Brian Schultz
Division of Biological Sciences
University of Michigan
Ann Arbor, Michigan 48109
U.S.A.

Received August 11, 1982; February 16, 1983

ABSTRACT: Literature concerning the use of Levene's test for differences in variation is discussed. The use of the median as the estimate of central location in the test statistic is robust and relatively powerful, but the popular uses of the mean or trimmed mean as described by Van Valen (1978) are not robust. Levene's test using the median appears to be the best of tests that have been carefully examined, including the Scheffé-Box test recommended by Sokal and Rohlf (1981). Potential problems with Levene's test as well as other tests of variation are also discussed.

* * *

INTRODUCTION

In Van Valen's (1978) discussion of statistics used to test for differences in variation among two or more samples, he noted that the popular F-test and the related Bartlett's test are too sensitive to nonnormality (cf. Box 1953; Martin and Games 1977; Scheffé 1959; Sokal and Rohlf 1981). They may be used in the analysis of variance, in which variances among means are compared, because means tend to be normally distributed by the central limit theorem. Raw data, however, need not be so well-behaved and these popular tests may be extremely inaccurate. The sensitivity seems to arise from the use of squared terms in test variances when deviations may be more frequent far from the mean than expected under normality (i.e., leptokurtosis). Tests for normality are not sensitive enough to distinguish invalid applications. Van Valen presented three alternatives, including Levene's test (Brown and Forsythe 1974; Levene 1960), which were presumed to be robust with respect to departures from normality.

Perhaps because of its simplicity and convenience, Levene's test (described below) appears to be the most popular alternative, at least in the biological literature. Many authors have subsequently followed Van Valen's recommendations or similar interpretations (e.g., Glass 1966) in the application of this test (e.g. Damuth 1981; Handford 1980; Jenkins 1980; MacKenzie and Sealy 1981; Sargent and Gebler 1980; Sokal and Braumann 1980). However most of these interpretations of Levene's test are incomplete, incorrect, or outdated, including those of Van Valen (1978) and the widely-cited simulation study by Brown and Forsythe (1974). Given its popularity, it seems important to correct common errors, especially since if properly used Levene's test appears to be among the best of the tests for differences in variation that have received careful examination. In this paper I present a more complete discussion of results concerning Levene's test, including some critiques and recent modifications. I will also briefly discuss some of the other popular tests and how they compare, and what seem to be some important questions that remain to be answered.

LEVENE'S TEST

Defining Levene's test first in broad terms, suppose one wants to examine differences in variation among a set of samples designated as the X_i's. The data are transformed to new values $Y_i = |X_i - M_i|$, where M_i is a measure of central location such as the mean, median, or a 'trimmed mean', in which some percentage (usually 10%) of the data are trimmed from each tail of each sample's distribution

* * * * * * *

Evolutionary Theory 6: 197-203
The editors thank F.L. Bookstein for help in evaluating this paper.

POTENTIAL PROBLEMS

The distinctions made by Brown and Forsythe were also missed by Martin and Games (1977), who noted the failure of the 10-percent trimmed mean for the skewed distribution and concluded that "None of the Levene statistics has shown the robustness, power, and flexibility necessary to be recommended as a general variance testing technique." They rejected the use of the median out of hand since "dispersion about a median is not a variance," ignoring the evidence presented by Brown and Forsythe that it nonetheless allows a well-behaved, robust test for differences in "spread." (cf. Conover 1981; Lemmer 1980; Martin 1976; O'Brien 1978). They therefore omit Levene's tests from their own simulations, comparing the performance of other statistics of variation (i.e. various forms of the Box test and the jackknife). This is unfortunate since they use χ^2 distributions (4 d.f. and 2 d.f.) as their "leptokurtic" distributions for examining the effects of nonnormality. These distributions are also positively skewed, and as Brown and Forsythe showed, skewness is where the use of Levene's test incorporating the median may be expected to perform particularly well.

Martin and Games say that Miller (1968) similarly rejected the use of the median, but recall that Miller also showed that the use of the median in Levene's test would be asymptotically distribution free, and it is clear from Miller's discussion that he did not use it in simulations mainly because he discovered this property after the study was otherwise completed!

While Levene's test using the median was sometimes conservative (for symmetrical distributions), referring to a low probability of rejecting the null hypothesis when it is in fact true, this does not necessarily mean it is not powerful, which refers to the ability to reject the null hypothesis when it is in fact false (cf. Games et al. 1979). Conover et al. (1981), comparing 56 tests of variation, found it to be one of the three best "...on the basis of robustness and power." O'Brien (1978) and Games et al. (1979) found it to be substantially more powerful than the Scheffé-Box test that was recommended by Martin and Games.

Sokal and Braumann (1980) describe Levene's test with use of the mean, and do not discuss at all the use of trimmed means or medians in Levene's test. They say that for parametric tests on the Y_i's one must still be concerned with the assumption of homogeneous variances among distributions of the Y_i transforms, even though nonnormality can be neglected. They suggest using non-parametric tests of the Y_is in this case. Nonhomogeneity in the variances (heteroscedasticity) seems as though it should have been a serious problem. If there are differences in variation among the X_i's there will probably also be differences in variation among the Y_i's since the transformation simply subtracts a constant from the data points in each distribution and then makes all differences positive in sign. Thus unequal variation in the X_i's will covary with unequal variation in the Y_i transforms. Yet the empirical results of the simulations by Brown and Forsythe and others, and theoretical studies by Miller (1968) and O'Brien (1978) remain that show Levene's test with the median to be robust in terms of frequencies of type I errors that are close to α. However, nonparametric tests for the Y_i's might be worth considering in any case since the Y_i distributions can thus be quite different from normal. Although the parametric tests may be robust to nonnormality when comparing means, non-parametric tests can actually be more powerful under nonnormality (Dixon and Massey 1969). But note that some nonparametric tests were among those rejected by Conover et al. (1981).

A more critical problem may be that of how well Levene's test (henceforth the use of the median is implied unless otherwise stated) performs with small, odd sample sizes. Martin and Games (1977) make note of a conference report by Fellers where Levene's test performs erratically for $n_i \leq 5$, but Martin and Games only use n_i as small as 7 in their simulations with other tests. O'Brien (1978) cites Martin's (1976) note that Levene's test is unduly conservative for small, odd sample sizes, but O'Brien only uses even sample sizes $n_i \geq 4$ in his simulations.

Levene's test was conservative and had low power when comparing four samples of n_i = 5,5,5,5 in simulation results by Conover et al. (1981) but the problem is not apparent in their results for n_i = 5,5,20,20 and no intermediate samples sizes appear in their results. In short, it is not clear what test is best for very small (odd) sample sizes, or what lower limits for various tests are. $n_i<5$ seems a bit small to allow a meaningful estimate of variance in general, but the problem may be of practical importance in testing the assumption of homogeneous variances in multiway ANOVA's where sample sizes are small for some number of cells.

O'Brien (1978) suggests randomly removing a value from small, odd samples to restore even sample sizes using Levene's test, and Conover et al. (1981) suggest removing a median value for some related tests that use a median as an estimate of central location (although this made Levene's test less robust).* Removing data. however is not a very satisfactory solution if data are rare or difficult to obtain! If in a multiway context there really is a problem, I suggest instead using standard techniques for adding an interpolated value to a sample which are usually used to restore balanced designs when some data are missing. One might also add median values rather than remove them, although this will tend to make the test more conservative. These methods of course stipulate that the added values contribute no degrees of freedom (Snedecor and Cochran 1967; Sokal and Rohlf 1981).

Briefly summarizing, Levene's test should be used with the median as the estimate of central location and not the mean or trimmed mean as is commonly practiced. The test then compares favorably with other well-known tests of variation both in terms of robustness and power. It should probably not be used for $n_i \leq 5$ without some modification as suggested above, but further clarification is needed concerning the testing of small odd sample sizes, and whether there is an alternative test that is better for such problems.

OTHER STATISTICS OF VARIATION

The test that performed best in the simulations performed by Martin and Games (1977) was the Scheffé-Box test. Various forms of this test start with samples being randomly divided into $\sqrt{n_i}$ subgroups, a sample variance calculated for each subgroup, and these subgroup variances used as data entries for the analysis of variance. This test was subsequently recommended by Sokal and Rohlf (1981; contrary to the note in Sokal and Braumann, Sokal and Rohlf do not discuss Levene's test). Given the importance of this text it seems important to point out problems with the Scheffé-Box test. Aside from being a bit complex, the test has been criticized and even rejected outright (Brown and Forsythe 1974; Conover et al. 1981) because it involves the arbitrary, random subgrouping of the data. This means extra variation in the results depending on how data chance to be grouped, and can tempt investigators to try other subgroupings if results aren't pleasing the first time** (Brown and Forsythe 1974; Conover et al. 1981; Games et al. 1979; O'Brien 1978). Furthermore, as noted above, the Scheffé-Box test is also well known to be less powerful than Levene's test (Conover et al. 1981; Games et al. 1979; O'Brien 1978).

The other statistics that were recommended by Van Valen were the jackknife and Smith's test. In the jacknife for variances one divides each sample i with sample size n_i into all possible subgroups of size $n_i - 1$ (there are $j = n_i$ subgroups for each n_i), and finds the variance for each subgroup, here designated s_{ij}^2, as well as for each complete set of data, the s_i^2. Pseudovalues σ_{ij}^2 are then constructed where $\sigma_{ij}^2 = n_i s_i^2 - (n_i - 1)s_{ij}^2$. One then performs an ANOVA or analagous test on the means of the σ_{ij}^2. Alternatively one first transforms the variances to their respective logarithms. The literature is ambiguous concerning

*Note that their recommendation of removing data for $n<19$ is again apparently based solely on results comparing n_i = 5,5,5,5 versus n_i = 5,5,20,20.

**Note that these criticisms would also apply to the random removal of data discussed above in connection with Levene's test.

the log-transformation of the jackknife. Van Valen describes without comment the untransformed jackknife even though most authors recommend or require the log-transformation (including those cited by Van Valen· Arveson and Schmitz 1970; Bissel and Fergusen 1975; Miller 1974). With logs it is well known to be nonrobust with respect to normality (cf. Martin and Games 1977; O'Brien 1978). However, O'Brien (1978) found the untransformed jackknife to be well-behaved and recommended it along with Levene's test (although for no stated reason he later recalls only Levene's test -- O'Brien 1981). He suggests that earlier authors were needlessly alarmed (and this test therefore subsequently neglected) because of negative values that can appear in confidence intervals about the the variances, but which could simply be truncated. Note that O'Brien used smaller cell sample sizes (sets of n = 4;8;12; 16;20;24) than deemed necessary for the jackknife by Van Valen. However, Arveson and Schmitz (1970) earlier concluded that

> The current paper shows the need to use the jackknife technique in conjunction with a variance stabilizing transformation in the case of moderate-sized samples.

because while it otherwise appeared to be robust with respect to nonnormality, it was not powerful with distributions that approximated normality. Thus the jackknife itself seems to require further clarification.

O'Brien's observations concerning the jackknife are merely mentioned by Games et al. and seem to have received little further critical examination. See Parr and Schucany (1980) for a bibliography of the statistical literature concerning the jackknife. Smith's test, other than its correction by Tamm (1980) and Van Valen (1980) has also apparently not been widely discussed. More recently, some new statistics have been suggested by O'Brien (1979; 1981), Lemmer (1978; 1979), and Conover et al. (1981) that also have yet to stand the test of time. If the tests discussed here are any indication, tests for differences in variation should require careful scrutiny before being widely accepted.

CONCLUSIONS

While the literature is filled with repeated demonstrations of many tests of variation that are not reliable (e.g. Gartside 1972; Levy 1975; Overall and Woodward 1974, and others above) there are number of tests that are new or otherwise have not received as much critical evaluation. These seem to include the untransformed jackknife (O'Brien 1978) and Smith's test, both as described by Van Valen (1978, 1980) and newer tests suggested by O'Brien (1979, 1981) Lemmer (1978, 1980) and Conover et al. (1981). These and any other newer tests should be compared thoroughly along with Levene's (median) test, and the older, invalid tests should finally be laid to rest.* The literature also seems to include many capricious choices concerning how many samples or cells are compared, the use of univariate or multivariate designs, and what underlying distributions are used. More systematic comparisons are needed to show what limitations or advantages each test may have. The problem of what tests are best for small sample sizes seems particularly important.

Levene's test itself needs further study. Its performance with small, odd sample sizes is unclear, as is how to best correct for any associated problems. Its use with nonparametric alternatives should also be investigated further. Finally, statistical researchers seem to have forgotten that the choice of the median and 10-percent trimmed mean by Brown and Forsythe (1974) was "arbitrary" and other estimates of central location could be investigated. (e.g., means with more of the data trimmed from the tails, or perhaps the geometric mean). Meanwhile, Levene's test using the median, but not the mean or 10-percent trimmed mean, seems to have stood up best thus far to the test of time both in terms of robustness and power.

* * *

*Games et al. (1979) note that a liberal test can be taken as good reason to accept the null hypothesis if differences in variation fail to appear significant.

ACKNOWLEDGEMENTS

I thank Jessica Bernstein, Mike Hansen, and John Vandermeer for critical comments on the manuscripts. This work was partially supported by NSF grant #DEB8108271 to John Vandermeer.

LITERATURE CITED

Arveson, J.N. and T.H. Schmitz. 1970. Robust procedures for variance component problems using the jackknife. Biometrics 26:677-686.

Bissell, A.F. and R.A. Ferguson. 1975. The jackknife-toy,tool, or two-edged weapon? The Statistician 24:79-100.

Box, G.E.P. 1953. Non-normality and tests on variances. Biometrika 40:318-335.

Brown, M.B. and A.B. Forsythe. 1974. Robust tests for the equality of variances. J. Amer. Stat. Assoc. 69:364-367.

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

Damuth, J. 1981. Population density and body size in animals. Nature 290·699-700.

Dixon, W.J. and F.J. Massey. 1969. Introduction to Statistical Analysis. McGraw Hill, N.Y.

Games, P.A., H.J. Keselman, and J.J. Clinch. 1979. Tests for homogeneity of variance in factorial designs. Psych. Bull. 86:978-984.

Gartside, P.S. 1972. A study of methods for comparing several variances. J. Amer. Stat. Assoc. 67:342-346.

Glass, G.V. 1966. Testing homogeneity of variances. Amer. Ed. Res. J. 3·187-190.

Handford, P. 1980. Heterozygosity at enzyme loci and morphological variation. Nature 286:261-262.

Jenkins, S.H. 1980. A size-distance relation in food selection by Beavers. Ecology 61:740-746.

Lemmer, H.H. 1978. A robust test for dispersion. J. Amer. Stat. Assoc. 73:419-421.

_____ 1980. An estimator for spread. J. Stat. Comp. Simul. 11:107-117.

Levene, H. 1960. Robust tests for equality of variances. In: Contributions to Probability and Statistics (I. Olkin et al., eds.), pp. 278-292. Stanford Univ.Press, Stanford.

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

Lewontin, R.C. 1966. On the measurement of relative variability. Syst. Zool. 15:141-142.

MacKenzie, D.I. and S.G. Sealy. 1981. Nest site selection in eastern and western kingbirds: a multivariate approach. Condor 83:310-321.

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 P.A. Games. 1977. ANOVA tests for homogeneity of variance· non-normality and unequal samples. J. Educ. Stat. 2:187-206.

Miller, R.G. 1968. Jackknifing variances. Ann. Math. Stat. 39:567-582.

_____ 1974. The jackknife-a review. Biometrika 61:1-15.

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

_____ 1979. A general ANOVA method for robust tests of additive models for variances. J. Amer. Stat. Assoc. 74:877-881.

_____ 1981. A simple test for variance effects in experimental designs. Psych. Bull. 89:570-574.

Overall, J.E. and J.A. Woodward. 1974. A simple test for heterogeneity of variance in complex factorial designs. Psychometrika 39:311-318.

Parr, W.C. and W.R. Schucany. 1980. The jackknife: a bibliography. Int. Stat. Rev. 48:73-78.

Sargent, R.G. and J.B. Gebler. 1980. Effects of nest site concealment on hatching success, reproductive success, and paternal behavior of the three spine stickleback, Gasterosteus aculeatus. Beh. Ecol. Sociobiol. 7:137-142.

Scheffé, H. 1959. The Analysis of Variance. Wiley, N.Y.

Snedecor, G.W. and W.G.Cochran. 1967. Statistical Methods. 6th ed. Iowa State Univ. Press, Ames.

Sokal, R.R. and C.A. Braumann. 1980. Significance tests for coefficients of variation and variability profiles. Syst. Zool. 29:50-66.

_____ and F.J. Rohlf. 1981. Biometry. 2nd ed. W.H. Freeman, San Fransisco.

Tamm, S. 1980. Bird orientation: single homing pigeons compared with small flocks. Beh. Ecol. Sociobiol. 7:319-322.

Van Valen, L. 1978. The statistics of variation. Evol. Theory. 4:33-43.

_____ 1980. Erratum. Evol. Theory. 4:202.

Note: Games et al. (1979) offered to copy a FORTRAN program for performing median Levene, jackknife, and Scheffé-Box tests onto computer tapes sent to Games. Sokal and Rohlf (1981) advertized a computing package available through Freeman Co.for Bartlett's, Scheffé-Box, and F_{max} tests.