Bulletin of the Psychonomic Society
1976, Vol. 7 (5), 417-420

The robustness of homogeneity of variance tests for asymmetric distributions: A Monte Carlo study

JAMES D. CHURCH and EDWARD L. WIKE
The University of Kansas, Lawrence, Kansas 66045

Robustness properties of three equality of variances tests, the jackknife procedure, the k-sample Box-Andersen test, and Levene's z test, were investigated and compared in a Monte Carlo study employing small samples from underlying gamma distributions and gamma mixtures. Overall, the jackknife test yielded rejection proportions usually closer to nominal levels under departures from normality, but the Box-Andersen test performed relatively well when departures from normality were not extreme. Four recommendations regarding testing variances were offered.

It has long been known that the classic F ratio and Bartlett's tests for evaluating homogeneity of variances may produce excessive Type I error probabilities under departures from normality for the underlying distributions. Accordingly, other tests for equality of variances have been devised and their effectiveness has been assessed by Monte Carlo studies (e.g., Levene, 1960). While this research has shown that certain variance tests are more robust than the classic tests, this conclusion must be qualified by the fact that most of the Monte Carlo studies have used symmetric distributions like the normal, rectangular, and double exponential. Thus, these studies may have limited relevance for the analysis of psychological data such as response times which are asymmetrically distributed.

The primary purpose of this Monte Carlo study was to evaluate the Type I error and power performance of three purportedly robust variance tests, the k-sample Box-Andersen test (Box & Andersen, 1955), Levene's z test (Levene, 1960), and the jackknife method (Miller, 1968; Mosteller & Tukey, 1968, pp. 133-144), with two series of asymmetric distributions. One series consisted of seven gamma distributions (Mood, Graybill, & Boes, 1975), and the other consisted of three mixtures of a pair of gammas which were bimodal as well as asymmetric. The decision to use gamma distributions was made by inspecting response distributions from published psychological experiments. Later we learned that Premack and Kintsch (1970) found that mixture of gamma distributions fit well much of their own data and of others. A second purpose was to assess test performance with more samples and smaller sample sizes than in previous studies. To this end, k = 3 and k = 6 samples were combined with sample sizes of n = 5 and n = 15 throughout. Finally, for comparison purposes, an approximately normal distribution and two nonrobust tests, Hartley's F maximum test and Bartlett's test, were included in the design.

METHOD

Tests

The hypothesis that the k underlying distributions have equal variances was tested against the alternative negation of this assertion. The F maximum and Bartlett tests are described in most statistics texts (e.g., Winer, 1971). The k-sample Box-Andersen test (1955) is a modification of Bartlett's test and uses as a statistic

$$[k(n-1)\log(\sum_j s_j^2/k) - (n-1)\sum_j \log s_j^2]/[1+\hat{\gamma}/2]$$

where $\hat{\gamma}$ is an estimate of kurtosis, $\gamma = [\mu_4/\sigma^4] - 3$. The estimate employed was

$$\hat{\gamma} = \left\{ [kn\sum_i\sum_j(x_{ij}-\bar{x}_j)^4]/[(n-1)\sum_j s_j^2]^2 \right\} - 3.$$

Both the Bartlett and Box-Andersen statistics are assumed to be approximately distributed as χ^2_{k-1} when the k variances are equal.

Levene's z test (1960) is performed by doing the usual one-way ANOVA test for equality of means, using as data the quantities

$$z_{ij} = |x_{ij} - \bar{x}_j|,\ i = 1, \cdots, n, j = 1, \cdots, k.$$

The jackknife procedure (Layard, 1973; Miller, 1968) is also based on a one-way ANOVA test with the data being the quantities

$$u_{ij} = n\log s_j^2 - (n-1)\log s_{j(i)}^2,$$

where $s_{j(i)}^2$ is the sample variance of the subsample obtained by deleting the ith observation from the jth sample.

Procedure

The density formula for the gamma distribution includes r, the shape parameter, and λ, the scale parameter (Mood et al., 1975). For positive integer r values, pseudorandom numbers from the gamma distribution were generated by setting

$$y = -\lambda\log(\prod_1^r x_i),$$

The authors gratefully acknowledge computer services provided by the University of Kansas Computer Center, Requests for reprints should be addressed to Edward L. Wike, Department of Psychology, The University of Kansas, Lawrence, Kansas 66045.

Table 1
Variance Ratios for the Power Experiments

(k,n)	$\sigma_1^2:\sigma_2^2:\cdots:\sigma_k^2$	Case	(k,n)	$\sigma_1^2:\sigma_2^2:\cdots:\sigma_k^2$	Case
(3,5)	1:4:9	2	(6,5)	1:2:3:5:7:9	5
(3,5)	1:25:81	3	(6,5)	1:9:16:36:49:81	6
(3,15)	1:2:3	1	(6,15)	1:1:2:2:3:3	4
(3,15)	1:4:9	2	(6,15)	1:2:3:5:7:9	5

where x_1, x_2, $\cdots$, x_r are pseudorandom numbers from the rectangular distribution on the interval from 0 to 1.

This family of distributions permits the study of the effects of asymmetry. For r = 1, the distribution is extremely "skewed" and for large r values it is approximately normal and symmetric. In each experiment with the gamma distributions, r was a fixed positive integer, and, for j = 1,2, $\cdots$, k, the jth sample was from a distribution with density (f(x;r,λ_j), where λ_j was chosen to be proportional to the desired standard deviation for this distribution.

To include some bimodal distributions, mixtures of two gamma distributions were used, with densities of the form

$$\theta f(x;8,\lambda/2) + (1 - \theta)\, f(x;12,\lambda),$$

where θ = .25, .50, or .75. In each experiment, θ was fixed, and again, for the underlying distribution for the jth sample, λ_j was chosen to be proportional to the desired standard deviation.

Samples were also obtained from approximately normal distributions in which each observation was generated by summing 12 pseudorandom numbers from the rectangular distribution on 0 to 1 and multiplying the result by the desired standard deviation.

Experiments with the following (k,n) were done: (3,5), (3,15), (6,5), and (6,15). In the equal variance studies, the following distributions were used: gamma distributions with r = 1,2,3,6,8,9, and 12, gamma mixtures with θ = .25, .50, and .75, and the normal distribution. The rejection proportions observed in 1,000 trials at α = .05 and α = .01 were obtained.

Power experiments were undertaken for the same four (k,n) pairs and for a subset of the distributions: gammas with r = 6,8,9, and 12, the gamma mixture with θ = .50, and the normal. Pilot studies were done to determine ratios of unequal variances which would produce informative power estimates. The variance ratios which were studied for the six distributions are shown in Table 1.

Power estimates, the rejection proportions in 500 trials, were obtained for each of the five tests at α = .05 and α = .01 for each of the distributions, for the four (k,n) pairs, and for the variance ratios above. A complete report of the rejection proportions for all null hypothesis and power experiments, as well as the results for other variance ratios, may be found in Church and Wike (1975).

RESULTS AND DISCUSSION

In the equal variance experiments, a total of 220 experiments were carried out with each experiment consisting of 1,000 runs. At α = .05, the observed rejection proportions were generally excessive; 184 exceeded the nominal value. Table 2 shows the rejection proportions at α = .05 for the five tests and 11 distributions averaged across the four (k,n) combinations. The nominal rejection proportion is .050. Thus, in the case of the F maximum test for the gamma distribution with r = 1, the null hypothesis was rejected almost seven times too often (348/1,000). It is obvious from Table 2 that both the tests and the nature of the underlying distributions exerted profound effects on the proportions of rejections. In the case of the gamma distributions, at r = 1 the Bartlett, F maximum, and Levene tests led to startlingly large proportions of rejections. With increasing r values, the results approached more closely α = .05. Considering the rejection proportions for the tests across the gamma distributions, the jackknife test was clearly superior, with the Box-Andersen test being next best.

In the experiments with the mixtures of gamma distributions, the θ = .75 mixture yielded excessive rejection proportions, and the θ = .25 mixture resulted in proportions which were slightly below α = .05 in the case of the jackknife, F maximum, and Bartlett tests. Across all the gamma mixtures, the jackknife test again clearly gave the best rejection proportions. While the Box-Andersen test did not yield as satisfactory a set of results as it did with the gamma distributions, it was the second best test. Once more, the Levene test led to too many rejections of the null hypothesis.

Considering all distributions, numbers of samples, and sample sizes, the jackknife was the "best" test; it resulted in a median rejection proportion of .059, with a range from .028 to .130. For the Box-Andersen test, the values were: median, .078, with a range from .048 to .308.

For Levene's test, the rejection proportions were dependent upon both the number of samples and sample size. For n fixed, the proportion was larger for k = 6 than for k = 3 in 21 of 22 cases. For k fixed, the proportion was larger for n = 5 than for n = 15 in 21 of 22 cases. For each of the 11 distributions, the maximum observed proportion was for the k = 6, n = 5 combination; these values ranged from .127 to .452 with a median of .151. Other Monte Carlo studies (Brown & Forsythe, 1974; Levene, 1960) suggest that Levene's test performs fairly well for two samples of Size 20 or more.

Table 2
Mean Rejection Proportions for the Tests and Distributions under Equal Variance Conditions at α = .05

	Gamma Distributions							Mixtures			Norm
Test	r = 1	2	3	6	8	9	12	θ = .75	.50	.25	Dist.
F maximum	.348	.196	.147	.101	.079	.080	.070	.325	.061	.033	.047
Bartlett	.397	.234	.175	.112	.090	.088	.076	.340	.052	.029	.049
Box-Andersen	.120	.089	.082	.079	.069	.076	.067	.203	.107	.078	.070
Levene	.253	.170	.146	.120	.100	.101	.090	.346	.139	.087	.082
Jackknife	.110	.081	.070	.057	.052	.053	.052	.094	.051	.037	.044

Table 3
Mean Rejection Proportions (Power) for the (k,n) Pairs and Different Variance Ratio Cases at $\alpha = .05$

	(3,5)		(3,15)		(6,5)		(6,15)	
Test	Case 2	Case 3	Case 1	Case 2	Case 5	Case 6	Case 4	Case 5
F maximum	.359	.958	.409	.952	.312	.930	.493	.947
Bartlett	.367	.947	.416	.952	.416	.913	.592	.964
Box-Andersen	.335	.860	.358	.862	.264	.676	.428	.835
Levene	.326	.641	.388	.879	.436	.739	.526	.909
Jackknife	.208	.683	.372	.874	.162	.591	.428	.870

The results of the present study suggest that its relative performance deteriorates for smaller sample sizes and larger numbers of samples.

The results for equal variance at $\alpha = .01$ were similar to those reported above.

Why do these variance tests yield such excessive rejection proportions in the equal variance case? If the distributions which produced the largest excesses, i.e., gammas with $r = 1,2$, and 3, and the gamma mixture with $\theta = .75$, are examined, it is apparent that they are long-tailed distributions. Thus, it is possible that the excessive rejections partly result from there being a substantial probability that some, but not all, of the k samples contain relative outliers and hence have inflated sample variances.

As indicated earlier, power experiments with 500 runs each were done only on a subset of the 11 distributions: gammas with $r = 6,8,9$, and 12, the $\theta = .50$ mixture, and the normal. The distributions which were omitted yielded such excessive rejection proportions under the equal variance case that it was pointless to study their power. It should also be noted that the remaining distributions are generally more like the normal distribution. Since the results were about the same for $r = 8$ and $r = 9$, the $r = 9$ case is excluded from the discussion. Again the discussion pertains to tests at the .05 level; comments about the .01 level are similar.

Table 3 presents the rejection proportions for the different variance ratios (see Table 1), the (k,n) combinations, and tests averaged across the distributions. It reveals that the power experiments gave results analogous to the equal variance experiments. That is, the F maximum and Bartlett tests showed the highest rejection proportions (best empirical power) and the Box-Andersen, Levene, and jackknife tests rejected less often than the traditional tests. There was one exception to this pattern: for the $\theta = .50$ mixture, for each k,n pair, at the lowest ratios of variances, the Levene test yielded observed powers greater than those for all other tests. The better power performances of the F maximum and Bartlett tests are consistent with the results of previous studies which have shown that departures from normality tend to increase rejection probabilities for these tests.

For the distributions for which both equal variance and power experiments were done, comparison of the jackknife and the Box-Andersen tests yields different conclusions for $n = 5$ and $n = 15$, the rejection proportions of the Box-Andersen test were closer to the nominal .05 level in 7 of 10 cases, and its observed power always exceeded 89.9% of the jackknife's. For $n = 5$, the proportions observed for the jackknife were closer to nominal in 9 of 10 cases, but its observed power exceeded 89.9% of the Box-Andersen's in only 3 of 20 cases. This suggests the possibility that for samples as large as 15, the Box-Andersen test may be preferable to the jackknife test in cases where departure from normality is not extreme.

The present study leads to several recommendations for testing the variances for equality. First, if the underlying distributions are normal or near normal, the F maximum or Bartlett test should be applied because they yield acceptable levels of Type 1 errors and have superior power. Since the performance of these two tests is very similar, the F maximum test is the preferred test because of its computational simplicity. Second, with asymmetric distributions like those investigated here, the jackknife test appears to be the best test, with the Box-Andersen test as a second choice. It should be noted, however, that both tests have relatively poor power when applied to normal or near normal distributions. Third, the Levene test is not recommended. It led to too many rejections under null conditions and deteriorated in performance in cases of large k and small n. Finally, in light of the demonstrated robustness of ANOVA (Glass, Peckham, & Sanders, 1972), the procedure of testing variances for homogeneity as a prelude to ANOVA appears questionable. In addition, if the F maximum or Bartlett test were to be employed with asymmetric distributions, the investigator would too often conclude that variances are unequal.

REFERENCES

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 the equality of variances. *Journal of the American Statistical Association*, 1974, **69**, 364-367.

Church, J. D., & Wike, E. L. *Homogeneity of variance tests with asymmetric distributions*. (Technical Report No. 3). University of Kansas Computation Center, 1975. (Available from the first author at the Mathematics Department, The University of Kansas, Lawrence, Kansas 66045).

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

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 equality of variances. In I. Olkin (Ed.). *Contributions to probability and statistics.* Palo Alto, California: Stanford University Press, 1960, 278-292.

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

Mood, A. M., Graybill, F. A., & Boes, D. C. *Introduction to the theory of statistics* (3rd ed.). New York: McGraw-Hill, 1975.

Mosteller, F., & Tukey, J. W. Data analysis, including statistics. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (Vol. II)(2nd ed.). Reading, Mass: Addison-Wesley, 1968, 80-203.

Premack, D., & Kintsch, W. A description of free responding in the rat. *Learning and Motivation*, 1970, **1**, 321-336.

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

(Received for publication January 16, 1976.)