m$) in balanced designs, Martin [1976] thoroughly reviewed the problems involved in selecting the value of m that properly balances the tradeoff between $\mathrm{Var}[\log s_{ijk}^2]$ and the number of spread observations produced. He recommended finding an m that is close to the square root of n and, hopefully, that is an even divisor of n.

Finally, the most troubling aspect of this procedure is that the test does not yield unique results for a given set of data because of the arbitrariness of the subgroupings.

The $\tilde{z}^2$ Variable

Levene [1960] proposed the spread variable

$$z_{ijk}^2 = (y_{ijk} - \bar{y}_{ij})^2. \tag{3}$$

Since

$$\mathcal{E}[z_{ijk}^2] = (n_{ij} - 1)\frac{\sigma_{ij}^2}{n_{ij}} \quad , \tag{4}$$

however, we shall consider the spread variable

$$\tilde{z}_{ijk}^2 = \frac{n_{ij}(y_{ijk} - \bar{y}_{ij})^2}{n_{ij} - 1} \tag{5}$$

which has an expected value of σ_{ij}^2. In fact, the sample mean of the $\tilde{z}_{ijk}^2$,

$$\sum_k \frac{z_{ijk}^2}{n_{ij}} = \sum_k \frac{(y_{ijk} - \bar{y}_{ij})^2}{n_{ij} - 1} = s_{ij}^2 \tag{6}$$

which is an excellent property since the between group effects are then computed directly from the unbiased sample variances.

Levene gave the variance and intraclass correlation of z_{ijk}^2, and these results can be modified to give

$$\mathrm{Var}[\tilde{z}_{ijk}^2] = \sigma_{ij}^4\left[2 + \frac{(n_{ij}^2 - 3n_{ij} + 3)\gamma}{n_{ij}^2 - n_{ij}}\right] \tag{7}$$

$$\rho[\tilde{z}_{ijk}^2] = \frac{2n_{ij} + (2n_{ij} - 3)\gamma}{2n_{ij}(n_{ij} - 1)^2 + (n_{ij} - 1)(n_{ij}^2 - 3n_{ij} + 3)\gamma} \quad . \tag{8}$$

Values for the intraclass correlation and variance of $\tilde{z}_{ijk}^2$ have been computed for various γ and n_{ij} and are given in Tables 1 and 2.

The patterns of variances suggest that unbalanced designs with platykurtic ($\gamma < 0$) parent distributions may tend to exhibit larger $P[EI]$ relative to balanced designs, because the larger cells will have the smaller variances [Glass, Peckham, & Sanders, 1972]. The opposite effect occurs for leptokurtic distributions.

Applying results given by Walsh [1947] and Basu, Odell, and Lewis [1974], it can be shown that if a dependent variable, say v, satisfies all the usual ANOVA assumptions except that there exists an intraclass correlation, ρ, then

$$\mathcal{E}\left[\sum_{k=1}^{n_{ij}} \frac{(v_{ijk} - \bar{v}_{ij})^2}{(n_{ij} - 1)n_{ij}}\right] = \frac{(1 - \rho)\mathrm{Var}[\bar{v}_{ij}]}{1 - \rho + \rho n_{ij}} \quad . \tag{9}$$

Thus, for $\rho > 0$, this usual estimate of the variance of the mean will tend to be too small regardless of the cell size. If H_0 is true and the design is balanced, the usual F-statistic, $F(\rho) = MS_H/MS_{WG}$, still has an F-distribution except for a scalar multiple adjustment, i.e.

$$\frac{(1 - \rho)F(\rho)}{1 - \rho + \rho n} \sim F(df_H, df_{WG}). \tag{10}$$

If $F(df_H, df_{WG})$ is used as the sampling distribution of $F(\rho)$, then $P[EI] > \alpha$ for $\rho > 0$. If $\rho < 0$, the variance of the mean tends to be overestimated and $P[EI] < \alpha$.

Using (10), it is straightforward to assess the effect of the intraclass correlation of $\tilde{z}^2$ on the balanced ANOVA. $P[EI]$ results have been computed for the A effect test in the 2×2 and 4×3 designs with various γ and n (Table 3). These results demonstrate the tendency for $\rho[\tilde{z}^2]$ to moderately inflate $P[EI]$ for $n \leq 12$. For example, with $n = 8$ and $\gamma = 6$ the $P[EI]$ values are .084 and .118 for the 2×2 and 4×3 designs respectively. For $n = 16$, however, these values sharply decrease to .065 and .073. It can also be seen that the larger design is considerably more affected. Although these calculations are based specifically on balanced designs, the intraclass correlation should similarly affect unbalanced designs: Equation (9) clearly shows that (under H_0) $\mathcal{E}[MS_{WG}]$ will be generally less than $\mathcal{E}[MS_H]$.

TABLE 1

Intraclass Correlation of Spread Variables as a Function of Cell Size and Parent Distribution.

Parent	Cell Size 4	8	12	16	20	24
Uniform (γ=-1.2)						
$\tilde{z}^2$	+.043	+.001	-.001	-.001	-.001	-.001
q	-.135	-.034	-.015	-.009	-.006	-.004
p	-.126	-.060	-.028	-.016	-.010	-.006
W50	-.044	-.035	-.020	-.013	-.010	-.007
Normal (γ=0)						
$\tilde{z}^2$	.111	.020	.008	.004	.003	.002
q	-.090	-.018	-.008	-.004	-.003	-.002
p	-.101	-.023	-.013	-.007	-.006	-.003
W50	-.012	-.010	-.006	-.005	-.005	-.003
Expontl (γ=6)						
$\tilde{z}^2$	.192	.036	.015	.008	.005	.003
q	-.034	-.006	-.002	-.001	-.001	-.000
p	-.003	+.022	+.015	+.010	+.007	+.004
W50	+.012	-.004	-.005	-.002	-.001	+.000

Note: Values for p and W50 are Monte Carlo estimates based on 1000 trials.

The kurtosis of the $\tilde{z}^2$ variable was estimated for various values of γ and n_{ij} using a simple Monte Carlo analysis. (See O'Brien, 1975, p. 23 for a description of the procedure.) These results are exemplified by the $n_{ij} = 12$ cases. For a uniform parent ($\gamma = -1.2$) the kurtosis of $\tilde{z}^2_{ijk}$ was estimated at 1.0; for the normal ($\gamma = 0$) it was 12.2; and the exponential ($\gamma = 6$) produced an extremely large value of 139.2. It should be noted that with a normal parent, $(y - \mu)^2/\sigma^2$ is a chi-square random variable which has a known kurtosis of 12. Thus, the kurtosis of $\tilde{z}^2_{ijk}$ is asymptotically 12 with a normal parent.

Box and Andersen [1955] showed that the usual one-way ANOVA F-statistic is distributed approximately as $F[\delta(df_H), \delta(df_{WG})]$ where

$$\delta = 1 + \frac{\gamma}{N} + O(N^{-2}). \tag{11}$$

Here $N = \sum n_{ij}$ and $O(N^{-2})$ is a term of order N^{-2}. When γ is large $P[EI]$ will decrease. Thus, the high kurtosis of $\tilde{z}^2$ should tend to counterbalance the effects of intraclass correlation. As the size of the design increases, however, the effects from the kurtosis of $\tilde{z}^2$ will diminish and the effects from intraclass correlation will dominate, thereby producing inflated $P[EI]$ values.

The p Variable

Mosteller and Tukey [1968] and Miller [1968] introduced the jackknife technique for use in the study of variances. Miller used the jackknife pseudovalues

$$p_{ijk} = n_{ij} \log s^2_{ij} - (n_{ij} - 1) \log s^2_{ij-k} \tag{12}$$

TABLE 2

Variance of Spread Variables as a Function of Cell Size and Parent Distribution. $\sigma^2_y = 4$.

	Cell Size					
Parent	4	8	12	16	20	24
Uniform (γ=-1.2)						
$\tilde{z}^2$	20.8	17.3	15.8	15.1	14.7	14.4
q	39.5	22.7	18.9	17.3	16.3	15.7
p	7.3	2.2	1.5	1.3	1.2	1.1
W50	1.3	1.4	1.3	1.2	1.2	1.2
Normal (γ=0)						
$\tilde{z}^2$	32.0	32.0	32.0	32.0	32.0	32.0
q	58.7	41.9	38.1	36.5	35.5	34.9
p	6.9	3.5	2.8	2.6	2.5	2.4
W50	1.4	1.6	1.6	1.5	1.5	1.5
Expontl (γ=6)						
$\tilde{z}^2$	88.0	105.7	112.7	115.9	118.7	120.2
q	154.7	137.9	134.1	132.5	131.5	130.1
p	12.7	9.2	8.4	7.9	8.0	7.9
W50	1.9	2.4	2.4	2.4	2.4	2.4

Note: Values for p and W50 are Monte Carlo estimates based on 1000 trials.

TABLE 3

Effect of Intraclass Correlation of $\tilde{z}^2$ and q Variables on the Test of A in Two-Way Balanced Designs.

	Cell Size					
	4	8	12	16	20	24
2 x 2 Design						
Uniform (γ=-1.2)						
$\tilde{z}^2$	.068	.052	.050	.050	.050	.050
q	.011	.025	.032	.036	.038	.039
Normal (γ=0)						
$\tilde{z}^2$	.100	.069	.063	.059	.057	.055
q	.021	.036	.041	.043	.044	.046
Expontl (γ=6)						
$\tilde{z}^2$	.178	.084	.070	.065	.062	.059
q	.038	.046	.047	.048	.049	.049
4 x 3 Design						
Uniform (γ=-1.2)						
$\tilde{z}^2$	.081	.053	.050	.050	.050	.050
q	.003	.016	.024	.029	.032	.034
Normal (γ=0)						
$\tilde{z}^2$	.144	.080	.071	.064	.060	.058
q	.011	.029	.036	.039	.041	.044
Expontl (γ=6)						
$\tilde{z}^2$	.242	.118	.083	.073	.069	.063
q	.032	.043	.046	.047	.048	.049

as spread variables, where s_{ij}^2 is the unbiased sample variance and s_{ij-k}^2 is the unbiased sample variance when the k^{th} observation is deleted from consideration. The sample mean of the p_{ijk} in the i, j^{th} cell is called the jackknifed estimate of log σ_{ij}^2 while the estimator that is jackknifed is log s_{ij}^2. It is a well-known property of the jackknife technique [Gray & Schacany, 1972] that since

$$\mathcal{E}[\log s_{ij}^2] = \log \sigma_{ij}^2 - (n_{ij} - 1)^{-1} - 2\gamma n_{ij}^{-1} + o(n_{ij}^{-1}), \tag{13}$$

it follows that

(14)
$$\mathcal{E}[p_{ijk}] = \log \sigma_{ij}^2 + o(n_{ij}^{-1}).$$

Jackknifing was developed to reduce estimation bias, which it does in this case. However, p is still biased to order less than n_{ij}^{-1}, thus its use in unbalanced designs may increase $P[EI]$ due to the production of artificial differences among the expected values of the spread means.

The intraclass correlation of the p variables was investigated through Monte Carlo analysis (Table 1). Those values are negative for both the uniform and normal parents, but are generally positive ($n_{ij} > 8$) for the exponential parent. This result explains the Monte Carlo results in the literature which consistently show conservative $P[EI]$ for uniform and normal distributions, but inflated $P[EI]$ for the exponential [Miller, 1968; Brown & Forsythe, 1974; Layard, 1973].

The variance properties of p were also investigated using Monte Carlo methods and those results are presented in Table 2. The variance decreases as n increases. Thus, both the heterogeneity of variance and biasedness of p should tend to inflate $P[EI]$ for unequal n cases. The limited Monte Carlo work on p with unequal n done by Brown and Forsythe [1974] supports this argument since the unbalanced two-group designs exhibited markedly increased empirical test sizes for nonnormal parent distributions.

Monte Carlo estimates of the kurtosis of p showed that for the uniform parent, $\gamma[p]$ decreases as n_{ij} increases (5.69 for $n_{ij} = 12$ and 1.02 for $n_{ij} = 24$). For the normal, $\gamma[p]$ is stable (averaged about 11). For the exponential, the $\gamma[p]$ increases as n_{ij} increases (31.12 for $n_{ij} = 12$ and 51.82 for $n_{ij} = 24$). Thus, the kurtosis of p will, to some extent, tend to make the ANOVA tests more conservative than otherwise. For the exponential parent, however, intraclass correlation effects should overwhelm this tendency.

The q Variable

An alternative jackknife pseudovalue

(15)
$$q_{ijk} = n_{ij}s_{ij}^2 - (n_{ij} - 1)s_{ij-k}^2$$

has been basically ignored in the literature although it was briefly considered by Miller [1968, 1974]. Because

(16)
$$\mathcal{E}[q_{ijk}] = n_{ij}\sigma_{ij}^2 - (n_{ij} - 1)\sigma_{ij}^2 = \sigma_{ij}^2,$$

it may be more appropriate than p for designs with unequal n. It is also easier to deal with theoretically.

Using Equation 7 from Miller [1968], it follows that

(17)
$$q_{ijk} = \frac{(n_{ij}^2 - n_{ij} - 1)\tilde{z}_{ijk}^2 - \sum_{k' \neq k} \tilde{z}_{ijk'}^2}{n_{ij}(n_{ij} - 1)}$$

which yields some simple computational formulae,

(18)
$$q_{ijk} = \frac{n_{ij}(y_{ijk} - \bar{y}_{ij})^2 - s_{ij}^2}{n_{ij} - 2} = \frac{(n_{ij} - 1)\tilde{z}_{ijk}^2 - s_{ij}^2}{n_{ij} - 2} .$$

It is easily demonstrated that

(19)
$$\sum_{k=1}^{n_{ij}} \frac{q_{ijk}}{n_{ij}} = s_{ij}^2;$$

thus, like $\tilde{z}^2$, the estimates of between group effects can be found directly from the within-cell variances. Therefore, the between-groups sums of squares based upon $\tilde{z}^2$ and q are

identical. As suggested by an associate editor, the within-group sum of squares using q,

$$SSW(q) = \sum_{i,j} \sum_{k=1}^{n_{ij}} (q_{ijk} - s_{ij}^2)^2 = \sum_{i,j} \sum_{k=1}^{n_{ij}} \frac{(n_{ij} - 1)^2(\tilde{z}_{ijk}^2 - s_{ij}^2)^2}{(n_{ij} - 2)^2} \quad . \tag{20}$$

Thus $SSW(q)$ will always be greater than $SSW(\tilde{z}^2)$. If the design is balanced,

$$F(\tilde{z}^2) = \frac{(n-1)^2 F(q)}{(n-2)^2} \quad . \tag{21}$$

Thus, the $P[EI]$ and power will be smaller for q than for $\tilde{z}^2$.

Since q_{ijk} is a linear combination of the $\tilde{z}_{ijk}^2$'s, the variance and intraclass correlation of q_{ijk} can be found. Temporarily dropping the i, j subscripts,

$$\mathrm{Var}[q_k] = \frac{\{(n^4 - 2n^3 - n^2 + 3n) - (2n^3 - 5n^2 + 3n)\rho[\tilde{z}^2]\}\mathrm{Var}[\tilde{z}^2]}{n^2(n-2)^2} \tag{22}$$

$$\rho[q_k] = \frac{(3n - 2n^2) + \{(n-1)^4 + (n-1)\}\,\rho[\tilde{z}^2]}{(n^4 - 2n^3 - n^2 + 3n) - (2n^3 - 5n^2 + 3n)\rho[\tilde{z}^2]} . \tag{23}$$

These equations have been evaluated for various values of n_{ij} and γ and are presented in Tables 1 and 2. Surprisingly, the intraclass correlation is negative throughout the range studied, thus making the ANOVA test basically conservative. Like $\rho[\tilde{z}^2]$, $\rho[q]$ is of order n_{ij}^{-2} and tends to increase as γ increases. Because $\rho[q]$ is negative, however, the larger γ will tend to make $P[EI]$ closer to α. Using the same logic that was applied to $\rho[\tilde{z}^2]$, the effects of $\rho[q]$ were calculated for the 2×2 and 4×3 balanced designs (Table 3). It can be easily seen that the $P[EI]$ is very conservative for small n and that the larger design is more affected. Using (9) again, it can be seen that this basic conservativeness should apply to unbalanced designs as well. However, the variance of q_{ijk} decreases as n_{ij} increases, which should tend to increase $P[EI]$ for unbalanced designs.

Using (18), it is immediately evident that

$$\lim_{n_{ij} \to \infty} (q_{ijk} - \tilde{z}_{ijk}^2) = 0 . \tag{24}$$

Kendall and Stuart [1969, p. 115] provide the "Second Limit Theorem" which can be used to show that the moments of q_{ijk} and $\tilde{z}_{ijk}^2$ are asymptotically equivalent. Thus in the limit, each variable has a variance of $\sigma_{ij}^4(2 + \gamma)$. Also, since $\tilde{z}_{ijk}^2$ has high kurtosis, so should q_{ijk}. Monte Carlo simulation of $\gamma[q]$ revealed it to be of the same magnitude as $\gamma[\tilde{z}^2]$ for $n_{ij} \leq 24$. Again, this ANOVA violation probably further reduces $P[EI]$ for designs with small N.

Miller [1974] criticized the q variable because its use in traditional two-tailed t-distribution confidence intervals may at times lead to intervals that include negative values for σ^2. Of course, the interval may simply be truncated openly at zero ($0 < \sigma^2$). This will not change the confidence level of the interval since only improper values of σ^2 are removed—if the larger interval contains the population σ^2, then so will the truncated interval.

The W50 Variable

Brown and Forsythe [1974] investigated

$$W50_{ijk} = |y_{ijk} - Md_{ij}| \tag{25}$$

where Md_{ij} is the median of the i, j^{th} cell. They found that it possesses excellent robustness in the balanced and unbalanced two-group designs. The sample mean of the $W50$ variable is a common measure of spread, the average absolute deviation [Hays, 1973, p. 243].

Brown and Forsythe suggested another variable, *W10*, the absolute deviation about the 10% trimmed mean. Their work suggests that while *W10* may be more powerful than *W50*, it may also be less robust. Only *W50* is examined here.

Little theoretical work has been done regarding the suitability of *W50* as an ANOVA variable. Because of the natural difficulties (absolute values and medians) of dealing with *W50* analytically, it was necessary to use Monte Carlo estimation to investigate its properties.

The expected value generally increases as a function of n and decreases as a function of γ. For example, with $\sigma_{ij}^2 = 4$ expected value estimates for $n_{ij} = 4$ were 1.38, 1.34, and 1.15 for uniform ($\gamma = -1.2$), normal ($\gamma = 0$) and exponential ($\gamma = 6$) parents respectively. Corresponding values for $n_{ij} = 24$ were 1.65, 1.54, and 1.34. Thus, unbalanced designs produce artificial differences in the spread means and therefore may give higher $P[EI]$ than balanced designs. The variances of *W50* (Table 1) are generally stable over different n_{ij}.

O'Brien [1975] suggested a correction factor for *W50*,

$$\tilde{W}_{ijk} = \left[\frac{n_{ij}}{n_{ij} - 1}\right]^{1/2} W50_{ijk} \cdot \tag{26}$$

While the expected values of this variable are stable, it has smaller variances for cells with larger n_{ij}, which will also inflate $P[EI]$. Of course, $\tilde{W}$ has the same intraclass correlation and kurtosis as *W50*. The correction factor $\tilde{W}$ was not included in the Monte Carlo study, although any *W50* results for equal n cases apply directly to $\tilde{W}$.

The intraclass correlation (Table 1) is generally negative indicating that *W50*, like q, will produce basically conservative ANOVAs. The kurtosis of *W50* for $n_{ij} = 12$ was estimated as $-.5$ for the uniform parent, 1.0 for the normal, and 12.7 for the exponential; and was quite stable across different n_{ij}.

It should be noted that all these simulations used even sample sizes. Martin [1976] suggested that *W50* is much more conservative (and hence less powerful) with odd n_{ij}'s: The resulting *W50* = 0.0 observations can radically increase the variance of *W50*. One obvious solution is to randomly delete an observation, so that n_{ij} is even.

Other Comments

The theory outlined above was used to evaluate the appropriateness of various spread variables by focusing nearly all attention on the within cell properties of those variables: the expected value, variance, intraclass correlation and kurtosis and their dependencies on the cell size and parent distribution. Little attention has been paid to the statistical model as a whole.

Since ANOVA is an additive model, the use of log σ_{ij}^2 type dependent variables effectively represents a multiplicative model for σ_{ij}^2. Thus log s^2 and p are variables that conform to a multiplicative model for the cell variances, while $\tilde{z}^2$ and q conform to an additive model for the cell variances. The use of *W50* produces an additive model for the average absolute deviation. The distinction between additive and multiplicative models has not been stressed in past research on tests for homogeneity of variance, because that work has only focused on the two-group and k-group (one-way) designs. In the case of two (or more) factors, the distinction becomes important if the researcher wishes to parsimoniously describe and test the groups' variances in terms of main effects and interactions. For example, if the model underlying the variances is multiplicative with no interactions, i.e.,

$$\sigma_{ij}^2 = \sigma^2 a_i b_j; \quad i = 1 \text{ to } I; \quad j = 1 \text{ to } J, \tag{27}$$

then the use of an additive statistical model will detect interaction effects that reflect the multiplicative model. A similar phenomenon will occur if data from an underlying

additive model is analyzed with a multiplicative statistical model. For those researchers who are untroubled by interactions, the use of an additive model may be preferable since it is most similar to the analysis of variance. It is important that researchers understand that the three sets of variables (log s^2, p; $\tilde{z}^2$, q; and $W50$) imply different statistical models.

Another point needs to be brought into focus. For every spread variable, greater parent kurtosis results in greater spread variance. This is the primary reason for the robustness of these tests. For example, it is well-known [Scheffé, 1959, p. 83] that

$$\text{Var}[s_{ij}^2] = \sigma_{ij}^4\left[\frac{2}{n_{ij}-1} + \frac{\gamma}{n_{ij}}\right]. \tag{28}$$

Normal theory tests implicitly use $\gamma = 0$ and are not robust. If appropriate spread variables are used in an ANOVA, Var[s^2] is estimated from the data and will "automatically" be properly sensitive to γ. For example, the asymptotic variance of both $\tilde{z}_{ijk}^2$ and q_{ijk} is $\sigma_{ij}^4(2+\gamma)$. Because both their intraclass correlations are asymptotically 0, the variance of the spread means, $\bar{\tilde{z}}_{ij}^2$ and $\bar{q}_{ij}$, will be (in the limit) properly estimated by $\sigma_{ij}^4(2+\gamma)/n_{ij}$. This relationship between parent kurtosis and spread variance also implies that even a "perfect" spread variable (unbiased, no intraclass correlation, stable variance over n_{ij}, normally distributed) will become less powerful as the parent kurtosis increases.

The Monte Carlo Simulation

In order to supplement and test the theoretical development, an extensive Monte Carlo simulation of these tests was conducted. The results of this study demonstrate and validate the conclusions made above. In addition, the powers of the tests are estimated and compared.

In addition to using the five ANOVA variables (log s^2, p, $\tilde{z}^2$, q, $W50$), Zelen's [1959, 1960] likelihood ratio tests of A, B, and AB (denoted M_1, M_2 and M_3 by Zelen) were also included in order to assess what sacrifices in power may result from using the more robust procedures. Since Bartlett's test has been customarily used in past research as a standard of power for the k-group homogeneity of variance tests, the "two-way Bartlett test" developed by Zelen seemed to be the best standard for this study. Only balanced designs could be considered for Zelen's technique, however, because it has not yet been extended to unbalanced designs. The Z-variance test was not studied, since it has already been shown to have unacceptable robustness [Overall & Woodward, 1974; Levy, 1975].

The Monte Carlo study was designed to estimate the probability of rejecting the usual A, B, and AB hypotheses for a variety of two-way factorial designs. In essence, a factorial Monte Carlo experiment was designed to study two-way analysis of spread designs. These two-way designs differed with respect to the dimension of the design (2×2 or 4×3), the type of parent distribution (uniform, normal, exponential), the average cell sizes (12, 24), the degree of imbalance of the cell sizes (balanced, "moderately" unbalanced, "severely" unbalanced), the type of underlying model for the cell variances (null, one main effect, two additive main effects, two multiplicative main effects, interaction effect). These nonnull models were varied with respect to their degree of effect ("low" and "high") which were operationally defined in order to obtain a suitable range for the power curves. Each of these factors was "crossed" with the other factors, except that only balanced designs were used for the nonnull underlying models. In other words, power was only studied for the balanced designs. For each unique combination of these factors, an estimate of the rejection rate was made using 1000 trials, except that in the cases with no underlying differences among the cell variances (null) 2000 trials were used. With 2000 trials, the standard error of the Type I error rate estimate is approximately .0048 when the true α-level is .05. Standard errors for true powers of .1 (or .9), .3(or .7) and .5 for 1000 trials are .0095, .0145, and .0158 respectively.

Initially, each of the tests was conducted using the nominal .05 critical values. In

addition, however, the experiment was repeated using empirical .05 critical values that were obtained by selecting the 100th highest test statistics from each of the null model designs (100/2000 = .05). Such a standardization allows for power comparisons that are not contaminated by the initial discrepancies in $P[EI]$. Techniques with good "empirical power" may provide a good basis for an adjusted test.

Random numbers were generated using the PDP-11 congruential function, RANDU, which was imbedded in a Marsaglia table scheme [Marsaglia & Bray, 1968]. This $U(0, 1)$ distribution was initially transformed to either a uniform ($\gamma = -1.2$), normal ($\gamma = 0$), or exponential ($\gamma = 6$) distribution with locations and scales standardized to $\mathcal{E}[X] = 0$ and $\text{Var}[X] = 1$. In order to obtain the standardized uniform parent, the linear transformation, $X = (U - .5)12^{1/2}$, was used. The Box-Muller normalization was used to generate the $N(0, 1)$ distribution. The transformation $X = -\ln(U) - 1$ produced the (standardized) exponential distribution, with density, $f(X) = \exp(-X - 1)$, $-1 \leq X \leq \infty$. The cell variances were then controlled through the transformation $Y = \sigma_{ij}X$. The unbalanced ANOVA tests were computed using only the "eliminating" tests ($AB \mid A, B$; $A \mid B$; and $B \mid A$) as described by Appelbaum and Cramer [1974].

Results for Type I Error Rates

It would not be practical to detail all of the results of the experiment here, but it is worthwhile to present the results for a demonstrative case. The null model results for the 4×3 design with average cell sizes $\bar{n} = 12$, are given in Table 4. The cell sizes ($n_{11}, n_{12}, \ldots, n_{43}$) were 8, 8, 16, 16, 12, 12, 12, 12, 16, 16, 8, 8 for the moderately unbalanced design and 4, 4, 20, 20, 12, 12, 12, 12, 20, 20, 4, 4 for the severely unbalanced design. This arrangement was selected so that the effects of biased variables would be manifested in the test of interaction.

The log s^2 variable produced excellent empirical test sizes with one unexplained exception. Of the 108 $P[EI]$ estimates generated from the entire study, only seven exceeded .060 (.076 maximum) and none were less than .040. Of these seven high values, however, five were from tests of severely unbalanced designs with the uniform parent. Equal subset sizes of $m = 4$ were used for all cases.

The $\tilde{z}^2$ variable produced basically inflated test sizes, as expected. The estimates generally increased as a function of γ due to the increasing intraclass correlation. The counterbalancing effect of the kurtosis of $\tilde{z}^2$ seems to have broken down in the 4×3 designs, although the 2×2 results showed a more stable pattern. There was some difficulty with unbalanced designs due perhaps to the heterogeneity of variance of $\tilde{z}^2$ and the high intraclass correlations that exist in the smaller cells.

The p variable also performed as expected. In agreement with its intraclass correlation, the test was conservative for uniform and normal parent populations and inflated for the exponential which gave quite high levels (.089 to .112 for the case presented here). This test was also unstable for unbalanced designs as shown by the high $P[EI]$ for the AB tests. For the normal parent, this case produced A, B, and AB test sizes of .052, .051, and .120 respectively. The exponential parent produced an AB test size of .174 in the severely unbalanced design.

The q variable produced the conservative test sizes that were expected from its negative intraclass correlations. This test was generally well-behaved in unbalanced designs. There were some increases for AB tests in severely unbalanced 4×3 designs, although this was only troublesome with the exponential parent (.097, for $\bar{n} = 12$ and .090 for $\bar{n} = 12$).

The W50 variable performed conservatively due to its negative intraclass correlation. Unbalanced designs did not substantially effect this test, thus the dependency between the expected value of $W50$ and n_{ij} had little effect.

TABLE 4

Empirical Probability of Type I Error Using the Nominal .05 Critical Values for the 4 x 3 Design with Average Cell Size of $\tilde{n}$=12

	Equal n			Moderate Imbal			Severe Imbal		
	A	B	AB	A	B	AB	A	B	AB
Uniform parent									
Zelen	.004	.002	.002						
log s^2	.043	.049	.042	.043	.047	.054	.052	.060	.076
p	.015	.016	.018	.021	.020	.022	.039	.067	.202
$\tilde{z}^2$	.048	.048	.050	.055	.047	.060	.069	.065	.105
q	.023	.026	.022	.030	.029	.022	.027	.036	.043
W50	.017	.017	.012	.028	.021	.020	.020	.029	.041
Normal parent									
Zelen	.051	.051	.040						
log s^2	.041	.049	.041	.045	.052	.051	.047	.042	.060
p	.041	.035	.027	.034	.045	.041	.052	.051	.120
$\tilde{z}^2$	.068	.056	.055	.052	.062	.079	.069	.073	.110
q	.036	.030	.027	.024	.034	.034	.032	.035	.052
W50	.032	.031	.026	.025	.038	.032	.032	.036	.033
Expontl parent									
Zelen	.437	.357	.649						
log s^2	.054	.053	.049	.048	.045	.052	.054	.047	.058
p	.090	.089	.112	.097	.076	.116	.101	.089	.174
$\tilde{z}^2$	.080	.064	.092	.079	.069	.116	.097	.098	.151
q	.044	.038	.045	.043	.037	.052	.045	.045	.097
W50	.047	.045	.045	.038	.036	.042	.053	.045	.052

The Zelen procedure was extremely nonrobust and showed the pattern typical of Bartlett's test.

Results Concerning Power

In order to present the general flavor of the results from the nonnull cases, the 4 × 3 design with the underlying model

$$\sigma_{ij}^2 = \sigma^2 a_i b_j; \quad \prod_{i=1}^{4} a_i = \prod_{j=1}^{3} b_j = 1; \quad a_i, b_j > 0 \tag{29}$$

was selected as a fair representative of the results from the entire study. The parameters for a_i and b_j were

low effect: $a = [7/10, 9/10, 10/9, 10/7]$ $b = [4/5, 1, 5/4]$
high effect: $a = [6/10, 9/10, 10/9, 10/6]$ $b = [4/5, 1, 5/4]$

with $\sigma^2 = 4$. Notice that the A effect was the only effect that was changing.

The powers using nominal critical values are given in Tables 5 and 6. Power for all the ANOVA tests decreased as a function of γ because of the corresponding increases in the variances of the spread variables.

TABLE 5

Empirical Probability of Rejection Using the Nominal .05 Critical Values for the 4 x 3 Design with Equal Cell Sizes of n=12.

	Null			Low Effect			High Effect		
	A	B	AB	A	B	AB	A	B	AB
Uniform parent									
Zelen	.004	.002	.002	.193	.092	.002	.638	.099	.002
log s^2	.043	.049	.042	.277	.190	.052	.285	.091	.007
p	.015	.016	.018	.430	.256	.019	.850	.260	.021
$\tilde{z}^2$	.048	.048	.050	.636	.414	.076	.920	.406	.090
q	.023	.026	.022	.508	.318	.048	.864	.299	.046
W50	.017	.017	.012	.299	.195	.016	.667	.179	.017
Normal parent									
Zelen	.051	.051	.040	.344	.232	.048	.657	.214	.060
log s^2	.041	.049	.041	.220	.150	.051	.204	.084	.004
p	.041	.035	.027	.263	.176	.032	.540	.152	.038
$\tilde{z}^2$	.068	.056	.055	.370	.251	.087	.638	.213	.100
q	.036	.030	.027	.257	.174	.039	.506	.152	.058
W50	.032	.031	.026	.239	.154	.032	.511	.146	.039
Expontl parent									
Zelen	.437	.357	.649	.627	.469	.606	.774	.467	.623
log s^2	.054	.053	.049	.124	.092	.041	.147	.073	.009
p	.090	.089	.112	.228	.139	.101	.297	.152	.107
$\tilde{z}^2$	.080	.064	.092	.181	.121	.091	.294	.131	.107
q	.044	.038	.045	.107	.068	.045	.195	.081	.061
W50	.047	.045	.045	.166	.095	.040	.297	.105	.044

TABLE 6

Empirical Probability of Rejection Using the Nominal .05 Critical Values for the 4 x 3 Design with Equal Cell Sizes of n=24.

	Null			Low Effect			High Effect		
	A	B	AB	A	B	AB	A	B	AB
Uniform parent									
Zelen	.001	.001	.001	.722	.328	.000	.992	.353	.000
log s^2	.052	.046	.050	.580	.345	.044	.580	.163	.002
p	.024	.022	.022	.968	.778	.024	.999	.760	.017
$\tilde{z}^2$	.047	.040	.051	.973	.786	.084	.999	.759	.131
q	.035	.030	.034	.965	.735	.057	.999	.718	.096
W50	.024	.023	.019	.798	.499	.019	.985	.501	.023
Normal parent									
Zelen	.049	.056	.062	.671	.465	.051	.953	.447	.057
log s^2	.043	.044	.056	.409	.281	.049	.452	.180	.003
p	.036	.041	.047	.619	.415	.043	.935	.415	.050
$\tilde{z}^2$	.050	.059	.064	.658	.447	.094	.947	.430	.094
q	.037	.043	.042	.616	.406	.058	.927	.384	.069
W50	.039	.042	.041	.586	.377	.049	.908	.369	.045
Expontl parent									
Zelen	.477	.387	.697	.790	.591	.697	.898	.594	.678
log s^2	.047	.048	.052	.254	.148	.054	.278	.104	.006
p	.073	.073	.104	.306	.187	.096	.493	.186	.106
$\tilde{z}^2$	.056	.058	.073	.253	.160	.076	.439	.141	.086
q	.041	.043	.049	.211	.135	.060	.382	.111	.065
W50	.043	.045	.054	.333	.197	.046	.592	.176	.068

For the uniform parent, the p and $\tilde{z}^2$ ANOVAs gave the highest power. The power of q was a close third in rank. Zelen's tests and the *W50* ANOVA were similar in power, ranking below q. The log s^2 ANOVA was much lower in power.

For the normal parent, the Zelen procedure was superior, although the (inflated) $\tilde{z}^2$ test was nearly as powerful. A group consisting of the p, q, and *W50* ANOVAs had less power, but were considerably more powerful than the log s^2 ANOVA.

For the exponential parent, the power of Zelen's procedure is deceptive due to the extreme test sizes. The most outstanding performance was that of the *W50* variable which

generally gave the highest power of the ANOVA procedures, even though its empirical test sizes were not inflated. The p and $\tilde{z}^2$ variables produced power near that of *W50*, but their $P[EI]$ values were quite high. The q and log s^2 ANOVAs were the lowest in power.

Because the results presented here were from the multiplicative underlying model, the $P[EI]$ values for the AB tests for the additive statistical models ($\tilde{z}^2$ and q) are expectedly greater than .05. Distributing the overall between-groups variance over main effects and interaction effects may have reduced the power of the main effects tests for $\tilde{z}^2$ and q to some degree. Over all underlying models, however, the relative powers of all ANOVA tests remained fairly stable.

With some minor exceptions, the ANOVA tests' power results based on the empirical .05 critical values were consistent with the nominal results. For the uniform parent, the p, $\tilde{z}^2$ and q ANOVAs produced the most power, although the Zelen procedure had similar power. The *W50* ANOVA trailed considerably, but it was still more powerful than the log s^2 ANOVA. For the normal parent, the Zelen test excelled, and the p, q, $\tilde{z}^2$, and *W50* ANOVAs were all quite similar and were all clearly more powerful than log s^2. With the exponential parent, the *W50* ANOVA was the most powerful test and was followed by Zelen's test. The power of the log s^2 variable was surprising in that it was greater than that of the p, q, or $\tilde{z}^2$ variables, especially with $n = 24$. Due to the fact that the F-statistics for $\tilde{z}^2$ and q are linearly related, their empirical powers were identical under all conditions.

Conclusion

It seems fair to conclude that the Monte Carlo results reasonably paralleled the expectations that were derived from the properties of the various spread variables. This fusion of empirical and theoretical studies enabled us to discover not only *what* happens when these spread variables are used in an ANOVA, but also *why* those things happen. Knowledge of the theoretical properties of these tests should provide a basis for their refinements.

Of the properties studied here, intraclass correlation seems to be the most important. Variables with positive ρ ($\tilde{z}^2$ and, at times, p) produce inflated $P[EI]$ and those with negative ρ (q, *W50*, and at times, p) produce conservative $P[EI]$. The influence of ρ on the behavior of these spread variables in ANOVA is strong and suggests that all candidates for spread variables should be examined closely with respect to intraclass correlation.

The other properties are also important. Because the expected values and variances are often dependent upon n_{ij}, designs with unequal n can behave differently than balanced designs. The kurtosis of $\tilde{z}^2$ and q can be very large, which might seriously reduce $P[EI]$ and power for designs with small total sample sizes.

Primarily because of their slightly negative intraclass correlation, q and *W50* emerge as the best spread variables examined in this study. The use of either $\tilde{z}^2$ and p will give inflated $P[EI]$. While the independence of the log s^2 variables produces reasonable $P[EI]$, this variable's lack of power and arbitrariness of subgroup formation make it undesirable.

The remaining attention will focus on q and *W50*. Of course, their usefulness depends on other factors besides $P[EI]$, i.e. power, ease of computations, and interpretability. When underlying assumptions are met, nonrobust procedures are often more powerful than robust procedures that do not "take advantage" of those assumptions. Here, for example, Zelen's test showed more power than all the ANOVA tests when the parent population was normal. Fortunately, the normal parent power of q and *W50* were not far behind the power of Zelen's test. Therefore, it is reasonable to use these robust procedures even in cases when normality can be "reasonably" assured. With other types of parent distributions, the q variable produces more power than *W50* for platykurtic ($\gamma < 0$) distributions, while *W50* is superior to all tests studied here when the distributions are quite leptokurtic. Thus some knowledge of the shape of the parent distribution will aid researchers in their choice of variables. Even cell sizes should be used for *W50*.

Both the q and *W50* variables are easy to compute, especially if one uses the computational formulae (18) for q. The abundance of ANOVA computer routines insures their usability. With respect to interpretability, many researchers will find q preferable to *W50*, because q produces an additive model for the most familiar measure of spread, the cell variances.

On balance, the q variable ANOVA can be recommended as a general tool, although *W50* should be used with extremely leptokurtic populations. (Remember not to use *W50* with odd numbers for cell sizes.) The use of either q or *W50* is a simple method to test hypotheses of homogeneity of spread in factorial designs. The resulting ANOVA tests are robust and relatively powerful.

REFERENCES

Appelbaum, M. I., & Cramer, E. M. Some problems in the nonorthogonal analysis of variance. *Psychological Bulletin,* 1974, *81,* 335–347.

Basu, J. P., Odell, P. L., & Lewis, T. O. The effects of intraclass correlation on certain significance tests when sampling from multivariate normal population. *Communications in Statistics,* 1974, *3,* 899–908.

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

Brown, M. B., & Forsythe, A. B. Robust tests for equality of variances. *Journal of the American Statistical Association,* 1974, *69,* 364–367.

Games, P. A., Winkler, H. R., & Probert, D. A. Robust tests for homogeneity of variance. *Educational and Psychological Measurement,* 1972, *32,* 887–909.

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

Glass, G. V., Peckham, P. D., & Sanders, J. R. Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research,* 1972, *42,* 237–288.

Gray, H. L., & Schucany, W. R. *The generalized jackknife statistic.* New York: Marcel Dekker, 1972.

Hays, W. L. *Statistics for the social sciences* (2nd ed.). New York: Holt, Rinehart, & Winston, 1973.

Kendall, M. G., & Stuart, A. *The advanced theory of statistics* (Vol. 1, 3rd ed.). London: Charles Griffin, 1969.

Layard, M. W. J. Robust large sample tests for homogeneity of variances. *Journal of the American Statistical Association,* 1973, *68,* 195–198.

Levene, H. Robust tests for the equality of variances. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann (Eds.), *Contributions to probability and statistics* Palo Alto: Stanford University Press, 1960.

Levy, K. J. An empirical comparison of the Z-variance and Box-Scheffé tests for homogeneity of variance. *Psychometrika,* 1975, *40,* 519–524.

Marsaglia, G., & Bray, T. A. One-line random number generators and their use in combinations. *Communications of the ACM,* 1968, *11,* 757–759.

Martin, C. G. Comment on Levy's "An empirical comparison of the Z-variance and Box-Scheffé tests for homogeneity of variance." *Psychometrika,* 1976, *41,* 551–556.

Miller, R. G., Jr. Jackknifing variances. *Annals of Mathematical Statistics,* 1968, *39,* 567–582.

Miller, R. G., Jr. The jackknife—a review. *Biometrika,* 1974. *61,* 1–15.

Mosteller, F., & Tukey, J. W. Data analysis, including statistics. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (Vol. 2, 2nd ed.). Reading, Mass.: Addison-Wesley, 1968.

O'Brien, R. G. Factorial designs for the analysis of spread. (Doctoral dissertation, University of North Carolina, 1975). *Dissertation Abstracts International,* 1976, *37,* 1328B. (University Microfilms No. 76-20,062)

Overall, J. E., & Woodward, J. A. A simple test for heterogeneity of variance in complex factorial designs. *Psychometrika,* 1974, *39,* 311–318.

Scheffé, H. A. *The analysis of variance.* New York: Wiley, 1959.

Zelen, M. Factorial experiments in life testing. *Technometrics,* 1959, *1,* 269–288.

Zelen, M. Analysis of two-factor classifications with respect to life tests. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann (Eds.), *Contributions to probability and statistics.* Palo Alto: Stanford University Press, 1960.

Walsh, J. E. Concerning the effect of intraclass correlation on certain significant tests. *Annals of Mathematical Statistics,* 1947, *18,* 88–96.

Manuscript received 6/28/76
First revision received 2/11/77
Second revision received 2/13/78
Final version received 4/3/78

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 pseudovalues 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, $z_{ij}^2 = (y_{ij} - \bar{y}_j)^2$, where y_{ij} is the ith observation in the jth 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 (γ_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, σ_j^2, rather than the means, μ_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

$$r_{ij}(w) = [(w + n_j - 2)n_j(y_{ij} - \bar{y}_j)^2 - ws_j^2(n_j - 1)]/[(n_j - 1)(n_j - 2)] . \quad (2.1)$$

Now

$$r_{ij}(0) = n_j(y_{ij} - \bar{y}_j)^2/(n_j - 1) = \tilde{z}_{ij}^2 , \quad (2.2)$$

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

$$\begin{aligned} r_{ij}(1) &= [n_j(y_{ij} - \bar{y}_j)^2 - s_j^2]/[n_j - 2] \\ &= n_j s_j^2 - (n_j - 1)s_{j-i}^2 = q_{ij} , \end{aligned} \quad (2.3)$$

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

Several investigators have considered z^2, $\tilde{z}^2$, and/or q (Levene 1960; Miller 1968; Games, Winkler, and Probert 1972; O'Brien 1978). When used in a regular ANOVA, z^2 and $\tilde{z}^2$ 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 $\tilde{z}^2$. The $r(w)$ variable is simply a weighted average of $\tilde{z}^2$ and q and provides a way to balance the inflated test sizes of $\tilde{z}^2$ 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,

$$\bar{r}_j(w) = \sum_{i=1}^{n_j} r_{ij}(w)/n_j = s_j^2 . \quad (2.4)$$

Thus, ANOVA tests using $r_{ij}(w)$ are readily interpretable because they conform to tests of additive models for

* Ralph G. O'Brien is Assistant Professor, Department of Psychology, University of Virginia, Charlottesville, VA 22901. This research was supported by a Wilson Gee Fellowship and was presented at the 1978 meeting of the American Statistical Association in San Diego, California. The author wishes to thank Jack Hahn for asking the right question.

© Journal of the American Statistical Association
December 1979, Volume 74, Number 368
Theory and Methods Section

σ_j^2 and linear contrasts among the σ_j^2. SSH, the numerator sums of squares for a given null hypothesis, is based only on the s_j^2 and n_j.

The sample variance of $\bar{r}_j(w)$,

$$\hat{V}[\bar{r}_j(w)] = \sum_{i=1}^{n_j} [r_{ij}(w) - s_j^2]^2/[n_j - 1]n_j = [n_j - 2 + w]^2 \sum_{i=1}^{n_j} [r_{ij}(0) - s_j^2]^2/[n_j - 1][n_j - 2]^2 n_j \quad (2.5)$$

Thus, the within-group sums of squares is affected by w:

$$\text{SSWG}(w) = \sum_{j=1}^{J} n_j[n_j - 1][n_j - 2 + w]^2 \hat{V}[\bar{r}_j(0)]/[n_j - 2]^2 . \quad (2.6)$$

Increasing w increases SSWG(w), which consequently decreases the usual ANOVA F statistic. For balanced designs,

$$F(w) = \frac{\text{SSH}/\text{df}_\text{H}}{\text{SSWG}(w)/\text{df}_\text{WG}} = \frac{\text{MSH}}{\text{MSWG}(w)} = \frac{[n - 2]^2 F(0)}{[n - 2 + w]^2}, \quad (2.7)$$

where df_H and df_WG are the usual degrees of freedom associated with SSH and SSWG.

Using Levene's (1960) derivation of $\text{var}[z_{ij}^2]$ and $\text{cov}[z_{ij}^2, z_{i'j}^2]$ in conjunction with the theory on the effects of intraclass correlation developed by Walsh (1947), it follows that

$$E\{\hat{V}[\bar{z}_j^2]\} = \{\text{var}[z_{ij}^2] - \text{cov}[z_{ij}^2, z_{i'j}^2]\}/n_j = \sigma^4[2n_j(n_j - 2) + (n_j - 2)^2\gamma_2]/n_j^3 \quad (2.8)$$

so that

$$E\{\hat{V}[\bar{r}_j(0)]\} = \sigma^4[2n_j(n_j - 2) + (n_j - 2)^2\gamma_2]/[n_j(n_j - 1)^2] . \quad (2.9)$$

By (2.5)

$$E\{\hat{V}[\bar{r}_j(w)]\} = \frac{\sigma^4[n_j - 2 + w]^2[2n_j + (n_j - 2)\gamma_2]}{n_j[n_j - 1]^2[n_j - 2]} . \quad (2.10)$$

It is well known that

$$\text{var}[\bar{r}_j(w)] = \text{var}[s_j^2] = \sigma^4[2/(n_j - 1) + \gamma_2/n_j] . \quad (2.11)$$

Kurtosis (γ_2), w^*, and the Limiting Value of the Kurtosis of r(w) for Several Parent Distributions

Parent	γ_2	w^* (n = 10, 20, 30)			$\tilde{\gamma}[r(w)]$
Uniform	−1.2	−.02	−.12	−.16	−1.3
Normal	.0	.49	.49	.50	12.0
$\chi^2_{(12)}$	1.0	.64	.65	.66	65.5
$\chi^2_{(6)}$	2.0	.72	.73	.74	99.3
Laplace	3.0	.77	.79	.79	84.7
Exponential, $\chi^2_{(2)}$	6.0	.85	.86	.87	213.0
$\chi^2_{(1)}$	12.0	.91	.92	.92	361.7

Equating (2.10) and (2.11) and solving for w produces the weighting factor that yields unbiased estimates for $\text{var}[\bar{r}_j(w)]$:

$$w^* = \left[\frac{2n_j(n_j - 1)(n_j - 2) + \gamma_2(n_j - 1)^2(n_j - 2)}{2n_j + (n_j - 2)\gamma_2}\right]^{\frac{1}{2}} - (n_j - 2) . \quad (2.12)$$

The table contains values of w^* for various parent distributions. The value of w^* increases as γ_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

$$E\{\text{MSH}\} = E\{\text{MSWG}(w^*)\} = \text{var}[s_j^2]n \quad (2.13)$$

for any testable null hypothesis.

In order to examine the relationships of $E\{\text{MSH}\}$ and $E\{\text{MSWG}(w^*)\}$ for unbalanced designs, consider the single degree of freedom contrast

$$H_0 = \sum_{j=1}^{J} c_j\sigma_j^2 = 0 \quad \text{where} \quad \sum_{j=1}^{J} c_j = 0 . \quad (2.14)$$

It can be shown that

$$E\{\text{MSH}\} = \sum_{j=1}^{J} c_j^2 \,\text{var}[s_j^2]/(\sum_{j=1}^{J} c_j^2/n_j) , \quad (2.15)$$

$$E\{\text{MSWG}(w^*)\} = \sum_{j=1}^{J} n_j(n_j - 1)\,\text{var}[s_j^2]/(N - J) . \quad (2.16)$$

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 $E\{\text{MSH}\}$ increases relative to $E\{\text{MSWG}(w)\}$ when the design is unbalanced, because $n_j\,\text{var}[s_j^2]$ 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 $E\{\text{MSH}\}$ increases relative to $E\{\text{MSWG}(w)\}$.

3. If some cells are not involved in the contrast ($c_j = 0$) yet have relatively low σ_j^2, their low $\text{var}[s_j^2]$ reduces $E\{\text{MSWG}(w)\}$ relative to $E\{\text{MSH}\}$, 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 $\text{df}_\text{H} > 1$, 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, $\bar{r}_j(w)$ and $\hat{V}[\bar{r}_j(w)]$ are not independent; in fact, they are highly correlated. Because an analytical expression for this correlation, $\tilde{\rho}_j$, was not obtainable, Monte Carlo estimates of $\tilde{\rho}_j$ were computed for several parent distributions with γ_2 varying from -1.2 to 3 and $n = 8$, 16, and 32. Estimates of $\tilde{\rho}_j$ 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 σ^2.

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

$$t = \sum_{j=1}^{J} c_j s_j^2 / \{[\sum_{j=1}^{J} c_j^2/n_j]\text{MSWG}(w)\}^{\frac{1}{2}} \qquad (2.17)$$

as a t random variable with $N - J$ degrees of freedom. It can be shown that

$$\text{corr}\{\sum_{j=1}^{J} c_j s_j^2, \text{MSWG}(w)\} = \frac{\sum_{j=1}^{J} c_j n_j (n_j - 1)\tilde{\rho}_j}{[(\sum_{j=1}^{J} c_j^2)\sum_{j=1}^{J} n_j^2(n_j - 1)^2]^{\frac{1}{2}}} . \qquad (2.18)$$

If the design is balanced (and all groups have the same distribution), then $\tilde{\rho}_j = \tilde{\rho}$ 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 $\bar{r}_j(w)$ and $\hat{V}[\bar{r}_j(w)]$ is so strong, the single group test of H_0: $\sigma^2 = \sigma_0^2$ using

$$t_{(n-1)} = \{\bar{r}(w) - \sigma_0^2\}/\{\hat{V}[\bar{r}(w)]/n\}^{\frac{1}{2}} \qquad (2.19)$$

and the associated confidence intervals for σ^2 should not be used. This explains why Lemmer (1978) obtained extremely low rejection rates when he used z^2 to test H_0: $\sigma^2 = \sigma_0^2$ versus H_a: $\sigma^2 > \sigma_0^2$.

It follows from (2.1) that

$$\lim_{n_j \to \infty} r_{ij}(w) = (y_{ij} - E[\bar{y}_j])^2 . \qquad (2.20)$$

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

$$\tilde{\gamma}[r(w)] = \frac{\mu_8 - 4\sigma^2\mu_6 + 6\sigma^4\mu_4 - 3\sigma^8}{(\mu_4 - \sigma^4)^2} - 3 , \qquad (2.21)$$

where μ_k is the kth central moment of the parent distribution. The table contains values of $\tilde{\gamma}[r(w)]$ for several parent distributions. If y is normally distributed, then the limiting distribution of $r(w)$ is $\chi^2_{(1)}$, which has a kurtosis of 12. For the exponential parent ($\gamma_2 = 6$), $\tilde{\gamma}[r(w)] = 213$. Monte Carlo estimates of $\tilde{\gamma}[r(w)]$ 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 $F = \text{MSH/MSWG}$ is approximately distributed as an F with $\delta \cdot \text{df}_\text{H}$ and $\delta \cdot \text{df}_\text{WG}$ degrees of freedom, where $\delta = 1 + \gamma_2/N$. If this result applies to $r(w)$, the high values of $\tilde{\gamma}[r(w)]$ reduce the empirical rejection rates for the customary F test, especially in cases with low 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, $\log \tilde{z}^2$ produces extremely inflated regular ANOVA tests (O'Brien 1975), and ANOVA's using $|\tilde{z}|$ are not asympotically 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 $\tilde{\gamma}[r(w)]$ 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 $E\{\hat{V}[\bar{r}_j(.5)]\}$ is nearly unbiased under the normal parent. Using similar logic, one would use $\delta_0 = 1 + 12/N$ and adopt the $F(\delta_0\text{df}_\text{H}, \delta_0\text{df}_\text{WG})$ distribution as the sampling distribution of $F(.5) = \text{MSH/MSWG}(.5)$. 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 δ to conform with their own criteria for robustness.

In the event that w^* and $\delta^* = 1 + \tilde{\gamma}[r(w)]/N$ can be easily determined, then of course they should be used. The consequences of using estimates of γ_2 and $\tilde{\gamma}[r(w)]$ to specify values for w and δ 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 γ_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 2×2 and 4×3 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, $F_N(\delta) = F(\delta\text{df}_\text{H}, \delta\text{df}_\text{WG}, .95)$. When various values

for w and δ 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 so closely paralleled the theoretical conclusions.[1] The $F_E[r(w^*)]$ values were reasonably close to $F_N(\delta = 1 + \tilde{\gamma}[r(w)]/N)$. These results also supported the conjecture that the utility test is acceptably robust. Comparing $F_E[r(.5)]$ 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 $\tilde{\gamma}[r(w)]$ 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 focused on the problem of obtaining proper rejection rates. The use of $r(.5)$ 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 σ_j^2 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_j, then

$$\text{var}[\bar{r}_j] = \sigma^4[\gamma_2/n_j + 2m/(m-1)n_j] \geq \text{var}[s_j^2] = \text{var}[\bar{r}_j(w)] . \quad (4.1)$$

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 characterization 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 σ^2.

[*Received March 1978. Revised April 1979.*]

[1] The original version of this article contained much detail concerning this empirical work, but the referees justifiably felt that these results were simply redundant with the theoretical work. Copies of this original version may be obtained from the author.

REFERENCES

Bartlett, M.S., and Kendall, D.G. (1946), "The Statistical Analysis of Variances—Heterogeneity and the Logarithmic Transformation," *Journal of the Royal Statistical Society*, Suppl. 8, 128–138.

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

———, 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*, Ser. B, 17, 1–26.

Brown, M.B., and Forsythe, A.B. (1974a), "Robust Tests for Equality of Variances," *Journal of the American Statistical Association*, 69, 364–367.

——— (1974b), "The ANOVA and Multiple Comparisons for Data With Heterogeneous Variances," *Biometrics*, 30, 719–724.

Games, P.A., Winkler, H.R., 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.

Kohr, R.L., and Games, P.A. (1977), "Testing Complex a Priori Contrasts on Means From Independent Samples," *Journal of Educational Statistics*, 2, 207–217.

Layard, W.M.J. (1973), "Robust Large Sample Tests for Homogeneity of Variances," *Journal of the American Statistical Association*, 68, 195–198.

Lemmer, H.H. (1978), "A Robust Test for Dispersion," *Journal of the American Statistical Association*, 73, 419–422.

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

Levy, K.J. (1975), "An Empirical Comparison of the Z-variance and Box-Scheffé Tests for Homogeneity of Variance," *Psychometrika*, 40, 519–524.

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

———, and Games, P.A. (1977), "ANOVA Tests for Homogeneity of Variance: Nonnormality and Unequal Samples," *Journal of Educational Statistics*, 2, 187–207.

Miller, R.G., Jr. (1968), "Jackknifing Variances," *Annals of Mathematical Statistics*, 39, 567–582.

Mosteller, F., and Tukey, J.W. (1968), "Data Analysis, Including Statistics," in *The Handbook of Social Psychology* (Vol. 2) (2nd ed.), eds. G. Lindzey and E. Aronson, Reading, Mass.: Addison-Wesley.

O'Brien, R.G. (1975), "Factorial Designs for the Analysis of Spread," unpublished PhD dissertation, University of North Carolina, Chapel Hill.

——— (1978), "Robust Techniques for Testing Heterogeneity of Variance Effects in Factorial Designs," *Psychometrika*, 43, 327–344.

Walsh, J.E. (1947), "Concerning the Effect of Intraclass Correlation on Certain Significance Tests," *Annals of Mathematical Statistics*, 18, 88–96.

Note: This test is available in JMP (Fit Y by X) and SAS (PROC GLM).

A Simple Test for Variance Effects in Experimental Designs

Ralph G. O'Brien
University of Virginia

Although experimental effects are usually assessed through contrasts of group means, there are situations in which differences among the groups' variances are also of interest. Such analyses are infrequently used in behavioral research, possibly because the most common methods are not robust to nonnormally distributed data. A procedure is presented that produces robust tests of the equality of cell variances by simply performing a regular analysis of variance using a transformation of the dependent variable. Special contrasts (e.g., simple effects, subeffects) are also discussed, and an example is given.

It is becoming increasingly recognized that statistical "effects" can be manifested in terms of differences among group variances as well as (or instead of) differences among group means. Thorngate (Note 1) proposed "that social psychologists should become less concerned with differences in central tendency and more concerned with differences in variability. . . . For example, questions about norms, roles, conformity and related topics are highly related to dispersion. Conformity results in uniformity, that is, lack of dispersion" (p. 12). Recently, Games (1978a, 1978b) included variance testing in his factor structure for parametric tests, and Games, Keselman, and Clinch (1979) compared several existing variance testing methods. This article describes a new method.

Researchers should be warned that the traditional homogeneity of variance tests, such as the $F = s_1^2/s_2^2$ test for two groups and Bartlett's χ^2, Hartley's F_{max}, and Cochran's C tests for one-way designs, are severely affected by the distributional form of the data, that is, they are not robust to nonnormality. For example, Miller (1968) reported that a .05 level two-group comparison using $F = s_1^2/s_2^2$ with $n = 25$ per group has a real Type I error rate of only about .007 if the data are uniformly distributed ("light" tails, kurtosis of -1.2) and about .127 if the data are double exponentially distributed ("moderately heavy" tails, kurtosis of 3). Other cases and tests can be even more affected. Even though some texts carry warnings against the use of these nonrobust methods (e.g., Glass & Stanley, 1970, p. 374; Hays, 1973, p. 451; Winer, 1971, p. 205), many statistical routines (including SPSS T-TEST, SAS T-TEST, SPSS ONEWAY, SPSS MANOVA, BMD13D, BMDP3D, BMDP9D) still incorporate them as their only tests of homogeneity of variance.

More robust methods are available. Typically, they involve the use of regular analysis of variance (ANOVA) tests on suitably transformed data. Let y_{ijk} be the kth observation in the i,jth cell of an I × J independent groups design. One well-known transformation, introduced by Levene (1960) and popularized by Glass and Stanley (1970, p. 375), is the absolute deviation about the cell mean,

$$z_{ijk} = |y_{ijk} - \bar{y}_{ij}|, \quad k = 1, 2, \ldots, n_{ij},$$

and is now incorporated into BMDP7D. This test is much more robust than the conventional methods, but it has received criticism. Miller (1968) recommended against it, because, unlike many other ANOVA-based methods, it is not asymptotically ($n_{ij} \rightarrow \infty$) distribution free. Brown and Forsythe (1974b) and Games, Winkler, and Probert

The author wishes to thank Mary Kister Kaiser for her particularly critical reading of an earlier draft of this article.

Requests for reprints should be sent to Ralph O'Brien, Department of Psychology, Gilmer Hall, University of Virginia, Charlottesville, Virginia 22901.

Psychological Bulletin, 1981, Vol. 89, No. 3, 570-574
 0033-2909/81/8903-0570$00.75

(1972) reported that z produced excessive Type I error rates.

An alternative to z is available. Replacing $\bar{y}_{ij}$ with the sample median (Md_{ij}) gives

$$W50_{ijk} = |y_{ijk} - Md_{ij}|,$$

which produces asymptotically distribution-free (Miller, 1968) and robust (Brown & Forsythe, 1974b) tests and conforms to the common definition of dispersion about the median, the average absolute deviation (Hays, 1973, p. 243). O'Brien (1978) found that $W50$ produces excellent power when the underlying distribution of the raw data has heavy tails (leptokurtic).

Another transformation method, developed by Box (1953) and Scheffé (1959, p. 83) and described by Winer (1973, p. 219), is to compute the logarithm of the sample variances of subgroups formed by randomly partitioning each cell's observations. While this test is robust, it lacks power (Games, et al., 1972; Layard, 1973; Levy, 1975; Martin & Games, 1977; Miller, 1968; O'Brien, 1978) and does not give unique test statistics for a given set of data. Power can be partially improved by optimally selecting subgroup sizes (Games et al., 1972; Martin, 1976; Toothaker, Hicks, & Price, 1978), but the lack of uniqueness of this method can produce confusion and doubt when it is applied in practice. Many other transformations have been investigated. (In addition to the above references, see Gartside, 1972.)

Testing Variance Equality With the r Transformation

A new method has been developed to compare cell variances (σ_{ij}^2), and it is directly analogous to the usual ANOVA tests on the cell means (μ_{ij}). The theoretical foundations of this transformation are contained in O'Brien (1979), but that work does not provide a nontechnical "how-to" summary. Using the following description, researchers who understand ANOVA can easily perform tests on group variances. Like its competitors, the r transformation method is not uniformly preferable to every other method in every situation. It is offered only as a general method that behaves acceptably in most situations commonly encountered in behavioral research. It appears to be (a) robust to departures from normality, (b) easy to apply—most statistical software packages can perform the computations, (c) relatively powerful, and (d) generalizable to factorial designs with equal or unequal numbers of observations in the cells.

For a two-way, I × J (fixed effects, completely randomized) design with n_{ij} observations in the i, jth cell, the basic steps of this method are as follows:

1. Compute the sample means, $\bar{y}_{ij}$, and the unbiased sample variances,

$$s_{ij}^2 = \sum_k (y_{ijk} - \bar{y}_{ij})^2/(n_{ij} - 1).$$

2. For every raw observation, y_{ijk}, compute

$$r_{ijk} = \frac{(n_{ij} - 1.5)n_{ij}(y_{ijk} - \bar{y}_{ij})^2 - .5s_{ij}^2(n_{ij} - 1)}{(n_{ij} - 1)(n_{ij} - 2)}.$$

3. Verify that the means for r are the variances of y: $\bar{r}_{ij} = s_{ij}^2$.

4. Use r as an ANOVA dependent variable. Any ANOVA on r effectively tests common linear hypotheses concerning the structure of σ_{ij}^2. General main effects and interactions can usually be tested using standard ANOVA practices. More specific hypotheses (contrasts, simple effects, etc.) or extremely unbalanced designs might require some nonstandard ANOVAs as described below.

This procedure extends to all fixed effects, completely randomized designs, since Steps 1 and 2 are done separately for each particular cell, no matter how those cells are factorially structured. At this time, the feasibility of using r in other types of designs, including those involving random factors or repeated measures, has not been investigated.

The rs are in general nonnormally distributed, the variances of r can be heterogeneous, and rs from the same cell are not independent. However, the particular distributional properties of r have been shown to be reasonably compatible with the analysis of variance.

If the n_{ij} are unequal, a nonorthogonal analysis must be selected that is compatible

Table 1
Means and Variances of y *and* r

A	B	n	$\bar{y}$	s^2	$\bar{r}$	$\hat{V}(r)$
1	1	11	34.82	6.56	6.56	43.0
1	2	11	37.91	37.89	37.89	1,995.6
1	3	10	37.50	20.28	20.28	252.7
2	1	12	35.75	12.57	12.57	455.8
2	2	10	34.20	54.62	54.62	2,473.0
2	3	12	34.83	56.52	56.52	4,991.4

with the goals and philosophy of the researcher. (Overall and Spiegel, 1969, and the ensuing series of *Psychological Bulletin* articles on nonorthogonal ANOVA are critiqued by Herr and Gaebelein, 1978, and Speed, Hocking, and Hackney, 1978.) Except for the fact that the use of r produces ANOVA models for σ_{ij}^2 rather than μ_{ij}, the issues concerning the definitions of the nonorthogonal mean squares for the main effects and interactions remain unchanged. Apart from these concerns about the between-groups variance, the properties of the within-groups variance (MS_w) of a nonorthogonal ANOVA on r are somewhat different than usual. Even under the null hypothesis, the variance of r_{ijk} decreases as n_{ij} increases—a pattern that tends to increase Type I error rates, since MS_w is biased downward. Fortunately, this tendency is not severe unless the n_{ij} are extremely unbalanced (say $\max(n_{ij})/\min(n_{ij}) \geq 4$) and the data are heavy tailed. Therefore, no modification to the ordinary computation of MS_w is usually needed for tests of general main effects and interactions. For troublesome situations, it is prudent to use a Welch-type ANOVA that assumes separate group variances (Brown & Forsythe, 1974a; Kohr & Games, 1974, 1977). BMDP7D computes such statistics for one-way designs.

It has been shown that the variance of s_{ij}^2 increases as a function of σ_{ij}^4. Thus, the variance of *any* spread variable, including r_{ijk}, can be extremely heterogeneous among the groups when real differences in the σ_{ij}^2 exist. Therefore, the overall MS_w may be an inappropriate error term for specific contrasts (or simple effects or subeffects) that do not involve all the cells of the design or have unequal absolute contrasting weights. For example, if the i', j'th cell is not included in the contrast and $\sigma_{r_{i'j'}}^2$ is generally larger than the average σ^2 of the cells being contrasted, then the overall MS_w of r will tend to overestimate the average variance of the rs in the contrasted cells. This produces conservative F statistics. If $\sigma_{r_{i'j'}}^2$ is smaller relative to the average σ^2 of the contrasted cells, inflated F statistics will result. Thus at the very least, the MS_w error term for any test should be based only on the rs that are actually involved in the contrast. Once again, the best solution is to use a Welch-type contrast method (available in SPSS ONEWAY for single degree of freedom contrasts). Similar issues and solutions also apply to multiple comparisons of $\bar{r}_{ij}$ (Howell & Games, 1974; Tamhane, 1977).

Because the sample means and variances of r are strongly correlated, it should not be used to test one-group hypotheses (H_0: $\sigma^2 = \sigma_0^2$) using $t = [\bar{r} - \sigma_0^2]/[\Sigma (r_k - \bar{r})^2/(n-1)n]$, nor should it be similarly used to compute confidence intervals for σ^2. Unfortunately, this problem is also present for all the other known transformation methods.

Example of r Transformation to Test Variances

Let us consider an artificial example of the use of the r transformation for ANOVA tests on variances. Table 1 contains the means and variances of a dependent measure, y, for a 2 × 3 fixed effects design with unequal numbers of observations in the cells.

Table 2
ANOVA Summaries for y *and* r

		Tests of means ($\bar{y}$)			Tests of variances ($\bar{r} = s^2$)		
Source	*df*	*MS*	*F*	*p*	*MS*	*F*	*p*
A	1	54.03	1.72	.19	6,340	3.62	.06
B	2	5.18	.17	.85	8,308	4.75	.01
AB	2	32.98	1.05	.36	1,306	.75	.48
Within cells	60	31.31			1,747		

Note. The method of weighted squares of means (Speed, Hocking, & Hackney, 1978) was employed because it tests *unweighted* hypotheses about the means. This is the same as Overall and Spiegel's (1969) complete least squares method and Herr and Gaebelein's (1978) standard parametric (STP) method.

A nonorthogonal analysis of variance on y (analysis of means) was completed and is summarized in Table 2. The A, B, and AB mean effects are not significant.

Because most computer programs (in this case, SPSS MANOVA) easily produce the cell means and variances (or standard deviations) of the dependent variable, the computation of r is straightforward. For example, the first raw observation is $y_{111} = 34$ so that

$$r_{111} = 9.5(11)(34 - 34.82)^2/90 - .5(6.56)10/90 = .416.$$

After completing the remaining transformations, the means, $\bar{r}_{ij}$, and variances, $\hat{V}_{ij}(r)$, can be computed (Table 1). Note that the means for r equal the variances for y. Using the same ANOVA design that was employed for the means analysis, the tests on r (Table 2) indicate strong support for a B main effect: The variances are differing over the levels of B. The absence of either an A or an AB effect makes the interpretation straightforward.

Paired comparisons among the marginal variances for the three levels of B were performed using three different 2 × 2 designs in order to obtain within-cells error terms that are unaffected by the cells not involved in the contrasts. Thus, the test of Level 1 of B versus Level 2 is not affected by the high variance in Level 3, nor is the Level 2 versus Level 3 test affected by the low variance in Level 1. The results in Table 3 show that if they are taken a priori, "1 versus 2" and "1 versus 3" are both significant at the .05 level. Using a Bonferroni level of significance ($\alpha = .05/3 = .0167$), as suggested by Morrison (1976, p. 33), only the "1 versus 2" comparison is significant with a family-wise error rate of .05.

Table 3
Paired Comparisons of Marginal Means for B

Comparison	MS_B	df_B	MS_w[a]	df_w	F	p
1 vs. 2	14,747	1	1,191	40	12.38	.001
2 vs. 3	660	1	2,549	39	.26	.614
1 vs. 3	9,298	1	1,527	41	6.09	.018

[a] The within-cells error terms are based only on those cells actually being compared.

Other Matters

Interpreting tests on variances is somewhat different from interpreting tests on means. Statistically significant tests on variances may reflect artifacts of the particular measure being used. There may be floor/ceiling effects and dependencies between the mean and the variance of the distribution of y (such as the Poisson or exponential). In these cases, the variance effects are not qualitatively different from the mean effects.

A significant variance effect can result if a factor is not included in the statistical design. For example, consider a 2 × 2 design with $\mu_{11} = \mu_{12}$, $\mu_{21} \neq \mu_{22}$ and $\sigma_{11}^2 = \sigma_{12}^2 = \sigma_{21}^2 = \sigma_{22}^2$. If the B factor is ignored, the variance in the first A "group" is less than the variance in the second "group." Thus, researchers should consider whether a variance effect simply reflects an interaction mean effect involving some ignored factor.

Finally, little has been mentioned here about the real Type I error rates of this test. Briefly, the real α level of the test increases as the kurtosis of the raw observations increases or the degree of cell-size imbalance increases. However, unless one has a severely unbalanced design with very heavy tailed data, there is little need to worry about excessive real α levels. The $W50$ variable can be used in such cases, and the r method has several refinements that are not reviewed here. Statistical zealots and researchers with special problems can refer to O'Brien (1978, 1979) for more details.

Reference Note

1. Thorngate, W. *The analysis of variability in social psychology* (Report 74-7). Edmonton, Canada: University of Alberta, Social Psychology Labs, Department of Psychology, 1974. (An extensive revision of this manuscript is entitled Support tests of dispersion and location differences in frequency tables. *Representative Research in Social Psychology*, 1975, *6*, 76–81.)

References

Box, G. E. P. Nonnormality and tests on variances. *Biometrika*, 1953, *40*, 318–335.

Brown, M. B., & Forsythe, A. B. The ANOVA and multiple comparisons for data with heterogeneous variances. *Biometrics*, 1974, *30*, 719–724. (a)

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

Games, P. A. A three-factor model encompassing many possible statistical tests on independent groups. *Psychological Bulletin*, 1978, *85*, 168–182. (a)

Games, P. A. A four-factor structure for parametric tests on independent groups. *Psychological Bulletin*, 1978, *85*, 661–673. (b)

Games, P. A., Keselman, H. J., & Clinch, J. J. Tests for homogeneity of variance in factorial designs. *Psychological Bulletin*, 1979, *86*, 978–984.

Games, P. A., Winkler, H. R., & Probert, D. A. Robust tests for homogeneity of variance. *Educational and Psychological Measurement*, 1972, *32*, 887–909.

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

Glass, G. V., & Stanley, J. C. *Statistical methods in education and psychology*. Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Hays, W. L. *Statistics for the social sciences* (2nd ed.). New York: Holt, Rinehart & Winston, 1973.

Herr, D. G., & Gaebelein, J. Nonorthogonal two-way analysis of variance. *Psychological Bulletin*, 1978, *85*, 207–216.

Howell, J. F., & Games, P. A. The effects of variance heterogeneity on simultaneous multiple-comparison procedures with equal sample size. *British Journal of Mathematical Statistical Psychology*, 1974, *27*, 72–81.

Kohr, R. L., & Games, P. A. Robustness of the analysis of variance, the Welch procedure, and a Box procedure to heterogeneous variances. *Journal of Experimental Education*, 1974, *43*, 61–69.

Kohr, R. L., & Games, P. A. Testing complex a priori contrasts on means from independent samples. *Journal of Educational Statistics*, 1977, *2*, 207–216.

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

Levene, H. Robust tests for the equality of variances. In I. Olkin (Ed.), *Contributions to probability and statistics*. Palo Alto, Calif.: Stanford University Press, 1960.

Levy, K. J. An empirical comparison of the Z-variance and Box-Scheffé tests for homogeneity of variance. *Psychometrika*, 1975, *40*, 519–524.

Martin, C. G. Comment on Levy's "An empirical comparison of the Z-variance and Box-Scheffé tests for homogeneity of variance." *Psychometrika*, 1976, *41*, 551–556.

Martin, C. G., & Games, P. A. ANOVA tests for homogeneity of variance: Nonnormality and unequal samples. *Journal of Educational Statistics*, 1977, *2*, 187–206.

Miller, R. G., Jr. Jackknifing variances. *Annals of Mathematical Statistics*, 1968, *39*, 567–582.

Morrison, D. F. *Multivariate statistical methods* (2nd ed.). New York: McGraw-Hill, 1976.

O'Brien, R. G. Robust techniques for testing heterogeneity of variance effects in factorial designs. *Psychometrika*, 1978, *43*, 327–344.

O'Brien, R. G. A general ANOVA method for robust tests of additive models for variances. *Journal of the American Statistical Association*, 1979, *74*, 877–881.

Overall, J. E., & Spiegel, D. K. Concerning least squares analysis of experimental data. *Psychological Bulletin*, 1969, *72*, 311–322.

Scheffé, H. A. *The analysis of variance*. New York: Wiley, 1959.

Speed, F. M., Hocking, R. R., & Hackney, O. P. Methods of analysis of linear models with unbalanced data. *Journal of the American Statistical Association*, 1978, *73*, 105–112.

Tamhane, A. C. Multiple comparisons in Model I one-way ANOVA with unequal variances. *Communications in Statistics—Theory and Methods*, 1977, *6*, 15–32.

Toothaker, L. E., Hicks, J. L., & Price, J. M. Optimum subsample sizes for the Bartlett-Kendall homogeneity of variance test. *Journal of the American Statistical Association*, 1978, *73*, 53–57.

Winer, B. J. *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill, 1971.

Received May 29, 1980 ■