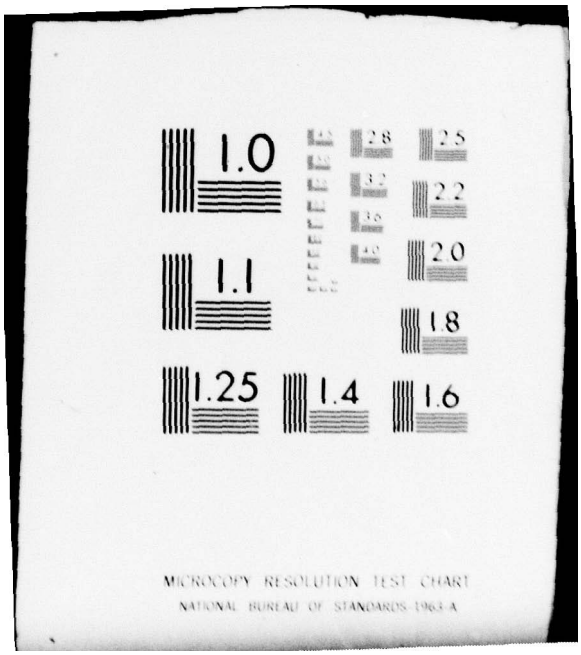
AD-A074 099 WALTER REED ARMY INST OF RESEARCH WASHINGTON DC F/G 12/1
ARE NON-PARAMETRIC TESTS DISTRIBUTION-FREE,(U)
1961 A LUBIN

UNCLASSIFIED NL

| OF |
AD
A074099

END
DATE
FILMED
10-79
DDC


1.0
1.1
1.25
1.4
1.6
2.8
3.2
3.6
4.0
2.5
2.2
2.0
1.8
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

Are Non-Parametric Tests Distribution-Free?

Ardie Lubin

Division of Neuropsychiatry
Walter Reed Army Institute of Research
Walter Reed Army Medical Center
Washington 12, D. C.

Introduction

An increasing use is being made of non-parametric tests, i.e., tests of significance which are not based on the usual assumption that all distributions are normal. Many psychologists have assumed that when non-normality or heterogenous variance is encountered in the data, then analysis of variance by ranks, the Mann-Whitney U Test (1947), Festinger's d (1946), or some similar non-parametric technique is appropriate.

The purpose of this paper is to point out that non-parametric tests also have basic assumptions which must be met, and that often the factors which bar the use of the parametric tests also violate the assumptions underlying most of the non-parametric tests. I am indebted to Professor Leon Festinger for calling this to my attention (personal communication).

Almost all non-parametric tests (in particular, those which use a rank-order procedure) assume that all distributions are identical. This means that if the variances are homogeneous, or third moments are not equal, etc., these non-parametric tests are inappropriate for testing the equality of means.

To fix our ideas, the following three non-parametric tests will be considered:

(1) The Kruskal-Wallis H Test (one-way analysis of variance by ranks) (1952)

(2) The Kendall W Test for differences in correlated means. (1948)

(3) The test for difference in medians. (Westenberg, 1948)

What are the assumptions underlying these procedures? When can they

DEPARTMENT OF THE ARMY
ARI FIELD UNIT, BENNING
U. S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES
P.O. BOX 2086, FORT BENNING, GEORGIA 31905

PERI-IJ

8 August 1979

SUBJECT: Shipment of Documents

Defense Documentation Center
Cameron Station
Alexandria, VA 22314
ATTN: Selection & Cataloging

The Documents in these shipments are approved for public release. The distribution is unlimited.

FOR THE CHIEF:

ALEXANDER NICOLINI
Major, Infantry
R&D Coordinator

appropriately be substituted for the usual parametric tests?

The Kruskal-Wallis H Test

Let us suppose that we have drawn c samples from the population and we wish to test the hypothesis that all c means are equal. The first step is to rank all N observations without regard to the sample from which they come. Then if there are no ties, compute

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

where c = the number of samples

n_i = the number of observations in the i^{th} sample

N = the total number of observations

R_i = the sum of the ranks in the i^{th} sample.

Kruskal and Wallis (1952, page 586) state that "If the samples come from identical continuous populations and the n_i are not too small..." H is distributed as chi-square with c-1 degrees of freedom. We may disregard the assumption of a large n_i, for when n_i is too small, it simply means that special distribution tables must be consulted. H is still a valid measure of the discrepancy between means. Later they state that "if the data are matched H is not appropriate..." and Friedman's chi-square test (or equivalently Kendall's W) should be used.

There then, are the two fundamental assumptions underlying the use of the H test:

(1) The samples come from identical continuous populations.

(2) The observations are independent of one another.

The first assumption means that the samples may be drawn from populations with a non-normal distribution, but all samples must be drawn from populations with the same non-normal distribution. For example, if we have drawn one sample from a population whose distribution has a positive skew and a second sample from a population whose distribution has a negative skew, the H test cannot be applied to see if the two sample means differ significantly.

The second assumption is the usual random sampling requirement of statistical independence. In particular, observations should be drawn so that no systematic way of matching the observations will produce a correlation (linear or non-linear) between the matched observations. For example, if we have obtained treatment scores A and B from the same individual, the H test cannot be used to test the difference between the means of Treatment A and Treatment B. Obviously, this independence assumption cannot be satisfied if the number of test scores exceeds the number of subjects. When there are two or more scores per individual the scores will, in general, be dependent upon one another, in a probability sense.

This, however, is the same as the usual analysis-of-variance assumption of independence of observations. Here too, the samples must be independently drawn. The use of three or more per subject will, in general, force the use of a much more complicated technique--the multi-variate analysis of variance (Rao, 1951, pp. 239-246; G. E. P. Box, 1954, pp. 484-498).

It has been customary for some workers to resort to non-parametric tests when the variances are heterogeneous. Obviously, the "identical distribution" assumption bars the use of the H test in this case. If we allow two samples to have very different variances, it is easy to construct an example where H is significant, even when the two means are identical!

In Table 1 consider the symmetric distribution of Treatment A and the skewed distribution of Treatment B. Here the H test is significant at the 1 percent level although the means are equal.

Table 1 about here

It must not be thought that the H test is simply inaccurate in this case and that some other non-parametric test would not give a significant difference. On the contrary the H test gives essentially the same results as any other test based on rank-order, runs, randomization, etc.

The reason for this is that all non-parametric rank-order tests seem to be based on assumption 1, that samples come from identical continuous populations. There do not seem to be any rank-order tests based on the notion that the samples come from several different populations (with respect to variance, skewness, etc.), where the population means are equal.

Significance tests do not always have to conform to their assumptions to be useful. For example, the t test assumes that the two samples are drawn from distributions with a common variance. Yet Welch (1938) was able to show that as long as the two samples were of equal size, one variance could be ten times the other without seriously disturbing the 5 percent level of t. The term "robust" has been used by Box (1954) to denote procedures like the t test which are maximally sensitive to differences in the crucial parameters being examined (such as the sample averages) but are fairly insensitive to differences in some of the irrelevant parameters (such as variances, fourth moments, etc.)

How "robust" is the H test? Under what conditions can the assumption of identical population distributions be ignored? Kruskal and Wallis conjecture that if two populations have symmetric distributions, then differences in variability will not affect the significance levels of the

H test very much, (1952, pp. 599-600).

The results of our inquiry into the H test can be summarized as follows:

(1) A significant value of H indicates that the populations differ, but the significance may be due to differences in variances, third moments, etc., and not necessarily to differences in the means.

(2) The H test for difference in means is probably robust when population distributions are symmetric.

Kendall's W Test for Differences in Correlated Means

It has already been mentioned that Kendall's W is appropriate when the data are matched. Suppose that n treatments have been applied to each of m subjects. A score matrix would result with m rows and n columns.

A row would represent the n treatment scores of one subject. These n treatment scores can be ranked. There will be m such rankings. If the treatment that is highest for one subject is highest for all subjects there will tend to be a significantly positive rank-order correlation between any two subjects.

In general, if one treatment tends to be consistently higher than another, the average rank order correlation among all m subjects will be significant. Kendall and Smith (1939) proposed a statistic, W, which is a monotonic function of the average Spearman rank-order coefficient. W equals one if the m rankings agree perfectly, and is zero or near-zero when the average rank-order coefficient is as low as possible.

Friedman (1937) had previously suggested a statistic, χ^2_r, roughly distributed as chi-square for large samples, which is also a monotonic function of the average rank-order correlation. Since the rank order of the n treatments for any subject is independent of the correlations

between the treatment scores, the Friedman chi-square statistic (or the Kendall W) can be used when the treatment means are correlated.

The fundamental assumptions of the Friedman chi-square test are:

(1) The n treatment samples come from identical populations

(2) The n rows of observations are independent of one another.

Essentially these are the same restrictions as those of the H test, and seem to lead to similar difficulties. For instance, if the c treatments produce unequal variances, a significant chi-square can be obtained even when the n means are identical. If Table 1 is analyzed as a Friedman 9 by 2 score matrix (with two scores per individual) then

$$\chi^2_r = \frac{12_i \sum_{i=1}^{c} \left[R_i - \frac{1}{2}(c+1)\right]^2}{rc\,(c+1)} = 5.444 \text{ with } (c-1)=1 \text{ degree of freedom.} \qquad (2)$$

This Friedman chi-square is significant at the .02 level. (The chi-square distribution is generally not an accurate approximation to the distribution of Friedman's chi-square for c = 2 and R. A. Fisher's exact binominal sign-test should be used. In this case the sign test rejects the null hypothesis of identical means at the .04 level.)

No investigations have been made of the robustness of Friedman's test. It seems likely that it is robust under the same conditions as the H test, i.e., when the n treatment distributions are symmetrical.

The Test for Differences in Medians

The mathematical statistician sometimes follows the adage "When you can't be near the girl you love, then you love the girl you're near." If the question about differences between means can't be answered satisfactorily, let's ask some other related question which can be answered. When the distributions are not symmetric, the median and mode

no longer coincide and it becomes a matter of preference, utility, and convenience as to which measure of central tendency should be used. Why not devise a test for the differences in medians? After all, if the distributions are badly skewed, most statisticians will be using the median, rather than the mean, for interpretative purposes.

Westenberg (1948) and Mood and Brown (1950) have discussed such a test for the differences in medians.

The basic step is to arrange the observations from the c samples in order of size. Then the $\frac{(N+1)^{th}}{2}$ observation is chosen as the best estimate of the population median. Now the c samples can be compared in terms of the numbers above the estimated population median. This results in a 2 by c contingency table. If the c samples differ markedly in the proportion of cases above the population median then chi-square, or the more exact multinomial technique, will show the significance of these differences.

For example in Table 1, the estimated population median is 158.5, the median for Treatment A is 155, and the median for Treatment B is 163. Table 2 shows the resulting 2 by c contingency table. Both the usual chi-square and Fisher's exact 2 by 2 test show significance at the .01 level.

Table 2 about here

The median test can be derived from two basic assumptions:

(1) The samples come from continuous populations with a common median

(2) The observations are independent of one another

It follows that for the i^{th} sample, the number of cases above the population median is distributed as the binomial

$(1/2+1/2\,x\,)$ (3)

where

the exponent of χ is the number above the population median in the i^{th} sample

n_i is the number of observations in the i^{th} sample

One of the chief difficulties is that the common population median is estimated from the total distribution of the samples. It does not seem to be known under what conditions such an estimate will be consistent and efficient. This in turn means that generally it is not known whether the median is biased or what its power is.

The chief advantage is that the assumption of identical population distributions is dropped. The samples can be drawn from any c populations and tested for a common median provided only that the observations are sampled in a random and independent manner, as usual.

The median test has the advantage over the H test that differences in variability, skewness, etc., apparently do not invalidate the procedure. Any set of diversely shaped distributions with a common median will have the binomial distribution of equation (3).

If the term "distribution-free" were reserved for those methods which do not assume the equality of irrelevant parameters (such as the variance, skewness, fourth moment, etc.) then presumably much misapprehension about the nature of these tests could be avoided. "Distribution-free" seems an appropriate adjective for a technique which assumes equality only for those crucial parameters being tested for equality.

Those methods which assume identical but non-normal distribution could possibly best be described as "normal-distribution-free," "non-normal," or "identical population" tests. At any rate, it seems rather inappropriate to lable as "non-parametric," tests which are completely dependent upon the values of the non-crucial parameters in the assumed non-normal population.

Recommendations

One purpose of this note was to set out the conditions when the non-parametric tests should and should not be applied.

If we are interested in differences between means, the t test and F ratio are appropriate when the distribution within the populations being sampled are normal and have homogeneous variance. These tests are "robust," i.e., they have approximately the tabled distributions even when these assumptions are not met exactly.

Case I--The population distributions are normal with differing variances.

When the distributions are normal and the number in each sample is large, then heterogeneous variances will not create much disturbance. If the sample sizes are small but equal, then heterogeneous variances will still make little difference, although for accuracy, a correction should be made reducing the degrees of freedom (G. E. P. Box, 1954, p. 300).

But if the sample sizes are small and unequal then corrections have to be applied to the F ratio as well as the degrees of freedom. Here the approximation as well as being more complex, tends to be less accurate then in the case of equal groups.

It is at this point that rank-order procedures like the H test can be applied. Although, strictly speaking, the heterogenous variances violate the assumption of identical population distributions, the H test is probably robust because the distributions are symmetric. Of course the median test could be used, for here the median coincides with the mean, but it is almost certain that the rank-order tests will be much more sensitive than any dichotomizing procedure.

Case II--The population distributions are identical but non-normal.

The straightforward, but painful, procedure in this case would be

to ascertain the nature of the common distribution and either devise an exact test or transform to the normal distribution.

If the population distribution is symmetrical, not much is gained in the process. With large samples, the tabled percentage points of the t and F ratios remain good approximations except in pathological cases such as the Cauchy distribution. With small sample sizes and symmetrical non-normal population distributions, the H test would probably be a good choice.

When the common population distribution is skewed, fairly large samples are needed before the t and F ratios become good approximations to the tabled percentage points. The H test is probably a safer test to use than the t or the F ratio for the usual sample sizes. The median test is not recommended. Since the distributions are identical, the medians will differ between groups in the same direction as the means. The more sensitive H test will tend to pick up small differences which the less powerful median test is unable to discriminate.

Case III--The population distributions are non-identical and non-normal.

Neither the t test nor the H test is strictly appropriate here. If the distributions are symmetric, the H test will probably be sensitive to differences in the means. If the distributions are not symmetric and have unequal variances, the median test can be used if the median is felt to be a useful measure of the central tendency.

Case IV--The observations are dependent.

Most often this is encountered when repeated measurements are taken for each subject. If the n scores for the m subjects form a multivariate normal distribution, then multivariate analysis-of-variance (Rao, 1952, pp. 240-244) is the most appropriate and sensitive test for differences between the n correlated treatment means.

Some psychological statisticians such as Edwards (1950), Lindquist (1940), and McNemar (1949) have recommended that the m by n score matrix should be treated by a two-way analysis of variance, using the interaction as the error term. For the "between treatments" comparison this will be exactly correct when n=2, and approximately so when n> 2 if all intercorrelations between the n treatments are equal (Box, 1954, p. 489). However when the correlations between treatments are unequal, it is likely that the F ratio for the treatment effect will be seriously biased. The "between-subjects" F ratio is even more seriously affected.

When the n treatment distributions are non-normal but identical, then Kendall's W is appropriate. The sensitivity of W increases very swiftly as n increases and somewhat more slowly as m increases.

So far as I know, no test has been proposed for the case when the treatment distributions are <u>not</u> assumed to be identical.

Summary

Most non-parametric rank-order tests such as the Kruskal-Wallis H test, Friedman's test for differences in correlated means, the Mann-Whitney test, etc., assume that samples are drawn from <u>identical</u> population distributions. Therefore, when heterogeneity of variance, differences in skewness, etc., are found, these non-parametric tests are <u>not</u> theoretically appropriate substitutes for the t test or F ratio. Strictly speaking, they should only be used with identical non-normal populations. However, if the population distributions are symmetric, the H test is probably a sensitive test of the differences in sample means.

The median test can be used without the assumption of identical population distributions, and therefore can always be applied. But the median is not always an appropriate measure of central tendency. At present,

very little is known of the bias and power of the median test when samples are drawn from a specified set of non-identical distributions.

It is suggested that the term "distribution-free" be reserved for those significance tests which do not assume that all populations are identical.

References

1. Box, G. E. P. (1954) Some theorems on quadratic forms applied in the study of analysis of variance problems: I. Effect of inequality of variance in the one-way classification. Ann. math. Statist., 25, 290-302. II. Effects of inequality of variance and of correlation between errors in the two-way classification. Ann. math. Statist., 25, 484-498.

2. Friedman, M. (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J. Am. statist. Assoc., 32, 675-701.

3. Kendall, M. G. and Smith, B. B. (1939) The problem of m rankings. Am. math. Statist., 10, 275-287.

4. Kruskal, W. H. and Wallis, W. A. (1952) Use of ranks in one-criterion variance analysis. J. Am. Statist. Assoc., 47, 583-621.

5. Mood, A. M. (1950) Introduction to the theory of statistics, New York, McGraw-Hill Book Company.

6. Rao, C. R. (1952) Advanced statistical methods in biometric research. New York, John Wiley and Sons.

7. Welch, B. L. (1938) The significance of the difference between two means when the population variances are unequal. Biometrika, 29, 350.

8. Westenberg, J. (1948) Significance test for median and interquartile range in samples from continuous populations of any form. Proceedings, Koninklijke Nederlandsche Akademie van Wetenschappen, 51, 252-261.

9. Kendall, M. G. (1948) Rank correlation methods, London, Charles Griffin and Company.

10. McNemar, Q. (1949) Psychological statistics. New York: John Wiley and Sons.

11. Lindquist, E. F. (1940) Statistical analysis in educational research. Boston: Houghton Mifflin.

12. Edwards, A. L. (1950) Experimental design in psychological research. New York: Rinehart and Company, Inc.

13. Festinger, L. (1946) The significance of differences between means without reference to the frequency distribution function. Psychometrika, 11, 97-105.

14. Mann, H. B. and Whitney, D. R. (1947) On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Statist., 10, 275-287.

Table 1

Example of a Significant H When Means are Identical

	Treatment A		Treatment B		N=18
	Score	Rank	Score	Rank	
	151	2	160	11	
	152	3	161	12	
	153	4	163	13	
	154	5	163	14	
	155	6	164	15	
	156	7	165	16	
	157	8	166	17	
	158	9	167	18	
	159	10	87	1	
Sum	1395	54			

$V = 1.796$

$$H = \frac{12}{18(19)} \quad \frac{54^2 + 117^2}{9} \quad -3(19) = 7.737$$

$$M = \frac{18^3 - 2(9^3)}{18(19)} = 12.789$$

$$F = \frac{7.737\ (11.789)}{5.052} = 18.055$$ Significant at the 1 percent level

$$f_1 = \frac{2\ (11.789-1.796)}{12.789\ (1.796)} = .870$$

$f_2 = 11.789\ (.870) = 10.256$

Table 2

	Treatment A	Treatment B	Sum
Cases above 158.5	1	8	9
Bases below 158.5	8	1	9
Sum	9	9	18

$$\chi^2_c = \frac{(|8(8)-1(1)|-\frac{18}{2})^2 18}{9(9)9(9)} = 12$$

Significant at the .01 level