Robust Tests for the Equality of Variances

MORTON B. BROWN and ALAN B. FORSYTHE*

Alternative formulations of Levene's test statistic for equality of variances are found to be robust under nonnormality. These statistics use more robust estimators of central location in place of the mean. They are compared with the unmodified Levene's statistic, a jackknife procedure, and a χ^2 test suggested by Layard which are all found to be less robust under nonnormality.

1. INTRODUCTION

In many statistical applications a test of the equality of variances is of interest. Examples are found in the lot-to-lot reproducibility of a manufacturing process and in the importance of the homogeneity of variances assumption in many statistical procedures. Unfortunately, the common F-ratio and Bartlett's test are very sensitive to the assumption that the underlying populations are from a Gaussian distribution [3, 12]. When the underlying distributions are nonnormal, these tests can have an actual size several times larger than their nominal level of significance.

Two recent papers [7, 9] have compared several alternative statistics for this problem. The statistics least sensitive to nonnormality were of the same form: each group is randomly divided into subgroups; the standard deviation of each subgroup is calculated; and a one-way ANOVA is calculated between groups using the logarithms of the subgroup variances as data points [2, 3]. Different results may be produced by alternate subdivisions of the data points into subgroups. These procedures were not considered because of the nonuniqueness of the results and the loss of power caused by subdividing the samples [12].

Levene [10] proposed a statistic for equal sample sizes which was subsequently generalized to unequal sample sizes [6]. The statistic is obtained from a one-way ANOVA between groups, where each observation has been replaced by its absolute deviation from its group mean. Miller [11] has pointed out that for very small samples the high correlations between deviations in the same group destroys the validity of the test. However, empirical sampling with ten or more observations per group does not indicate that this is the problem. In our sampling this statistic was not robust when the underlying populations were skew. This led us to consider the median and ten percent trimmed mean, more robust estimates of central locations [1, 8, 13], as alternates to the mean in the calculation of absolute deviations.

The ten percent trimmed mean is the mean of the observations after deleting the ten percent largest and ten percent smallest values in that group. The median can be considered a 50 percent trimmed mean and the mean is a zero percent trimmed mean. The choice of ten percent trimming is arbitrary.

This investigation of the robustness of these statistics is limited to two groups since their robustness or lack of it is well illustrated in this simple case. When the underlying distribution is Gaussian, the usual F-test is a check on the sampling experiment. For purposes of comparison, the jackknife procedure of Miller [12] and a χ^2 test presented by Layard [9] are included.

2. DEFINITION OF THE STATISTICS

Let $x_{ij} = \mu_i + \epsilon_{ij}$ by the jth $(j = i, \cdots, n_i)$ observation in the ith group $(i = 1, \cdots, g)$ where the means μ_i are neither known nor assumed equal. The ϵ_{ij} are independent and similarly distributed with zero mean and possibly unequal variances. For each group the sample mean $(\bar{x}_i)$ and sample variance (s_i^2) are estimated in the usual manner. The statistics compared are:

The usual F-test; $F = s_1^2/s_2^2$. The critical values of F are taken from the Snedecor F-table with $n_1 - 1$ and $n_2 - 1$ degrees of freedom (df) for the $\alpha/2$ and $1 - \alpha/2$ percentiles. Bartlett's test statistic [9] yielded essentially the same results for the two-sample case and therefore is not reported.

Levene [6, 10] suggested the following statistic: Let $z_{ij} = |x_{ij} - \bar{x}_i|$. Then form the one-way ANOVA statistic

$$W_0 = \frac{\sum_i n_i(\bar{z}_{i\cdot} - \bar{z}_{\cdot\cdot})^2/(g-1)}{\sum_i \sum_j (z_{ij} - \bar{z}_{i\cdot})^2/\sum_i (n_i - 1)},$$

where

$$\bar{z}_i = \sum z_{ij}/n_i \quad \text{and} \quad \bar{z}_{\cdot\cdot} = \sum\sum z_{ij}/\sum n_i .$$

The critical values of W_0 are obtained from the Snedecor F-table with $g - 1$ and $\sum_i (n_i - 1)$ df. Alternate formulations considered by Levene were to replace z_{ij} by $\sqrt{z_{ij}}$ or by $\log(z_{ij})$. Since, in our empirical sampling, both are less powerful than using z_{ij} (although $\sqrt{z_{ij}}$ is quite similar), only W_0's results are reported.

Replacing the mean $\bar{x}_i$ by the median x_1' in forming z_{ij}

* Morton B. Brown is senior lecturer, Tel-Aviv University, Tel-Aviv, Israel, formerly visiting research statistician, Department of Biomathematics, University of California, Los Angeles (1972–73). Alan B. Forsythe is supervising statistician, Department of Biomathematics, Health Sciences Computing Facility—AV 111, Center for Health Sciences, University of California, Los Angeles, Calif. 90024. This research was supported by NIH Special Research Resources Grant RR-3. Both authors contributed equally to the research in this article and are listed alphabetically.

© Journal of the American Statistical Association
June 1974, Volume 69, Number 346
Applications Section

defines W_{50}. By replacing the mean $\bar{x}_i$ by $\tilde{x}_i$ where $\tilde{x}_i$ is the ten percent trimmed mean of the ith group, W_{10} is defined.

Layard [9] suggested a χ^2 test statistic which is a function of the kurtosis. The kurtosis is estimated by pooling the numerators and denominators of the individual estimates of kurtosis for each group. Using his notation let

$$S' = \sum (n_i - 1) \left[\log s_i^2 - \frac{\sum (n_i - 1) \log s_i^2}{\sum (n_i - 1)} \right]^2 \bigg/ \hat{\tau}^2,$$

where

$$\hat{\tau}^2 = 2 + \{1 - (g/\sum n_i)\} \hat{\gamma},$$

and

$$\hat{\gamma} = \frac{(\sum n_i) \sum \sum (x_{ij} - \bar{x}_i)^4}{\{\sum \sum (x_{ij} - \bar{x}_i)^2\}^2} - 3,$$

is the estimate of the kurtosis. Then S' is asymptotically distributed as a χ^2 variable with $(g - 1)$ df.

Miller's [12] jackknife procedure was generalized by Layard [9]. Let

$$s_{i(j)}^2 = \left(\frac{1}{n_i - 2}\right) \sum_{k \neq j} (x_{ik} - \bar{x}_{i(j)})^2,$$

where

$$\bar{x}_{i(j)} = \left(\frac{1}{n_i - 1}\right) \sum_{k \neq j} x_{ik} \quad (i = 1, \cdots, g)$$

and let

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

Compute an F-statistic from a one-way analysis of variance on the U_{ij}, namely,

$$J = \frac{\sum n_i (\bar{U}_{i\cdot} - \bar{U}_{\cdot\cdot})^2 / (g - 1)}{\sum \sum (U_{ij} - \bar{U}_{i\cdot})^2 / \sum (n_i - 1)},$$

and test H_0 by approximating the null distribution of J by the $F_{g-1, \Sigma(n_i - 1)}$ distribution

3. DESIGN OF THE SAMPLING EXPERIMENT

Pseudorandom numbers were generated from four underlying distributions: the Gaussian by the Box-Muller method [4]; the chi-square distribution with four df as proportional to the logarithm of the product of two uniform random numbers; the Student-t with four df as proportional to the ratio of a Gaussian to the square root of a chi square with four df; and the Cauchy as the ratio of two Gaussians. The chi-square distribution is skewed to the right while the last two distributions are symmetric but long-tailed. The long-tailedness of the Cauchy is more pronounced than that of the t_4. For each distribution a different series of random numbers was used.

A set of 80 random numbers were generated at one time. Pairs of groups of sample sizes (40, 40), (20, 40), (10, 10), and (10, 20) were selected from the set of 80 observations and used in the calculation of the test statistics. The observations in the second sample were rescaled to reflect the ratios of population variances of the two groups; i.e., 1:1, 1:2, 1:4, 2:1, and 4:1. All the test statistics were computed on the same pairs of groups. These were repeated 1,000 times for each distribution in ten blocks of 100 trials. The numbers of rejections at both the nominal five-percent and one-percent levels of significance were recorded.

4. RESULTS AND CONCLUSIONS

The results of the empirical sampling from the Gaussian, Student's-t with four df and the chi-square with four df are summarized in the table. Only the results at the normal five-percent level of significance are presented since those at the one-percent level are similar. The standard error of each of the probabilities was estimated for ten blocks of trials and was well approximated by $\hat{p}(1 - \hat{p})/1{,}000$ where $N = 1{,}000$ is the number of simulations. Therefore, the standard error for five percent can be taken as .7 percent. To avoid distracting the reader, the power results are not shown for tests whose empirical sizes are greater than eight percent and are parenthesized for tests whose sizes are between seven and eight percent.

When there are equal variances in the two groups, the results for the Gaussian distribution (table, part A) indicate that W_{50} is conservative for small sample sizes. The results for the other test statistics are not inconsistent with the sampling error. The powers of these statistics do not differ greatly when the differences in their empirical size are taken into account.

When the distribution is long-tailed (table, part B), the F-test rejects far too often. For unequal groups, the size of the jackknife statistic is larger than it should be. This may be due to the lack of robustness of the ANOVA when the within group variances are unequal [5]. The Layard χ^2 deviates from its level of significance for the smaller sample sizes. Since an estimate of kurtosis is used in the calculation of the statistic, this result is not surprising. Of the three Levene-type procedures, W_{10} appears the most robust. It varies in size less than W_0 and is nearer to 5 percent than W_{50}.

Sampling from the Cauchy distribution emphasized the preceding departures from the nominal size. For the four sets of sample sizes considered here, the rejection rates were excessive for all but W_{10} and W_{50}. The rates for a nominal five-percent level were 60–80 percent rejection for the F-test, approximately 20 percent for the jackknife, 28–36 percent for Layard's χ^2, and 15–21 percent for Levene's W_0. For W_{10} the observed rates were 2.8–5.9 percent and for W_{50} 1.8–3.5 percent. Again W_{10} was closer to five percent than W_{50}.

The results of sampling from the chi square with four df are reported in the table, part C. Only W_{50} maintains its size near the five-percent level of significance. The others depart from their nominal sizes by rejecting far too often.

The preceding results indicate that the equality of variances in long-tailed distributions can best be tested by a statistic of the form of W_{10} and asymmetric distributions by a statistic similar to W_{50}. Therefore, when

Empirical Size and Power ($\alpha = 5\%$)

n_1, n_2	$\sigma_1^2:\sigma_2^2$	F	*Jackknife*	*Layard* χ^2	*Levene* (W_0)	W_{10}	W_{50}
			A. Gaussian distribution				
40, 40	1:1	6.3	5.8	6.5	6.4	6.1	5.1
	2:1	57.5	54.2	56.3	51.1	50.8	48.4
	4:1	98.2	98.1	98.2	97.1	96.9	96.7
10, 10	1:1	5.4	4.6	6.3	5.5	4.9	2.9
	2:1	16.7	13.9	19.5	15.8	14.7	9.8
	4:1	51.3	42.1	51.1	44.3	41.4	31.7
20, 40	1:1	5.8	5.7	6.2	5.8	5.2	4.5
	2:1	43.4	41.3	37.8	38.8	37.5	32.9
	4:1	92.0	89.4	88.4	88.5	88.0	85.7
	1:2	36.6	38.0	42.9	33.2	32.9	31.1
	1:4	92.0	90.5	93.3	86.3	85.3	83.9
10, 20	1:1	4.8	5.2	6.6	5.7	5.2	4.0
	2:1	24.3	19.7	17.7	21.5	20.4	15.7
	4:1	71.7	63.0	59.8	62.2	59.8	52.7
	1:2	16.1	18.4	24.3	15.8	14.8	12.1
	1:4	57.0	57.4	66.0	49.9	48.6	43.0
			B. Student's t on 4 df				
40, 40	1:1	24.1	6.3	4.8	5.0	4.8	4.4
	2:1		33.7	31.2	36.4	34.9	33.2
	4:1		74.3	79.3	87.2	86.4	85.8
10, 10	1:1	15.9	6.8	8.4	5.9	5.3	3.5
	2:1		13.2		12.2	10.3	7.1
	4:1		31.9		32.1	28.6	22.4
20, 40	1:1	21.9	8.2	6.0	5.0	4.7	4.4
	2:1			19.4	27.5	25.7	22.9
	4:1			60.9	75.9	74.2	71.3
	1:2			25.6	21.4	21.1	20.2
	1:4			70.1	66.9	65.2	63.3
10, 20	1:1	16.7	9.0	7.7	6.5	5.2	3.6
	2:1			(14.6)	18.4	15.9	12.4
	4:1			(40.5)	50.3	47.2	42.0
	1:2			(21.9)	10.9	10.0	7.8
	1:4			(49.7)	33.5	30.6	26.7
			C. Chi square on 4 df				
40, 40	1:1	18.4	7.5	6.0	9.7	6.6	3.6
	2:1		(35.7)	35.2		42.1	34.8
	4:1		(80.7)	86.8		92.6	88.4
10, 10	1:1	13.9	7.9	11.2	12.9	9.9	5.6
	2:1		(16.1)				10.3
	4:1		(34.0)				23.8
20, 40	1:1	18.7	7.9	7.2	10.9	7.3	4.7
	2:1		(26.8)	(23.9)		(30.2)	24.4
	4:1		(63.6)	(63.4)		(76.0)	71.1
	1:2		(25.6)	(29.8)		(27.9)	22.2
	1:4		(65.3)	(73.9)		(71.3)	65.1
10, 20	1:1	15.4	9.6	9.9	11.8	8.7	5.2
	2:1						14.8
	4:1						40.1
	1:2						9.3
	1:4						28.1

departures from normality are anticipated, the estimate of the mean for each group in the Levene statistic should be replaced by a more robust estimate of central location. The loss in power that occurs when W_{10} is used in place of W_0 is small relative to the increased probability of a false rejection of the null hypothesis caused by non-normality.

These results also indicate that future research into robust tests of homogeneity of variances must take into consideration robust estimates of location.

[*Received July 1973. Revised January 1974.*]

REFERENCES

[1] Andrews, D.F., *et al.*, *Robust Estimates of Location: Survey and Advances*, Princeton, N.J.: Princeton University Press, 1972.

[2] Bartlett, M.S. and Kendal, D.G., "The Statistical Analysis of

Variance-Heterogeneity and the Logarithmic Transformation," Supplement to the *Journal of the Royal Statistical Society*, Ser. B, No. 1 (1946), 128–38.

[3] Box, G.E.P., "Non-Normality and Tests on Variances," *Biometrika*, 40, No. 1 and 2 (1953), 318–35.

[4] ——— and Muller, M.E., "A Note on the Generation of Normal Deviates," *Annals of Mathematical Statistics*, 29, No. 2 (1958), 610–11.

[5] Brown, M.B., and Forsythe, A.B., "The Small Sample Behavior of Some Statistics which Test the Equality of Several Means," *Technometrics* 16, No. 1 (1974), 129–32.

[6] Draper, N.R. and Hunter, W.G., "Transformations: Some Examples Revisited," *Technometrics*, 11, No. 1 (1969), 23–40.

[7] Gartside, P.S., "A Study of Methods for Comparing Several Variances," *Journal of the American Statistical Association*, 67, No. 338 (June 1972), 342–6.

[8] Huber, P.J., "Robust Statistics: A Review," *Annals of Mathematical Statistics*, 43, No. 4 (1972), 1041–67.

[9] Layard, M.W.J., "Robust Large-Sample Tests for Homogeneity of Variances," *Journal of the American Statistical Association*, 68, No. 341 (March 1973), 195–8.

[10] Levene, H., "Robust Tests for Equality of Variances," in I. Olkin, ed., *Contributions to Probability and Statistics*, Palo Alto, Calif.: Stanford University Press, 1960, 278–92.

[11] Miller, R.G., Jr., Appeared in "Letters to the Editor," *Technometrics*, 14, No. 2 (1972), 507.

[12] ———, "Jackknifing Variances," *Annals of Mathematical Statistics*, 39, No. 2 (1968), 567–82.

[13] Tukey, J.W. and McLaughlin, D.H., "Less Vulnerable Confidence and Significance Procedures for Location Based on a Single Sample: Trimming/Winsorization I," *Sankhyā*, Ser. A., 25, No. 3 (1963), 331–52.