A NOTE ON LEVENE'S TESTS FOR EQUALITY OF VARIANCES

Raymond J. Carroll*

Helmut Schneider**

*University of North Carolina at Chapel Hill. Research supported by the Air Force Office of Scientific Research Contract AFOSR-F 49620 82 C 0009, with partial support by the Sonderforschungsbereich 123 at the University of Heidelberg.

**Free University of Berlin. Research supported by the Air Force of Scientific Research Contract AFOSR- F 49620 82 C 0009 and by the Deutsche Forschungsgemeinschaft.

ABSTRACT

Consider testing for equality of variances in a one-way analysis of variance. Levene's test is the usual F-test for equality of means computed on pseudo-observations, which one defines as the absolute deviations of the data points from an estimate of the group "center". We show that, asymptotically, Levene's test has the correct level whenever the estimate of group "center" is an estimate of group median. This explains why published Monte-Carlo studies have found that Levene's original proposal of centering at the sample mean has the correct level only for symmetric distributions, while centering at the sample median has correct level even for asymmetric distributions. Generalizations are discussed.

Some Key Words and Phrases: Heteroscedasticity, Analysis of Variance, Asymptotic Theory.

1. Introduction

Consider testing for equality of variances in a one-way analysis of variance with K groups. This problem has a long history and there are many possible tests. For example, in a recent simulation study, Conover, Johnson and Johnson (1981) compare the properties of 56 tests. They conclude that three tests "appear to be the best tests to use on the basis of robustness and power". One of these tests is what we shall call Levene's test, described below.

If X_{ij} represents the ith observation in the jth group, the Levene class is simply the ordinary one-way analysis of variance F-test computed on the terms $|X_{ij}-\hat{\theta}_j|$, where $\hat{\theta}_j$ is an estimate of the "center" of the distribution for the jth group. The Levene mean test centers at the sample mean of the jth group, while the Levene median test centers at the sample median. Conover, et al recommend centering at the median. In their simulations, as well as those of Brown and Forsythe (1974), one sees that the level of Levene's mean test is near the nominal 5% for symmetric data, but not asymmetric data. On the other hand, Levene's median test had approximately the correct level, although it seems to be somewhat conservative for small sample sizes. See Table 1 for a summary of these simulations, which show the dramatic effect of asymmetry on Levene's mean test.

The purpose of this note is to address two questions. First, why is it that both Levene's mean and median tests have approximately correct level for symmetric distributions? Second, why does Levene's median test have approximately correct level for asymmetric distributions, while Levene's mean test does not?

Using simple asymptotic arguments, we obtain a precise answer to these questions. Specifically, a test in the Levene class will have correct asymptotic level only when the centering estimates are estimating the population median. Thus, for symmetric distributions, almost any sensible centering will do, especially either the mean or the median since here they estimate the same quantity. For asymmetric distributions, our result implies that, asymptotically, the only permissible centering will be by the sample median, and Levene's mean test will have the wrong level.

We also consider testing based on the use not merely of $|X_{ij}-\hat{\theta}_j|$ but also based on $G(|X_{ij}-\hat{\theta}_j|)$, where $G(.)$ is a smooth monotone function. This enables us to give some explanation of the Monte-Carlo behavior of other tests considered by Conover, et al.

2. Outline of the Main Result

We now sketch the proof of the result discussed in the previous section. The discussion will be concerned only with the null hypothesis case of equal variances. The model in this null hypothesis case is

$$X_{ij}=\theta_j+\varepsilon_{ij} \ : \ j=1,\ldots, K \text{ and } i=1,\ldots, N_j$$
$$Z_{ij}=|X_{ij}-\theta_j| \ ; \ N=\sum_{j=1}^{K} N_j. \tag{2.1}$$

Here, the (θ_j) represent the true "centers" of the distribution, such as the population mean, median or 10% trimmed mean. The terms (ε_{ij}) are independent and identically distributed, since in this paper we are only considering the levels and not the powers of the tests. It is assumed that estimates of the (θ_j) are available with the property that

$$N_j^{\frac{1}{2}}(\hat{\theta}_j-\theta_j) \Rightarrow \text{Normal } (0, \ \xi^2 <\infty). \tag{2.2}$$

We consider a general class of tests based on

$$R_{ij}=G(W_{ij}), \text{ where} \tag{2.3}$$
$$W_{ij}=|X_{ij}-\hat{\theta}_j|. \tag{2.4}$$

Further, let

$$\ell_{ij}=G(Z_{ij}) \tag{2.5}$$

and suppose G(.) has derivative g(.). Let $F_N(\theta_j)$ denote the usual F-statistic computed on the ℓ_{ij} of (2.5) while $F_N(\hat{\theta}_j)$ is the F-statistic computer on the R_{ij} of (2.3) and given by

$$F_N(\hat{\theta}_j)=MSR(\hat{\theta}_j) \ / \ MSE(\hat{\theta}_j), \tag{2.6}$$

$$MSR(\hat{\theta}_j)=\sum_{j=1}^{K} N_j(\overline{R}_{\cdot j}-\overline{R}_{\cdot\cdot})^2 \ / \ (K-1) \ ,$$

$$\mathrm{MSE}(\hat{\theta}_j)=\sum_{j=1}^{K}\sum_{i=1}^{N_j}(R_{ij}-\bar{R}_{\cdot j})^2/(N-K).$$

Computation of the null hypothesis asymptotic distribution of $F_N(\hat{\theta}_j)$ is facilitated by comparison with the test $F_N(\theta_j)$. Before making this comparison, it is useful to note the null hypothesis limit distribution of the latter test. We have the following result, see Arnold (1980).

Lemma. Assume that for j=1,..., K,

$$N_j/N \to a_j \;:\; 0<a_j<1.$$

Then $F_N(\theta_j)$ is asymptotically chi-squared as $N\to\infty$ with K-1 degrees of freedom.

Assuming that G(.) is Lipschitz it is easy to use (2.2) and the fact that $|W_{ij}-Z_{ij}| \le M|\hat{\theta}_j-\theta_j|$ for some constant M to show that

$$\mathrm{MSE}(\hat{\theta}_j) \xrightarrow{p} \sigma^2=\mathrm{Var}(\ell_{ij}).$$

Let $X_{ij}-\theta_j=\varepsilon_{ij}$ have distribution function F(.) and density f(.). Results of Bickel (1975) or Carroll and Ruppert (1982) can be used to show that

$$N_j^{\frac{1}{2}}(\bar{R}_{\cdot j}-\bar{\ell}_{\cdot j}) - \gamma N_j^{\frac{1}{2}}(\hat{\theta}_j-\theta_j) \xrightarrow{p} 0, \tag{2.7}$$

where

$$\gamma=\int_0^\infty g(v)\{f(v)-f(-v)\}dv \tag{2.8}$$

is independent of the index j.

A simple calculation yields the main result.

Theorem 1. Under the previously stated conditions,

$$F_N(\hat{\theta}_j)-F_N(\theta_j)-Q_N \xrightarrow{p} 0, \text{ where}$$

$$Q_N=\sum_{j=1}^{K}\{\gamma^2H_j^{\,2}-2\gamma H_jN_j^{\frac{1}{2}}(\bar{\ell}_{\cdot j}-\bar{\ell}_{\cdot\cdot})\}/\{\sigma^2(K-1)\},$$

$$H_j=N_j^{\frac{1}{2}}\{(\hat{\theta}_j-\theta_j)-\sum_{\ell=1}^{K}(N_\ell/N)(\hat{\theta}_\ell-\theta_\ell)\}.$$

3. Specific Examples

First consider the class of Levene tests based on computing F-tests using $|X_{ij}-\hat{\theta}_j|$, where $\hat{\theta}_j$ is an estimate of the "center" θ_j. In this case, γ of (2.8) reduces to

$$\gamma=\Pr\{X_{ij}\geq\theta_j\}-\Pr\{X_{ij}\leq\theta_j\}. \tag{3.1}$$

Thus, as a general rule, Levene's test will have asymptotically correct level only when γ=o, i.e., θ_j is the median of the distribution of $\{X_{ij}\}$. This explains why for symmetric distributions, where γ=o, using the mean or the median ought not to have a tremendous effect on the level of the test. For asymmetric distributions, $\gamma\neq$o unless θ_j is the median. Thus, the theory gives some explanation for the Monte-Carlo results observed in Table 1.

A second interesting special case occurs by choosing $G(v)=v^2$. Conover, et al call resulting tests based on mean and median centering LEV2 and LEV2:med respectively. In this case we get

$$\gamma=2(EX_{ij}-\theta_j), \tag{3.2}$$

so that in large samples for asymmetric distributions, only centering at the mean will give tests which have the correct level. In the simulations reported by Conover, et al the LEV2 test with mean centering tended to have levels which were slightly too high at the $(\text{Normal})^2$ = chi-squared with one degree of freedom and the $(\text{double exponential})^2$ distributions. This may well be an indication that asymptotics need larger sample sizes before they reflect reality when dealing with squared residuals from hightly asymmetric, heavy-tailed distributions.

The behavior of the LEV2:med test using $G(v)=v^2$ and median centering is interesting. The simulations of Conover, et al seem to indicate that, for heavy-tailed asymmetric distributions, this test will be conservative. Since $\gamma\neq$o for median centering the Theorem 1 tells us that the level will in general be wrong for large samples, but much more work is needed to tell whether it will be conservative with respect to level.

When there are only two populations and $X_{ij}-\theta_j=\varepsilon_{ij}$ has the distribution of a multiple of a chi-squared random variable with one degree of freedom, it is possible to use the results in the technical appendix to Ruppert and Carroll (1980) to show directly that the LEV2:med test with median censoring is conservative in large samples. We expect that this phenomenon holds fairly generally, but we are unable to provide comprehensive theoretical results.

<u>TABLE 1</u>

The levels of Levene's mean and median tests in simulations reported by Brown and Forsythe (1974) and Conover, Johnson and Johnson (1981).

Distribution	Sample Sizes	Levene Mean Test	Levene Median Test
Normal	(5,5,5,5)	.083	.002
	(10,10,10,10)	.064	.025
	(20,20,20,20)	.058	.039
	(5,5,20,20)	.060	.032
Double Exponential	(5,5,5,5)	.097	.008
	(10,10,10,10)	.077	.033
	(20,20,20,20)	.068	.039
	(5,5,20,20)	.087	.035
Chi-squared, 4 degrees of freedom	(40,40)	.050	.044
	(10,10)	.059	.035
	(20,40)	.050	.044
	(10,20)	.065	.036
Chi-squared, 1 degree of freedom	(10,10,10,10)	.349	.054
	(5,5,20,20)	.293	.043
$(\text{Double Exponential})^2$	(10,10,10,10)	.473	.048
	(5,5,20,20)	.384	.060
Chi-squared, 4 degrees of freedom	(40,40)	.097	.036
	(10,10)	.129	.056
	(20,40)	.109	.047
	(10,20)	.118	.052

REFERENCES

Arnold, S.F. (1980). Asymptotic validity of F tests for the ordinary linear model and the multiple correlation model. J. Am. Statist. Assoc. 75, 890-894.

Bickel, P.J. (1975). One-step Haber estimates in the linear model. J. Am. Statist. Assoc. 70, 428-434.

Brown, M.B. and Forsythe, A.B. (1974). Robust tests for the equality of variances. J. Am. Statist. Assoc. 69, 364-367.

Carroll, R.J. and Ruppert, D. (1982). Robust estimation in heteroscedastic linear models. Ann. Statist. 10, 429-441.

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

Ruppert, D. and Carroll, R.J. (1980). Trimmed least squares estimation in the linear model. J. Am. Statist. Assoc. 75, 828-838.