EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1974, 34, 789-799.

AN EMPIRICAL COMPARISON OF THE ANOVA F-TEST, NORMAL SCORES TEST AND KRUSKAL-WALLIS TEST UNDER VIOLATION OF ASSUMPTIONS

BETTY J. FEIR-WALSH
University of South Carolina
LARRY E. TOOTHAKER
University of Oklahoma

The present research compares the ANOVA *F*-test, the Kruskal-Wallis test, and the normal scores test in terms of empirical alpha and empirical power with samples from the normal distribution and two exponential distributions. Empirical evidence supports the use of the ANOVA *F*-test even under violation of assumptions when testing hypotheses about means. If the researcher is willing to test hypotheses about medians, the Kruskal-Wallis test was found to be competitive to the *F*-test. However, in the cases investigated, the normal scores test was not consistently better than the *F*-test or the Kruskal-Wallis test and could not be recommended on the basis of this research.

A common problem in applied research is to decide whether or not sample differences in central tendency reflect true differences in parent populations. It is appropriate to use the one-way fixed effects ANOVA *F*-test for the *k*-sample case (two or more groups) if assumptions of normality, homogeneity of variance, and independence of errors are met. When normality and/or equality of variance are doubtful, current literature recommends the use of nonparametric statistical procedures. Two nonparametric counterparts to *F* are the Kruskal-Wallis rank test (Kruskal, 1952) and the expected normal scores test, which used normalized observations in the place of ranks (McSweeney and Penfield, 1969).

There are several types of normal scores tests which have been

developed for the two-sample and the k-sample cases (Hoeffding, 1951; Terry, 1952; Van der Waerden, 1953; Hájek and Šidák, 1967; Puri, 1964). McSweeney and Penfield (1969) have presented a review of the literature, as well as rationale for and derivation of the k-sample case. The Terry-Hoeffding form of the k-sample normal scores test requires the use of special tables (Harter, 1961) to transform ranked data into expected normal order statistics. The Van der Waerden form replaces ranks with inverse normal statistics which can be computed from any standard normal table. Normal scores tests were derived to test the hypothesis of equal populations but are sensitive to location shifts; underlying continuous distributions are assumed and observations are assumed to be drawn randomly and independently from their respective populations. The calculation of the test statistic is performed on the expected normal scores,

$$W = \frac{(N-1)\sum_{i=1}\frac{(\sum_{j} W_{ij})^2}{n_i}}{\sum_{i=1}\sum_{j=1} W_{ij}^2} \tag{1}$$

where:

n_i = the number of observations in the ith sample,
$N = \sum n_i$, the number of observations in all samples combined,
W_{ij} = the jth expected normal order statistic in the ith sample.

rather than on the ranks or the original data. The test is asymptotically distributed under the null hypothesis as chi-square with $k - 1$ degrees of freedom, where k is the number of treatment levels or samples. Large values of the test statistic lead to the rejection of the null hypothesis.

The Kruskal-Wallis test is based on ranks and is suitable for the k-sample case. It is a direct generalization of the two-sample Mann-Whitney U test (Kruskal, 1952; Kruskal and Wallis, 1952). The Kruskal-Wallis statistic tests

$$H = \frac{12}{N(N+1)}\sum_{i=1}\frac{R_i^2}{n_i} - 3(N+1) \tag{2}$$

where:

n_i = the number of observations in the ith sample,
$N = \sum n_i$, the number of observations in all samples combined,
R_i = the sum of the ranks in the ith sample.

hypotheses of equal populations and is sensitive to location shifts. Under the null hypothesis, the Kruskal-Wallis test is also asymptotically distributed as chi-square with $k - 1$ degrees of freedom. It is assumed that sampling is random, that samples are drawn from populations with continuous distributions, and that populations are infinite or sampling is with individual replacement. Large values of the statistic lead to the rejection of the null hypothesis.

The most common index for comparing nonparametric tests to parametric tests is asymptotic relative efficiency or ARE. This index compares the power of one test to the power or efficiency of the other, by using mathematical computations based on extremely large sample sizes and extremely small central tendency or location differences. In fact, sample size is permitted to approach infinity while at the same time location differences approach zero. The ARE of the normal scores test as compared to the F-test has a value of unity for the normal distribution and a lower bound ARE of unity, for non-normal distributions. Therefore, asymptotically the normal scores test can be said to be at least as powerful as F, and when ANOVA violations are present can be more efficient than F. The Kruskal-Wallis test as compared to F has an ARE of .95 for the normal distribution and a lower bound ARE of .864. Thus asymptotically, the Kruskal-Wallis H-test is 95% as powerful as the F-test for the normal distribution and can never asymptotically be less than 86% as powerful. Therefore, with no further information, the normal scores test would appear to be quite competitive to F.

In addition, McSweeney and Penfield (1969) have shown that upon comparing the Kruskal-Wallis test and the normal scores test with samples from both normal and uniform distributions that "the small sample power of the normal scores test is clearly superior to that of the Kruskal-Wallis test in those marginal cases in which a test at a moderate significance level is used to detect small differences in location among non-normal distributions." They state "that the comparison is dependent on the significance level of the test, the location parameter, and sample sizes as well as on the distributions sampled."

Both the enticing ARE and the favorable comparison of the normal scores test to the Kruskal-Wallis test as cited by McSweeney and Penfield (1969) have shown the need for further research in this area. Keeping in mind that asymptotic relative efficiencies are computed for unrealistically large sample sizes with minuscule differences in measures of location, it would seem profitable to the researcher to be aware of the small, medium, and large sample size

performance of the normal scores test. In addition, there has been no comparison of the k-sample normal scores test to its parametric analogue, F, and neither the normal scores test nor the Kruskal-Wallis test have been compared to the ANOVA F-test for skewed distributions.

Further, current literature (Bradley, 1968; Kendall and Stuart, 1961) refers to the nonparametric sensitivity to detect location differences without stating whether mean or median differences will be equally detected. Therefore, a Monte Carlo comparison of the three statistical tests was completed for realistic location differences and realistic sample sizes from a normal distribution and two exponential distributions. One of the exponential distributions was scaled to have equal means under the null hypothesis to investigate the sensitivity of the three tests in detecting mean differences. The other exponential distribution was scaled to have equal medians under the null hypothesis in order to investigate sensitivity of the tests to median differences.

Procedure

Random numbers were selected using a pseudo-random number generator. Depending upon the assumption violation, the numbers were selected from either a normal distribution or from one of two exponential distributions. The random deviates were allocated to four treatment levels that comprised a one-way fixed effects analysis of variance situation.

The observations from the normal distribution were derived by a technique developed by Box and Muller (1958), which generates pseudo-random variables distributed N (0, 1). For the null situation, the means of the four treatment levels were zero. The non-null situation was established by defining values of α_j, $j = 1, 2, 3, 4$, such that the power for the ANOVA F-test would be about .86 for the equal variance condition for the normal distribution. Then the defined α_j's were used for all three statistical procedures, for all three distributions, and for both equal and unequal variance conditions. Specification of the α_j's for the normal distribution was made through the non-centrality parameter, θ, (Pearson and Hartley, 1951) where

$$\theta = \sqrt{\frac{\sum n_j \alpha_j^2}{J \sigma_e^2}} \tag{3}$$

Setting $\sigma_e^2 = 1$ and $J = 4$ and using probability of a Type 1 error equal to .05, the values of α_j were found such that the power was about .86. Since the equal sample size and unequal size cases would

lead to different values of α_j for each of three sample sizes, the values of α_j were calculated for both equal and unequal sample sizes. Values of α_j for the total sample sizes of 28, 68, and 200 are presented in Table 1. The appropriate α_j's were added to the samples in each of the four treatment levels for the non-null situation. Variance differences were established for particular cases by utilizing unequal variances in the ratio of 1:2:3:4, with the average variance equal to unity. The variances used were .4, .8, 1.2, and 1.6. When equal variance cases were desired, the variances were all given a value of unity.

The exponential distributions were derived by a method given by Lehman and Bailey (1968):

$$f(t) = pe^{-pt} \tag{4}$$

with $p = 1$, $E(t) = 1/p = 1$, and $\text{var}(t) = 1/p^2 = 1$. Pseudo-random exponential variables were generated by multiplying the negative of the mean, $-E(t) = -1$, times the natural logarithm of uniform random variates distributed on the unit interval (IBM, 1968). The exponential variates were then scaled so that either the medians would be zero or the means would be zero depending upon which of the two exponential distributions was desired. The resulting skewed populations had either mean or median equal to zero, a variance of σ_j^2, a skewness measure of $\gamma_1 = 2$, and a kurtosis measure of $\gamma_2 = 6$.

For the exponential distribution scaled to have equal means of zero value under the null distribution, the mean of unity was subtracted from every score. Thus the median of .69315 also had the value of unity subtracted from it, yielding a median of −.30685 when variances were equal. When variances were unequal and means

TABLE 1

Values of n_j and α_j
(n_j = Number of Observations per Treatment Level, and N = Total Number of Observations)

		$N = 28$		$N = 68$		$N = 200$	
	j	n_j	α_j	n_j	α_j	n_j	α_j
$= n_j$'s	1	7	−1.0050	17	−.6000	50	−.3360
	2	7	−.3350	17	−.2000	50	−.1120
	3	7	.3350	17	.2000	50	.1120
	4	7	1.0050	17	.6000	50	.3360
$\neq n_j$'s	1	4	−1.0605	11	−.6225	32	−.3495
	2	6	−.3535	15	−.2075	44	−.1165
	3	8	.3535	19	.2075	56	.1165
	4	10	1.0605	23	.6225	68	.3495

were equal and of zero value, the median for group j was $-.30685\sigma_j$; thus the medians were $-.19407$, $-.27445$, $-.33614$, and $-.38814$.

For the exponential distribution scaled to have equal medians of zero value under the null hypothesis, the means will be nonzero. For equal variances, the value of the means was .30685. For unequal variances and equal medians of zero value the mean for group j is $.30685\sigma_j$; thus the means were .19407, .27445, .33614, and .38814.

In order to simulate null and non-null conditions in the exponential distributions, the values of α_j were identical to those used in the normal distributions as shown in Table 1. The variance for equal variance cases for the two exponential distributions was, as in the normal distribution, equal to unity, and for unequal variance cases were equal to .4, .8, 1.2, and 1.6.

Comparisons among the F-test, the normal scores test, and the Kruskal-Wallis test were made on five combinations of sample sizes and variances. These combinations were as follows: (1) equal sample sizes and equal variances, (2) equal sample sizes and unequal variances, (3) unequal sample sizes and equal variances, (4) unequal sample sizes and unequal variances which were positively related, and (5) unequal sample sizes and unequal variances which were negatively related. For each of the five cases, 1000 experiments were performed using observations from the normal distribution, the exponential distribution scaled to have equal means under the null hypothesis, and the exponential distribution scaled to have equal medians under the null hypothesis, where an experiment consisted of computation of each statistical test. The proportion of rejections in 1000 experiments when there were no location differences was referred to as empirical alpha. When differences in location were specified, the proportion of rejections was referred to as empirical power. Theoretical alpha (level of significance) was set at .05. The three statistical tests were then compared in terms of empirical alpha and empirical power for total sample sizes of 28, 68, and 200. It should be noted that the equal sample size, equal variance case for the normal distribution was included in the present study for the purpose of establishing validity of the Monte Carlo method and has been established for the statistics in prior studies.

Results

Normal Distribution

For a total sample size of 28, the ANOVA F-test surpasses the performance of both the Kruskal-Wallis test and the normal scores

test in terms of approximating theoretical alpha and having greater power in all but one case of assumption violation (negatively related sample sizes and variances). The empirical alphas and power of the nonparametric methods were comparable to each other for $N = 28$, with the Kruskal-Wallis test being more preferred than the normal scores test in the unequal variance situations. Current literature has suggested that the normal scores procedures are far less sensitive to heterogeneity of variance than are the parametric or rank procedures (McSweeney and Penfield, 1969), but this robustness was not substantiated for $N = 28$ as shown in Table 2. In general, for this sample size neither of the nonparametric methods compete favorably with F, except for negatively related sample sizes and variances.

The ANOVA F-test is generally the most powerful technique for the larger sample sizes ($N = 68$, $N = 200$), but at the expense of making a few more Type 1 errors than the nonparametric methods. For $N = 68$, the Kruskal-Wallis test provides the best overall approximation to theoretical alpha when variances are unequal. For $N = 200$, while the normal scores test generally provides the best approximation to alpha, the Kruskal-Wallis test also gives a good approximation to alpha with comparable or better power. All three of the tests are competitive for the larger sample sizes.

TABLE 2

Normal Population

		$N = 28$			$N = 68$			$N = 200$		
		F	KW	NS	F	KW	NS	F	KW	NS
$= n'$s	α	.058	.045	.047	.056	.058	.054	.052	.054	.052
$= \sigma^2$	$1 - \beta$	.878	.837	.840	.864	.838	.847	.850	.820	.839
$= n'$s	α	.057	.038	.039	.066	.058	.060	.058	.057	.056
$\neq \sigma^2$	$1 - \beta$	.860	.841	.818	.865	.859	.820	.853	.847	.806
$\neq n'$s	α	.056	.038	.043	.057	.059	.061	.059	.056	.055
$= \sigma^2$	$1 - \beta$	.888	.836	.836	.877	.850	.863	.837	.815	.832
$\neq n'$s	α	.029	.026	.023	.041	.043	.032	.042	.037	.039
$\neq \sigma^2$	$1 - \beta$	.827	.823	.781	.848	.859	.809	.790	.823	.752
positively related										
$\neq n'$s	α	.093	.058	.065	.093	.078	.083	.080	.058	.055
$\neq \sigma^2$	$1 - \beta$	.916	.869	.848	.916	.874	.856	.887	.872	.842
negatively related										

Note.—Entries are the proportion of rejections in 1,000 experiments for the ANOVA F-test (F), the Kruskal-Wallis Test (KW) and the Normal Scores Test (NS) in terms of probability of a Type 1 error (α) and power ($1 - \beta$). Nominal alpha was set at .05, N = total sample size, and n = sample size per treatment level.

Exponential Distribution: Scaled to Have Equal Means under the Null Hypothesis

For this distribution, the ANOVA F-test consistently outperforms the nonparametric methods. As sample size increases ($N = 68$, $N = 200$), the nonparametric tests begin to make far too many Type I errors when variances are unequal. For example, in Table 3, the negatively related sample sizes and variances case for $N = 200$, shows an $\alpha = .440$ for the normal scores test, and an $\alpha = .381$ for the Kruskal-Wallis test while for the F-test, $\alpha = .092$. This extreme increase in empirical alphas is thought to be caused by the nonparametric sensitivity to the unequal medians. Scaling to equal means of zero value under the null hypothesis for a skewed distribution leaves nonzero medians. If the variances are equal, then the medians (though nonzero) are equal; however when variances are unequal, the medians are also unequal. Thus for cases where variances were equal, the nonparametric tests approximated theoretical alpha fairly well with good power. However, the F-test still provided the best approximation to theoretical alpha in most cases for both equal and unequal variance situations.

Exponential Distribution: Scaled to Have Equal Medians under the Null Hypothesis

For this distribution for unequal variance cases, the empirical alphas of the Kruskal-Wallis test and the normal scores test show

TABLE 3

Exponential Population, Scaled to Have Equal Means under the Null Hypothesis

		$N = 28$			$N = 68$			$N = 200$		
		F	KW	NS	F	KW	NS	F	KW	NS
$= n$'s	α	.030	.039	.036	.039	.040	.039	.049	.054	.048
$= \sigma^2$	$1 - \beta$	.863	.913	.893	.857	.974	.972	.861	.991	.994
$= n$'s	α	.047	.073	.072	.045	.133	.125	.053	.377	.389
$\neq \sigma^2$	$1 - \beta$	.927	.973	.977	.920	.978	.987	.883	.816	.859
$\neq n$'s	α	.052	.042	.042	.051	.045	.049	.046	.040	.043
$= \sigma^2$	$1 - \beta$	.885	.924	.916	.868	.971	.977	.853	.992	1.000
$\neq n$'s	α	.033	.042	.040	.031	.091	.077	.034	.272	.278
$\neq \sigma^2$	$1 - \beta$	.859	.934	.942	.865	.962	.980	.842	.970	.915
positively related										
$\neq n$'s	α	.088	.109	.119	.093	.177	.206	.092	.381	.440
$\neq \sigma^2$	$1 - \beta$	.899	.932	.910	.878	.986	.977	.865	.999	.999
negatively related										

Note.—Entries are the porportion of rejections in 1,000 experiments for the ANOVA F-test (F), the Kruskal-Wallis Test (KW) and the Normal Scores Test (NS) in terms of probability of a Type I error (α) and power ($1 - \beta$). Nominal alpha was set at .05, N = total sample size, and n = sample size per treatment level.

a marked decrease as compared to the exponential distribution scaled to have equal means. For the case cited in section 3.2, $N = 200$, negatively related sample sizes and variances, the empirical alpha for the Kruskal-Wallis test has dropped from .381 to .095 and for the normal scores test from .440 to .121. This decrease in empirical alphas with equality of medians under the null hypothesis substantiates the nonparametric sensitivity to median differences. When there exist no median differences, the nonparametric procedures (especially the Kruskal-Wallis, see Table 4) do well in approximating theoretical alpha, and provide good power in detecting median differences when they are specified. The Kruskal-Wallis test provides the best approximation to theoretical alpha with high power for all sample sizes, however the F-test is a good competitor with the normal scores test falling in close proximity.

Conclusion and Summary

When normality and/or homogeneity of variance is doubtful, the ANOVA F-test is the recommended procedure for testing hypotheses about means. The researcher does have the option of testing hypotheses about medians with the assurance that if a significant F-value is obtained, both mean and median differences will be present. When using the Kruskal-Wallis test or the normal scores test in investigating mean differences, with non-normality and heterogeneity of variances, the researcher might very well reject the null hypothesis due

TABLE 4

Exponential Population, Scaled to Have Equal Medians under the Null Hypothesis

		$N = 28$			$N = 68$			$N = 200$		
		F	KW	NS	F	KW	NS	F	KW	NS
$= n'$s	α	.043	.047	.043	.042	.031	.038	.050	.056	.050
$= \sigma^2$	$1 - \beta$	.873	.940	.914	.843	.970	.971	.850	.988	.996
$= n'$s	α	.054	.047	.040	.066	.065	.056	.102	.086	.103
$\neq \sigma^2$	$1 - \beta$	.956	.978	.977	.971	.997	.999	.985	1.00	1.00
$\neq n'$s	α	.036	.041	.037	.039	.050	.043	.049	.050	.049
$= \sigma^2$	$1 - \beta$	.893	.937	.932	.864	.978	.983	.846	.992	.997
$\neq n'$s	α	.029	.031	.029	.047	.046	.042	.060	.065	.073
$\neq \sigma^2$	$1 - \beta$	.927	.966	.971	.953	.996	.999	.981	.999	.999
positively related										
$\neq n'$s	α	.078	.068	.078	.091	.073	.092	.141	.095	.121
$\neq \sigma^2$	$1 - \beta$	.832	.889	.859	.776	.952	.945	.618	.987	.991
negatively related										

Note.—Entries are the proportion of rejections in 1,000 experiments for the ANOVA F-test (F), the Kruskal-Wallis Test (KW) and the Normal Scores Test (NS) in terms of probability of a Type I error (α) and power ($1 - \beta$). Nominal alpha was set at .05, N = total sample size, and n = sample size per treatment level.

to median differences, when means are in fact equal. Thus, the researcher in the case of finding a significant value for the Kruskal-Wallis test or the normal scores test has little assurance that mean differences actually are present. The F-test can further be recommended on the grounds that the F-distribution is more extensively tabled, and that the F-test is the most easily used of the three tests for large sample sizes. There are computer programs readily available for ANOVA, and the tedious task of transforming data is not necessary, as it is for the nonparametric tests.

With non-normality and inequality of variances, the Kruskal-Wallis test might be considered to be the recommended procedure. The researcher, however, must be aware that he is testing for median differences, and must state his null hypothesis in these terms. The Kruskal-Wallis test for large samples does require the tedious chore of ranking, but computer programs are becoming more and more accessible.

The normal scores test, despite its enticing asymptotic relative efficiency, cannot be recommended on the basis of this study. In none of the cases investigated does the normal scores test consistently outperform the Kruskal-Wallis test or the ANOVA F-test. Only in isolated cases could the normal scores test be recommended, and even then, only with reserve, because of the difficulty in transforming data from ranks to expected normal scores.

REFERENCES

Box, G. E. J. and Muller, M. A note on the generation of random normal deviates. *Annals of Mathematical Statistics,* 1958, 29, 610–611.

Bradley, J. V. *Distribution-free statistical tests.* London: Prentice-Hall, 1968.

Hájek, J. and Šidák, F. *Theory of rank tests.* Prague: Academic Press and Academia, 1967.

Harter, H. L. Expected values of normal order statistics. *Biometrika,* 1961, 48, 151–165.

Hoeffding, W. 'Optimum' nonparametric tests. *Proceedings of Second Berkeley Symposium of Mathematical Statistics and Probability.* Berkeley and Los Angeles: University of California Press, 1951.

International Business Machine Corporation. Random number generation and testing. Reference Manual c 20-8011, 1959.

International Business Machine Corporation. System/360, Scientific Subroutine Package. Programmer's Manual H20-0205-3, 77, 1968.

Kendall, M. G. and Stuart, A. *The advanced theory of statistics.* London: Griffin and Company, 1961.

Kruskal, W. H. A nonparametric test for the several sample problem. *Annals of Mathematical Statistics,* 1952, 23, 525–540.

Kruskal, W. H. and Wallis, W. A. Use of ranks in one criterion variance analysis. *Journal of Ameriacn Statistics Association,* 1952, 47, 583–621.

Lehman, R. S. and Bailey, D. E. *Digital computing.* New York: Wiley, 1968.

McSweeney, M. and Penfield, D. The normal scores test for the c-sample problem. *British Journal of Mathematical and Statistical Psychology,* 1969, 22, 177–192.

Pearson, E. S. and Hartley, H. O. Chart of the power function for analysis of variance tests, derived from the non-central *F*-distribution. *Biometrika,* 1958, 38, 112–130.

Puri, M. L. Asymptotic efficiency of a class of c-sample tests. *Annals of Mathematical Statisitcs,* 1964, 35, 102–121.

Terry, M. E. Some rank order tests which are most powerful against specific parametric alternatives. *Annals of Mathematical Statistics,* 1952, 23, 346–366.

Van Der Waerden, B. L. Ein neuer Test für das Problem der zwer Stichproben. *Mathematical Annals,* 1953, 126, 93–107.