Bulletin of the Psychonomic Society
1982, Vol. 20(3), 165-167

Nonrobustness in F tests: 1. A replication and extension of Bradley's study

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

Portions of Bradley's Monte Carlo study on the effects of violations of assumptions on the F test were replicated. The empirical ps agreed well with those observed in the original study. Reciprocal, $\log_e$, and square-root transformations failed to reduce the differences between the empirical ps and nominal αs. An analysis of the relative importance of the factors in the Monte Carlo design suggested other statistical procedures that might be applied to ameliorate the difficulties posed by Bradley's results.

The careful review on violations of assumptions for analysis of variance by Glass, Peckham, and Sanders (1972) concluded that departures from normality and heterogeneity of variance have only slight effects upon p under the null hypothesis when ns are equal. In contrast, Bradley's (1980a) Monte Carlo study revealed some gross differences between empirical ps and nominal αs with equal ns. Bradley used various combinations of four finite populations: X, termed "L-shaped," with a thin, highly peaked distribution and a small, widely dispersed set of outliers to the right; Y, a near-normal population; and x and y, scale-transformed versions of X and Y, respectively, with σs that were one-half those in X and Y. The σs of X and Y were equal and all populations had the same μ. Bradley obtained empirical ps for nominal αs = .05, .01, and .001 based upon 30,000 random samples per population combination for k = 3 and 4 and n = 8.

The aims of the present study were (1) to assess the replicability of Bradley's (1980a) results, (2) to determine the effects of reciprocal, $\log_e$, and square-root transformations of the populations, and (3) to evaluate the influence of the factors of (a) transforming the population values, (b) the shape of the first population, (c) homogeneity of variance, (d) homogeneity of shape, and (e) the number of treatments and their interactions upon the empirical ps. These aims were realized in a Monte Carlo study with a 4 by 2 by 2 by 2 by 2 factorial plan that included 14 of the population combinations investigated by Bradley. Two other combinations of near-normal populations, YYY and YYYY, completed the design and provided checks upon the programming and the random sampling. The transformations were studied because investigators are frequently advised to employ them to achieve normality and to stabilize the variances of positively skewed distributions.[1]

METHOD

The Populations

The basic populations, X and Y, each consisted of 100,000 integers.[2] Approximately 90,000 values in the L-shaped X population fell between 89 and 103; the remaining values in the right-hand distribution of outliers fell between 104 and 218. Thus, the bulk of the values conformed to a discrete, symmetric, leptokurtic distribution. The X population was intended by Bradley (1980a) to model the reaction times that had been recorded for a single subject. The integers in the Y population were bell-shaped and symmetric and fell between 89 and 150. Histograms of X and Y are presented in Bradley (1980b, p. 29). The x and y populations were created by moving the integers in X and Y halfway to the mean. The four populations had a common μ, the σs of X and Y were equal, and σs of x and y were one-half of those in X and Y.

Sampling

The random sampling of values from the appropriate populations and the subsequent computations were done on a Honeywell 66/60 computer. The sampling methods duplicated those described by Bradley (1980a, p. 275f), except that 10,000 test statistics were computed for each population combination instead of 30,000 in the original study. For example, for the XXX population combination, 10,000 random samples of k = 3 and n = 8 were drawn, and the resulting 10,000 F ratios were compared with the critical values of F for α = .05, .01, and .001 in order to determine the empirical ps at these three nominal levels. Thus, an empirical p was the number of rejections of the null hypothesis per 10,000 runs at a specified nominal α. This procedure was repeated for k = 4 and n = 8 and for the remaining 14 population combinations. After replicating Bradley's study, the entire procedure was repeated three times upon the reciprocal-, $\log_e$-, and square-root-transformed integers.

Experimental Design

Earlier we (Wike & Church, 1977) advocated the orthogonal design of Monte Carlo studies so that statistical analysis of the results might illuminate the importance of the factors involved in a study. The addition of YYY and YYYY population combinations to 14 combinations from the Bradley (1980a) study (see Table 1) led to a 4 by 2 by 2 by 2 by 2 design. The first factor (1) was the three transformations and the identity transformation that yielded the original populations, (2) was X vs. Y as a first population (i.e., Combinations 1-8 vs. 9-16), (3) was homogeneity vs. heterogeneity of variance (i.e., 1-4, 9-12 vs. 5-8, 13-16), (4) was homogeneity vs. heterogeneity of shape (i.e., 1, 2, 5, 6, 9, 10, 13, 14 vs. 3, 4, 7, 8, 11, 12, 15, 16), and (5) was the number of treatments, k = 3 vs. k = 4 (i.e., 1, 3, 5, . . ., 15 vs. 2, 4, 6, . . ., 16). By performing analyses of variance upon the observed ps and calculating percents of accounted-for variance, the influence of the five factors and their interactions could be ascertained. It should be noted that when the original values are transformed, the μs and σs for the transformed populations are equal only for Population Combina-

0090-5054/82/090165-03$00.55/0

Table 1
Empirical ps for F Tests at the Nominal α = .05 as a Function of the Original and Transformed Values and Population Combinations

Population	Original	Reciprocal	Log_e	SQRT
(1) XXX	0242	0258	0248	0248
(2) XXXX	0292	0320	0311	0309
(3) XYY	0567	0568	0572	0575
(4) XYYY	0540	0523	0537	0540
(5) Xxx	1306	1610	1455	1375
(6) Xxxx	0995	1273	1131	1057
(7) Xyy	1310	1274	1284	1300
(8) Xyyy	1116	1071	1089	1102
(9) YYY	0479	0477	0647	0477
(10) YYYY	0495	0498	0701	0497
(11) YXX	0570	0618	0588	0573
(12) YXXX	0513	0592	0546	0529
(13) Yyy	0638	0657	0484	0645
(14) Yyyy	0695	0699	0501	0696
(15) Yxx	0699	0762	0725	0717
(16) Yxxx	0725	0818	0773	0747

Note–Leading decimal points have been omitted for all entries.

tions 1, 2, 9, and 10. Therefore, in the transformed populations, generalized moments may be needed to describe the meaning of the levels of the factors in the experimental design.

RESULTS

The empirical ps for the original populations and the three sets of transformed data at the nominal αs of .05, .01, and .001 are displayed in Tables 1-3. The replications of Bradley's (1980a) work, shown in the first column of each table, agreed well with his ps (1980a, Table 3, p. 278). At the .05 level across the 14 population combinations (Bradley did not report his results for YYY and YYYY), the correlation between the ps for the two studies was .999; at .01 it was .997, and at .001 it was .992. Other evidence for the replicability of Bradley's findings came from t tests, with the population combinations as blocks, between the studies. The t values for α = .05, .01, and .001 were −.08, −1.67, and .52, respectively. At α = .01, where p ≅ .12, our ps tended to be slightly smaller, but the mean difference was only about .0005.

Inspection of Tables 1-3 discloses that four population combinations, 5-8, had excessive numbers of rejections of the null hypothesis. On the other hand, two combinations, 1 and 2, generally had ps that were too small. In the cases of excessive rejections, the first population was L-shaped and the combinations involved heterogeneity of variance or heterogeneity of variance and shape. It is also noteworthy that the ratios of empirical ps to nominal αs often varied with the nominal level (e.g., at .05 with the Xyy populations, the ratio was 2.62; at .01 it was 4.46, and at .001 it was 8.50).

The empirical ps for the transformed data are shown in the last three columns of Tables 1-3. At α = .05, the effect of the transformations was to increase the ps in approximately 75% of the cases. Since most of the ps prior to transformation were too large, transforming the populations generally led to poorer results. Although such increases did not occur as regularly at α = .01 and α = .001, it is obvious from an inspection of the empirical ps that the use of these common transformations did not ameliorate the discrepancies between the ps and αs and, frequently, made them worse.

Finally, the observed ps in each table were subjected to analyses of variance, and percents of accounted-for variance were calculated (Hays, 1963, p. 407) to determine how the factors in the design influenced the ps. In Table 1 at the nominal α = .05, three factors accounted for significant amounts of the variance. In order of magnitude, the factors were homogeneity of variance (51.37%), the interaction of the shape of the first population and homogeneity of variance (26.23%), and the shape of the first population (9.63%). No

Table 2
Empirical ps for F Tests at the Nominal α = .01 as a Function of the Original and Transformed Values and Population Combinations

Population	Original	Reciprocal	Log_e	SQRT
(1) XXX	0032	0032	0031	0030
(2) XXXX	0045	0052	0051	0048
(3) XYY	0155	0156	0155	0152
(4) XYYY	0133	0130	0129	0132
(5) Xxx	0501	0571	0501	0518
(6) Xxxx	0323	0397	0323	0336
(7) Xyy	0446	0441	0439	0441
(8) Xyyy	0314	0313	0313	0313
(9) YYY	0111	0102	0172	0111
(10) YYYY	0113	0096	0207	0104
(11) YXX	0228	0241	0238	0232
(12) YXXX	0165	0191	0171	0172
(13) Yyy	0169	0175	0108	0170
(14) Yyyy	0205	0214	0102	0204
(15) Yxx	0226	0246	0237	0230
(16) Yxxx	0251	0299	0273	0265

Note–Leading decimal points have been omitted for all entries.

Table 3
Empirical ps for F Tests at the Nominal α = .001 as a Function of the Original and Transformed Values and Population Combinations

Population	Original	Reciprocal	Log_e	SQRT
(1) XXX	0002	0002	0002	0002
(2) XXXX	0006	0005	0006	0007
(3) XYY	0029	0028	0031	0030
(4) XYYY	0022	0023	0023	0023
(5) Xxx	0205	0203	0204	0206
(6) Xxxx	0116	0122	0118	0117
(7) Xyy	0085	0089	0088	0084
(8) Xyyy	0061	0061	0060	0060
(9) YYY	0008	0008	0037	0008
(10) YYYY	0010	0012	0035	0012
(11) YXX	0073	0078	0075	0075
(12) YXXX	0070	0071	0072	0071
(13) Yyy	0035	0040	0008	0038
(14) Yyyy	0054	0052	0012	0055
(15) Yxx	0068	0076	0073	0071
(16) Yxxx	0073	0093	0085	0078

Note–Leading decimal points have been omitted for all entries.

other main effect or interaction accounted for more than 5% of the variance. On the average, the population combinations with heterogeneous variances had larger ps than did those with homogeneous variances. On the average, the population combinations in which the first population was L-shaped had larger ps than those in which the first population was normal. The interaction between shape of the first population and homogeneity of variance was due largely to the excessive ps from those population combinations with an L-shaped first population and heterogeneous variances (i.e., Population Combinations 5-8). The same three factors were influential at $\alpha = .01$ (46.36%, 26.03%, and 5.19%, respectively).

At $\alpha = .001$, homogeneity of variance (32.53%) and the interaction of shape of the first population and homogeneity of variance (20.11%) were important, as well as the interaction of the shape of the first population and homogeneity of shape (16.44%) and the interaction of homogeneity of variance and shape (9.90%). Population combinations with an L-shaped first population and homogeneous shapes had excessive ps, in contrast with those in which the first population was normal; in populations with heterogeneous shapes, those with a normal first population had greater ps than did those with an L-shaped first population. The interaction between homogeneous variances and shapes resulted from the fact that the average p was .0010 for the combinations with homogeneous variances and homogeneous shapes but excessive in varying degrees in the other three cells in the interaction table.

Finally, analyses of variance and percents of accounted-for-variance computations were performed on the replication portion of the study (i.e., the ps in the first column of Tables 1-3). In each instance, the same factors and approximately the same percents of accounted-for variance were found as were reported above for the analyses of the complete tables.

DISCUSSION

Following the stated aims of the study, we shall consider questions regarding the replicability of Bradley's (1980a) findings, the effects of the transformations, and the influence of the various independent factors in the design. Comparisons of the empirical ps in the Bradley study with those in the present study leave little doubt as to the reliability of Bradley's findings. Four population combinations produced ps that were disturbingly excessive. Two others produced ps that were too small. Earlier, Glass et al. (1972, p. 273) concluded that empirical ps were likely to be smaller than nominal αs with leptokurtic populations. It might be argued that the latter case is not as serious. But, as Bradley has noted, when the empirical ps are too small, the investigator has less power.

The application of the common transformations, reciprocal, $\log_e$, and square root, did not improve the empirical ps. In fact, the transformations frequently increased ps that were already excessive. While such transformations may be beneficial with a set of simple, positively skewed distributions, they did not improve matters when L-shaped distributions like Bradley's (1980a) occurred alone or in combination with normal populations. What features of the L-shaped distributions are critical is unknown. The excessive ps could be due to the unusual degree of kurtosis in the left-hand portion of the distribution or to the widespread tail of outliers, or perhaps both are critical. These features need to be varied systematically in order to evaluate their importance.

The independent factors that generally accounted for most of the variance in the obtained ps were homogeneity of variance, the shape of the first population, and their interaction. These findings suggest some other possible approaches to reducing the differences between the nominal αs and the observed ps. Since inequality of variances appears to be important, the procedures proposed by Brown and Forsythe (1974) might be beneficial. If the X distributions are troublesome because of outliers, then trimmed-means methods (Yuen, 1974; Yuen & Dixon, 1973) might be a possible solution. Another approach would be to do nonparametric analyses of variance upon samples from Bradley's (1980a) population combinations. We will report on the outcome of the application of these four techniques in a subsequent publication.

REFERENCES

Bradley, J. V. Nonrobustness in classical tests on means and variances: A large-scale sampling study. *Bulletin of the Psychonomic Society,* 1980, **15**, 275-278. (a)

Bradley, J. V. Nonrobustness in one-sample Z and t tests: A large-scale sampling study. *Bulletin of the Psychonomic Society,* 1980, **15**, 29-32. (b)

Brown, M. B., & Forsythe, A. B. The small sample behavior of some statistics which test the equality of several means. *Technometrics,* 1974, **16**, 129-132.

Glass, G. V., Peckham, P. D., & Saunders, 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.

Hays, W. L. *Statistics for psychologists.* New York: Holt, Rinehart & Winston, 1963.

Wike, E. L., & Church, J. D. Analysis of variance methods for the design and analysis of Monte Carlo statistical studies. *Bulletin of the Psychonomic Society,* 1977, **10**, 131-133.

Yuen, K. K. The two-sample trimmed t for unequal population variances. *Biometrika,* 1974, **61**, 165-170.

Yuen, K. K., & Dixon, W. J. The approximate behavior and performance of the two-sample trimmed t. *Biometrika,* 1973, **60**, 369-373.

NOTES

1. However, the utility of such transformations has been questioned by Glass et al. (1972, p. 241).

2. We are indebted to James V. Bradley for supplying us with complete details regarding his populations.

(Received for publication July 19, 1982.)