Bulletin of the Psychonomic Society
1982, Vol. 20(3), 168-170

Nonrobustness in F tests: 2. Further extensions of Bradley's study

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

In further extensions of Bradley's work on the effects of violations of assumptions in analysis of variance, random samples of k=3 and 4 and n=8 were drawn from 16 population combinations. The obtained values were subjected to four statistical procedures: the trimmed-means F_T test, the Kruskal-Wallis H test, Welch's W test, and the Brown-Forsythe F* test. The F_T test performed more poorly (had greater discrepancies between empirical ps and nominal αs) than the regular F test. The F* test was the best procedure, but, like the F test, it had excessive ps with population combinations that had a starting L-shaped population and heterogeneous variances. It was concluded that none of the four statistical procedures wholly resolved the problems posed by Bradley's results.

In a Monte Carlo study upon the effects of violations of assumptions in analysis of variance (ANOVA), Bradley (1980a) observed some gross differences between empirical ps and nominal αs. Wike and Church (1982), in a partial replication of the Bradley study, verified the reliability of his results. Data transformations—reciprocal, $\log_e$, and square root—failed to resolve the problems posed by Bradley's findings. Since the Wike-Church study involved a factorial design, the empirical ps were subjected to ANOVAs, and the percents of accounted-for variance were calculated. Three factors, homogeneity of variance, the shape (L or nearly normal) of the first population, and their interaction, generally accounted for most of the variance in the empirical ps. These factors suggested alternative statistical procedures that might be applied to reduce the differences between the empirical ps and nominal αs.

Bradley's (1980a) L-shaped distribution (X) had a wide tail to the right. Thus, the excessive empirical ps could have resulted from the presence of outliers. Yuen (1974) and Yuen and Dixon (1973) have advocated a trimmed t test for two independent samples that are long-tailed or are contaminated by outliers. Accordingly, we generalized their trimmed t test to a trimmed F_T test for k samples.

When violations of assumptions for the F test are encountered, investigators are often encouraged to apply nonparametric tests. An appropriate nonparametric test here is the Kruskal-Wallis H test (Hollander & Wolfe, 1973). It should be stressed, however, that this test is not fully assumption free. In order to make inferences regarding k locations, the test assumes random sampling from the same continuous population (Hollander & Wolfe, 1973, p. 115, Assumption 3). Despite the finite nature of the populations and possible violations of the shape-homogeneity assumption, it was deemed worthwhile to investigate the performance of the Kruskal-Wallis test with Bradley's (1980a) populations.

Finally, two procedures for correcting for heterogeneity of variance, Welch's W test and Brown and Forsythe's (1974) F* test, were applied. In summary, the purpose of the present study was to determine whether or not the discrepancies between the empirical ps and nominal αs observed when the F test was applied to Bradley's (1980a) population combinations could be minimized by the use of the trimmed F_T test, the Kruskal-Wallis H test, Welch's W test, or Brown and Forsythe's F* test.

METHOD

Populations and Sampling

The basic populations, X and Y, each consisted of 100,000 integers.[1] The L-shaped X population was a symmetric, highly peaked distribution of approximately 90,000 integers with a symmetric, wide tail of outliers to the right. The Y distribution was nearly normal and had the same σ as X. The x and y distributions were achieved by moving every integer in X and Y halfway to the mean. Thus, the σs in x and y were equal and were one-half the size of σ in X and Y. These four populations, which had a common μ, are described more fully in Bradley (1980b).

Bradley's (1980a, p. 275f) sampling procedures were followed exactly, except that 10,000 random samples of k = 3 and 4 and n = 8 were drawn, instead of 30,000 in the original study.[2] The sampling and calculations were done on a Honeywell 66/60 computer. At the nominal levels of α = .05, .01, and .001, the numbers of rejections of the null hypothesis were counted. So, at each nominal level, the empirical p was the number of rejections per the number of random samples.

Statistical Tests

The trimmed F_T test was a generalization of the trimmed t test proposed by Yuen (1974) and Yuen and Dixon (1973). The eight values for each treatment were ordered, and the largest and smallest values were trimmed. The resulting trimmed means, based on the remaining six values, were used in calculating the between-treatments sum of squares. The within-treatments sums of squares were based upon the Winsorized sum of squared deviations in each treatment (Yuen & Dixon, p. 369, Formula 1.3). In other words, the original eight values in each treatment were Winsorized by replacing the smallest value by the next-to-the-smallest value and the largest value by the next-to-the-largest value. The dfs for the F_T tests were appropriate for

0090-5054/82/090168-03$00.55/0

the trimmed samples [i.e., $df_B = k - 1$, $df_W = k(h - 1)$, where $h = n - 2g$ and g = the number of times the samples were trimmed]. Because the samples were trimmed once, the $df_W = k(n-3)$. These dfs represent an extension of the two-sample procedure of Yuen and Dixon (1973). Another way of calculating dfs for the trimmed t test has been proposed by Yuen (1974, p. 167), but it was not implemented here.

The application of the second procedure, the Kruskal-Wallis H test, was straightforward. The 24 values for the k = 3 treatments and 32 values for k = 4 treatments were each ranked, and H was computed from the rank sums for the treatments. Since k = 4 and n = 8 exceed the tabled values of H, critical values were obtained by approximating the H distribution by a chi-square distribution with df = k−1. Corrections for ties were not included because the effects of such corrections are minimal.

The other two statistical procedures, Welch's W test and Brown and Forsythe's F* test, were designed to correct ANOVA for heterogeneity of variance. The W statistic incorporates modifications in both the numerator and denominator of the usual F ratio and is approximately distributed as an F statistic with df = k−1 and f. The df for the denominator, f, is dependent upon n_j/s_j^2 and, therefore, must be calculated individually for every ANOVA. The Brown-Forsythe F* test has the usual numerator but a modified denominator. It is approximately distributed as an F statistic with df = k − 1 and f. However, the f for this procedure follows the Satterwaite approximation and, again, must be computed for every ANOVA, as it is a function of n_j and s_j^2. The formulas for the Welch and Brown-Forsythe tests were taken from Brown and Forsythe (1974, p. 130). Both tests performed better than the regular F test in their Monte Carlo study with k = 4 and n = 4 and n = 11 under conditions of unequal variances.

Experimental Design

The population combinations in the Monte Carlo study formed a 2^4 factorial design. The factors were (1) the shape of the first population (Population Combinations 1-8 vs. 9-16 in Table 1), (2) homogeneity of variance (1-4, 9-12 vs. 5-8, 13-16), (3) homogeneity of shape (1, 2, 5, 6, 9, 10, 13, 14 vs. 3, 4, 7, 8, 11, 12, 15, 16), and (4) number of treatments (1, 3, 5, . . ., 15 vs. 2, 4, 6, . . ., 16). By subjecting the obtained ps to ANOVAs and subsequent determination of percents of accounted-for variance, the relative importance of the four factors, of another factor of the four statistical procedures, and of their interactions could be ascertained.

RESULTS

The empirical ps for the nominal αs = .05, .01, and .001 as a function of the statistical procedures and the population combinations are shown in Tables 1-3. The ps obtained with the F test in our previous study are displayed in the first column of each table. Across the three nominal levels, the F test produced ps that were too low in the first two population combinations (X-only populations). The ps were typically too high elsewhere, especially in Population Combinations 5-8, which started with an X population and had heterogeneous variances. The trimmed-means F_T test generally led to higher ps than the F test, and, consequently, the differences between the empirical ps and nominal αs were larger than for the F test in 42 of 48 comparisons. Both the Kruskal-Wallis H test and the Welch W test outperformed the F_T test, especially on Population Combinations 13-16, which started with a Y population and had heterogeneous variances. But both the H and W tests

Table 1
Empirical ps at the Nominal α = .05 as a Function of Statistical Procedures and Population Combinations

Population	F	F_T	H	W	F*
(1) XXX	0242	0277	0400	0144	0099
(2) XXXX	0292	0263	0408	0162	0125
(3) XYY	0567	0856	0752	0770	0468
(4) XYYY	0540	0759	0680	0904	0458
(5) Xxx	1306	3027	3908	1979	0924
(6) Xxxx	0995	2622	3700	2031	0557
(7) Xyy	1310	1872	1650	1918	1138
(8) Xyyy	1116	1573	1518	2272	0976
(9) YYY	0479	0531	0436	0464	0451
(10) YYYY	0495	0529	0444	0492	0460
(11) YXX	0570	1283	0822	0617	0414
(12) YXXX	0513	1232	0772	0588	0315
(13) Yyy	0638	0692	0548	0469	0557
(14) Yyyy	0695	0719	0520	0513	0593
(15) Yxx	0699	1019	0702	0395	0530
(16) Yxxx	0725	1097	0682	0374	0540

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

Table 2
Empirical ps at the Nominal α = .01 as a Function of Statistical Procedures and Population Combinations

Population	F	F_T	H	W	F*
(1) XXX	0032	0056	0054	0024	0014
(2) XXXX	0045	0060	0060	0026	0012
(3) XYY	0155	0260	0144	0198	0094
(4) XYYY	0133	0221	0126	0250	0091
(5) Xxx	0501	1771	1688	0701	0385
(6) Xxxx	0323	1646	1638	0709	0187
(7) Xyy	0446	0646	0454	0768	0313
(8) Xyyy	0314	0462	0352	0976	0247
(9) YYY	0111	0112	0070	0115	0099
(10) YYYY	0113	0120	0070	0108	0095
(11) YXX	0228	0626	0196	0144	0114
(12) YXXX	0165	0688	0184	0148	0083
(13) Yyy	0169	0185	0098	0109	0142
(14) Yyyy	0205	0224	0118	0115	0171
(15) Yxx	0226	0456	0182	0096	0115
(16) Yxxx	0251	0558	0158	0080	0131

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

Table 3
Empirical ps at the Nominal α = .001 as a Function of Statistical Procedures and Population Combinations

Population	F	F_T	H	W	F*
(1) XXX	0002	0005	0000	0001	0000
(2) XXXX	0006	0005	0000	0002	0000
(3) XYY	0029	0063	0006	0040	0012
(4) XYYY	0022	0043	0002	0057	0009
(5) Xxx	0205	0693	0338	0163	0146
(6) Xxxx	0116	0740	0344	0170	0078
(7) Xyy	0085	0133	0018	0141	0043
(8) Xyyy	0061	0097	0020	0203	0034
(9) YYY	0008	0009	0000	0007	0007
(10) YYYY	0010	0009	0000	0012	0007
(11) YXX	0073	0238	0020	0029	0028
(12) YXXX	0070	0331	0020	0029	0023
(13) Yyy	0035	0039	0004	0013	0026
(14) Yyyy	0054	0051	0004	0016	0041
(15) Yxx	0068	0171	0022	0010	0020
(16) Yxxx	0073	0238	0020	0012	0026

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

had excessive ps for Population Combinations 5-8. The "best" test was the Brown-Forsythe F* test. Its p values were closer to α than were those of the F test in 9 of 16 cases at the nominal .05 level, 14 of 16 cases at the .01 level, and 12 of 16 cases at the .001 level. The F* test was the only procedure that did better than the F test on Population Combinations 5-8. Unfortunately, the ps for these troublesome combinations were still excessive. Finally, like the W test, the F* test did worse than the F test on Combinations 1 and 2 and better than the F test on Combinations 13-16.

Finally, in an effort to understand the role of the independent variables in the design in the determination of the empirical ps, these p values were subjected to ANOVAs at each nominal level and the percents of accounted-for variance were calculated. At $\alpha = .05$, four factors–homogeneity of variance (22.57%), the nature of the first population (14.44%), the interaction of these two (22.20%), and the four statistical procedures (9.42%)–were the most influential. At $\alpha = .01$, these same factors accounted for 18.22%, 11.01%, 18.43%, and 10.02%, respectively, of the variance in the ps. Population combinations with heterogeneous variances had higher ps on the average than did those with homogeneous variances. Population combinations beginning with an X population produced larger ps on the average than did those starting with Y. The interaction was largely due to excessive ps for the populations starting with X and having heterogeneous variances in contrast with the remaining three cells in the interaction table. The F_T and H tests generated the highest rejection rates; the F* test generated the lowest, and the W test fell in between. These same conclusions were valid for the four factors at the $\alpha = .01$ level.

At the $\alpha = .001$ level, the relationships were far more complex, with eight factors accounting for more than 5% of the total variance of the ps. Four factors were the same as those above for $\alpha = .05$ and .01, but because of the complexity of the relationships, further details will be omitted.

DISCUSSION

Our previous investigation demonstrated the replicability of Bradley's (1980a) results and the failure of three common transformations to rectify the problems posed by those results. In the present study, none of the four statistical procedures was wholly satisfactory. The results with the trimmed-means F_T test were particularly disappointing. Two modifications might, however, enhance its performance. First, the alternative method for computing f suggested by Yuen (1974) could be tried. Second, the possibility exists that a twice-trimmed means procedure might have been better than the once-trimmed procedure used. Rocke, Downs, and Rocke (1982) have evidence to show that trimming 20% or 25% of the observations from each end of an array of scores yielded more efficient location estimates than did less severe trimming. The present results with the Kruskal-Wallis H test demonstrated that a nonparametric test is not a panacea for all situations in which the assumptions for ANOVA cannot be met. The variance-adjusting tests, Welch's W and Brown and Forsythe's F*, did improve the ps when the starting Y population was combined with heterogeneity of variance, but the troublesome distributions in which X was combined with heterogeneity of variance remained intractable.

Why do Bradley's (1980a) distributions produce such results? This question cannot be answered until the distributions themselves are systematically investigated. It should be pointed out that his X population is basically a mixed distribution that is both highly leptokurtic ($\beta_2 = 13.85$) and positively skewed ($\beta_1 = 3.18$). It is also noteworthy that the "leptokurtic" population investigated in Norton's classic study (Lindquist, 1953) was less peaked ($\beta_2 = 7.0$) and symmetric ($\beta_1 = 0.0$). Clearly, further studies will be needed to understand what features of the X population are critical in the production of too many rejections of the null hypothesis. How typical is the X distribution in reaction time studies? If it is a common occurrence, then investigators should be aware of the possibility of encountering ps less than nominal αs when sampling from X populations and excessive ps when sampling from combinations of X and non-X populations with unequal variances.

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.

HOLLANDER, M., & WOLFE, D. A. *Nonparametric statistical methods.* New York: Wiley, 1973.

LINDQUIST, E. F. *Design and analysis of experiments in psychology and education.* Boston: Houghton Mifflin, 1953.

ROCKE, D. M., DOWNS, G. W., & ROCKE, A. J. Are robust estimators really necessary? *Technometrics,* 1982, **24**, 95-101.

WIKE, E. L., & CHURCH, J. D. Nonrobustness in F tests: 1. A replication and extension of Bradley's study. *Bulletin of the Psychonomic Society,* 1982, **20**, 165-167.

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. We wish to thank James V. Bradley for supplying complete details regarding his populations.

2. In the case of the Kruskal-Wallis H test, 5,000 runs were used.

(Received for publication July 26, 1982.)