EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1984, 44

F- AND H-TEST ASSUMPTIONS REVISITED[1]

KLAUS BOEHNKE[2]
Technische Universität Berlin

The effects of some restraints not included in the classical assumptions of the *F*- and *H*-test (e.g., correlation of mean and sample size) were examined in a simulation design of 1000 samples per condition. Also simulated was a situation in which two assumptions were not met simultaneously. The major conclusions were: *H* was not an appropriate alternative for *F* with samples of $N \geqq 20$; in all cases of unequal variances combined with unequal sample sizes *H* should be applied; and neither *H* nor *F* should be applied if more than one assumption of either test is not met.

TRADITIONALLY, to test several samples simultaneously for differences in location Fisher's *F*-test is applied. For example, a typical application of the *F*-statistic would be to compare the effects of several different teaching methods on the mean achievement of students as measured by a standardized achievement test.

F-tests in one-way ANOVA designs like this are based on three major assumptions: normality of the population distributions, homogeneity of the population variances, and additivity of treatment and error components. According to most textbooks, the *F*-test is fairly robust to violations of these assumptions (Bortz, 1979; Clauss and Ebner, 1971; Weber, 1980). However, if these assumptions are not met, Kruskal-Wallis' *H*-test is a possible alternative. According to Lienert (1973), there are also three assumptions of the *H*-test:

[1] The study on which this paper is based was conducted as part of the thesis requirements of the diploma thesis in psychology—Advisor J. Bortz.

[2] The author wishes to express his thankfulness to J. R. Nesselroade, R. K. Silbereisen, and especially Ellen Skinner and Amy J. Michéle, who gave invaluable assistance in preparing the manuscript.

continuous measurement of the dependent variable; homomerity, i.e., treatments are allowed to effect only the means but not the shape of population distributions; and additivity of treatment and error components. The basic issue is: when parametric assumptions are *not* met, under what conditions should the *H*-test be applied, and under what conditions should the *F*-test be applied.

The present study focused on two major aspects of this question: (a) to assess to what extent certain conditions (sample size, correlation of sample size and sample mean, etc.) not usually included in the above mentioned classical assumptions influence the accuracy of the tests; (b) to assess the results when two of the classical assumptions are not met simultaneously. Although much research has been done on the comparison of *F*- and *H*-tests, (cf. Keselman and Rogan, 1977; Scheirer, Hare, and Schmitt, 1978), previous studies have rarely concentrated on the questions raised above; moreover their conclusions have been somewhat contradictory (compare Bradley, 1968, vs. Smith, Note 1, concerning e.g., the influence of sample size on the power of the respective test).

Method

To answer these questions, a 1000-samples-per-condition simulation was performed using the SPSS random number generator (Nie et al. 1975). Three conditions were systematically varied. The conditions were chosen to enable cross-validation with results from studies by Bradley (1968), Keselman and Rogan (1977), and Smith (Note 1) concerned primarily with the influence of sample size on the power of the *H*- as well as the *F*-test. Furthermore results not provided by those studies, namely the influence of two parametric assumptions not being met simultaneously and of neither parametric nor non-parametric assumptions being met, were desired.

The three conditions can be described as follows: First were conditions under which all the parametric assumptions had been met. Included here were varying sub-conditions, i.e., number of treatments (k), level of nominal α, and number of subjects under each treatment (n_j); second were conditions under which all the parametric assumptions had *not* been met but non-parametric assumptions had been. Again, included here were varying sub-conditions, i.e., non-normal distributions of different shapes combined with homogeneous variances (σ^2), normal distributions combined with unequal variances (σ^2), and non-normal distributions of different shape combined with non-homogeneous variances (σ^2). Within these sub-conditions, number of subjects (n_j) under each treatment

(k), correlation (r) between sample size (n_j) and population variance (σ^2), correlation (r) between sample size (n_j) and population means (μ) additionally varied. Third, were conditions under which parametric as well as non-parametric assumptions had not been met. Included again here were varying sub-conditions, i.e., non-continuous dependent variables from equally shaped distributions, and non-continuous dependent variables from non-homomeric distributions.

Each of the conditions was simulated under the null as well as under the alternative hypothesis. Under the null hypothesis, means of all populations were chosen to be $\mu = 4$. Under the alternative hypothesis, means were chosen to be $\mu_1 = 2$, $\mu_2 = 3$, $\mu_3 = 7$, $\mu_4 = 4$, and $\mu_5 = 4$. This selection of means is based on an example for one-way ANOVA given by Bortz (1979). Population variances were then adjusted so that the F-test would have an expected power of .6 according to the tables generated by Cohen (1977). Cohen's power coefficient was chosen because it is the most extensively tabulated coefficient available.

Results

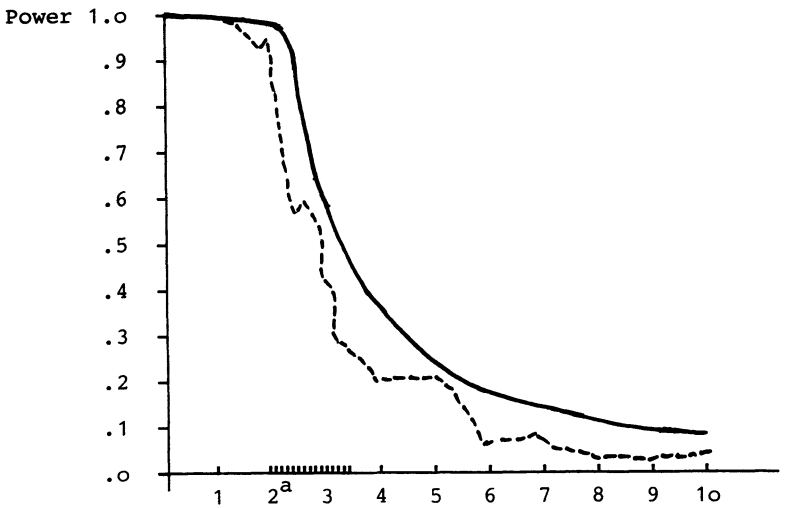
Condition 1: All parametric assumptions met

Under condition 1, the simulation showed that the H-test is less powerful and more conservative than the F-test under all sub-conditions tested. The influence of sample size (n_j) may be seen in Table 1. As can be seen, in general as sample size increases, relative efficiency also increases. (See Footnote e, Table 1 for qualifications to this conclusion.)

TABLE 1
Empirical Power of H and F at Different Sample Sizes

n_j[a]	$1 - \beta_H$[b]	$1 - \beta_F$[c]	$1 - \beta_H/1 - \beta_F$[d]
3	.485	.600	.808
4	.554	.604	.917
5[e]	.550	.588	.935
6	.516	.577	.894
8	.541	.591	.915
11	.554	.600	.923
31	.560	.594	.943
51	.551	.578	.953
250	.614	.637	.964
average	.547	.597	.917

[a] Number of treatments held constant at $k = 3$.
[b] Empirical power of H.
[c] Empirical power of F.
[d] Local relative efficiency of H compared to F.
[e] Exact critical values were used for $n_j \leqq 5$, χ^2-approximations were used for $n_j > 5$.



$1 - \beta_F = \frac{}{}$ [b]

$(1 - \beta_{H(emp)}) \cdot (1 - \beta_{F(Cohen)})/(1 - \beta_{F(emp)}) = ----$ [c]

[a] Between $\sigma = 2.0$ and $\sigma = 3.5$ first decimals were probed in addition to full numbers only probed otherwise.

[b] Power of F-test according to the Tables of Cohen (1977).

[c] Power of the H-test standardized by the ratio of the expected power of the F-test and the empirical power of the F-test.

Figure 1. Power curves of H and F.

The power curves of the H- and F-tests are presented in Figure 1. Examination shows that the power curves of H and F are not parallel. As can be seen, the distance between these two curves varies (unsystematically) as a function of the size of the standard deviation of the population. This was expected because the H-statistic has a discrete distribution. In order to obtain power curves for H and F, the standard procedure for obtaining an alternative hypothesis with the expected power of .6 was applied and then population standard deviations (σ) were systematically varied. Empirical power values of H and F were afterwards standardized on the basis of Cohen's expected power coefficients, i.e., the empirical power values of H were divided by the empirical power values of F

TABLE 2
Empirical Type I Errors of H and F for $p \leqq .05$

n_j	α_H[a]	α_F[b]
3	.048	.046
4	.053	.057
5[c]	.052	.057
6	.047	.054
8	.052	.053
11	.035	.039
31	.047	.045
51	.052	.052
250	.048	.049
average	.0482	.0502

[a] Empirical percentage of false rejections of the null hypothesis by the *H*-test.
[b] Empirical percentage of false rejections of the null hypothesis by the *F*-test.
[c] Exact critical values were used for $n_j \leqq 5$, χ^2-approximations were used for $n_j > 5$.

and multiplied by the expected power of *F* according to Cohen (1977).

Under the null hypothesis, *H* was conservative under all subconditions. Table 2 shows the percentage of false rejections of the null hypothesis by *H* and *F*, on a nominal α-level of $p \leqq .05$ with $k = 3$ and n_j varying. As can be seen, on the average, the empirical Type I error of the *F*-test is closer to the expected Type I error than is that of the *H*-test.

Three additional results obtained under condition 1 were somewhat unexpected. First, in a substantial number of samples (4% of all samples drawn), the false null hypothesis was rejected by the *H*-test, while at the same time it was falsely retained by the *F*-test; an example of this kind of sample is given in Table 3.

TABLE 3
Parametric Sample for Which H Rejects While F Retains the False Null Hypothesis

	A_1	A_2	A_3	A_4
	2.000	3.000	10.618	.523
	2.537	5.688	7.000	2.260
	.389	.675	6.322	4.000
	6.121	5.688	7.905	6.146
	−1.047	−.051	3.155	7.071
$\sum_{i=1}^{n_j} x_{ij}$	10.000	15.000	35.000	20.000
$\bar{x}_j$	2.000	3.000	7.000	4.000

Second, with large samples ($n_j \geqq 250$), the empirical power of H is consistently higher than the power expected on the basis of asymptotic theory (see Pitman, Note 2): In this case, it is $1 - \beta = .97$ as compared to an expected $1 - \beta = .955$. Third, with unequal sample sizes under each treatment, the power of both H and F is considerably reduced especially when population means are negatively correlated with sample size. This loss is as high as 65% under the most extreme conditions.

Condition 2: Parametric assumptions not met and non-parametric assumptions met

Results under condition 2 of the simulation are presented by sub-condition.[3]

Under the first sub-condition, namely of homogeneous variances with non-normal distributions, three results were obtained: (a) the shape of the distribution has almost no effect on Type I and Type II errors as long as the distribution is symmetrical, (b) the relative efficiency of H compared to F does not increase under a rectangular distribution, as would have been predicted by asymptotic theory; and (c) the H-test gains considerably in relative efficiency with asymmetric distributions; it even exceeds 1.0 if the distribution is log-normal.

Under the second sub-condition, namely of the normal distribution with unequal variances, the F-test is slightly liberal, even with equal sample sizes, i.e., .08 at a nominal α-level of .05. In addition, the H-test leads to fewer false decisions under the null as well as under the alternative hypothesis when variances and sample sizes are negatively correlated. Finally, if population means and variances are strongly correlated, the power of both H and F is considerably reduced, e.g., from expected $1 - \beta_F = .6$ and $1 - \beta_H = .57$ to .12 and .09 respectively.

Under the third sub-condition, namely of non-normal distributions with unequal variances, the simulation shows that neither the F- nor the H-test should be used, because under certain conditions (e.g., an inverted j-shaped distribution[4] combined with slightly unequal variances) both tests lead to approximately 60% false decisions under the null hypothesis.

---

[3] Detailed Tables are omitted here, they may be obtained from the author, Klaus Boehnke, Technische Universität Berlin, Institut für Psychologie, Dovestr. 1-5, D 1000 Berlin 10, Federal Republic of Germany.

[4] A distribution of the absolute scores of the normal distribution.

Condition 3: Neither parametric nor non-parametric assumptions met

Again, the results of this part of the simulation are presented by sub-condition.

Under the sub-condition of discontinuous but homomeric distributions: (a) discontinuity has hardly any influence whatsoever as long as the distribution is symmetrical; and (b) H gains considerably in relative efficiency if applied to a Poisson-distribution. H retains power of .95 even at very small sample sizes (e.g., $k = 4$, $n_j = 5$) compared to .85 under perfectly parametric conditions with the same sample size.

Under the sub-condition of discontinuous and non-homomeric populations, the simulation again shows that both tests should not be applied in cases where more than one assumption is not met. However, this is true for different reasons. The F-test loses its power almost completely; at the same time under the null hypothesis, frequencies of correct decisions come close to the nominal α. H, on the other hand, retains its power to an extent that comes close to that expected from asymptotic theory, but results in about 50% false decisions when the null hypothesis is true.

Discussion

The effects of some restraints not included in the classical assumptions of the F- and H-test, (e.g., sample size, correlation of variance or mean with sample size) were examined in a simulation design. As has been discussed previously (Illers, Note 3), such parameters may have a major impact on the accuracy of non-parametric tests, especially in comparison to their parametric analogs. Also of interest was the accuracy of F- and H-tests under conditions where two assumptions of either of the tests were not met simultaneously.

Three major conclusions can be drawn from the results of this study.

1. With sample sizes of $N\ (k \cdot n_j) \leqq 20$, the H-test should not be used as a regular alternative for the F-test, even if parametric assumptions are not completely met.

2. The H-test should be used instead of the F-test if unequal variances are combined with unequal sample sizes, especially if both are negatively correlated.

3. Neither of the two tests should be applied if more than one assumption of either test is not met.

The results from which conclusion 1 is derived directly oppose the position taken by Bradley (1968) who claimed that non-parametric tests are only slightly less efficient with extremely small sample sizes and lose power with increasing sample sizes. Smith (Note 1) claimed the opposite and the present results definitely support her view. The results underlying the second conclusion are consistent with results of studies by Box (1954) and by Keselman and colleagues in recent years (e.g., Keselman and Rogan, 1977). The third conclusion finds no counterpart in the literature. In agreement with Bradley (1968) it is concluded that robustness of the *F*-test is a "myth" if more than one assumption of the test is not met; these results suggest that it is also a myth for the *H*-test if more than one assumption of that test is not met.

The reasons for the occurrence of samples like the one in Table 3 still remain unclear. This kind of sample has already been referred to by Games (1971). More information could be gained by an analytical study of the rejection areas of both the *H*- and the *F*-test. The present study, being purely empirical, could not attempt this analysis. Higher order ANOVA designs as well were not within the scope of the present study but are also an area for future additional research.

Whatever the results of further research, it is clear that a more careful examination of the parametric *as well as* the non-parametric assumptions is needed if serious errors in future research are to be avoided.

Reference Notes

1. Smith, M. A. K. *A Monte-Carlo development of constant power functions of the Kruskal-Wallis H-statistic for sampling from selected distributions.* Unpublished educational dissertation, Memphis State University, 1976.
2. Pitman, E. J. G. *Notes on non-parametric statistical inference.* Unpublished manuscript, Columbia University, 1947.
3. Illers, W. *Der Mann-Whithney-Wilcoxon-U-Test—Untersuchungen zur Robustheit gegen Streuungsungleichheiten bestimmter nicht-symmetrischer Verteilungen.* Unveröffentlichte Diplomarbeit, Universität Karlsruhe, 1977.

REFERENCES

Bortz, J. (1979). *Lehrbuch der Statistik.* Berlin: Springer.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in study of analysis of variance problems: I—Effect of inequality of variance on the one-way classification. *Annals of Mathematical Statistics*, 25, 290–302.

Bradley, J. V. (1968). *Distribution-free statistical tests*. Englewood Cliffs, NJ: Prentice Hall.

Clauß, G. and Ebner, H. (1971). *Grundlagen der Statistik für Psychologen, Pädagogen und Soziologen*. Frankfurt a. M.: Deutsch.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Games, P. (1971). Multiple comparisons of means. *American Educational Research Journal*, 8, 531ff.

Keselman, H. J. and Rogan, J. (1977). An evaluation of some nonparametric and parametric tests for location equality. *British Journal of Mathematical and Statistical Psychology*, 30, 125–133.

Lienert, G.-A. (1973). *Verteilungsfreie Methoden der Biostatistik*. Meisenheim am Glan: Hain.

Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., and Bent, D. H. (1975). *SPSS—Statistical Package for the Social Sciences*. New York: McGraw-Hill.

Scheirer, C. J., Hare, N., and Schmitt, J. C. (1978). Effect of some violations of the normality assumption on the power of the Kruskal-Wallis-H-test and the ANOVA in the two sample equal *n* case. *Catalog of Selected Documents in Psychology*, 8 (2), 1626.

Weber, E. (1980). *Grundriß der biologischen Statistik*. Jena: Gustav Fischer.