DOCUMENT RESUME

ED 065 559 TM 001 705

AUTHOR Penfield, Douglas A.
TITLE A Comparison of Some Nonparametric Tests for Scale.
PUB DATE Apr 72
NOTE 16p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, Illinois, April 1972)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Comparative Analysis; *Nonparametric Statistics; *Statistical Studies; *Test Construction; *Tests
IDENTIFIERS Mood Test; Normal Scores Test; Siegel Tukey Test

ABSTRACT

Three different nonparametric tests for scale--the Siegel-Tukey (S-T), the Mood (M), and the Normal Scores (NS)--are compared in order to contrast varying methods of scale test development and usage. Procedures for developing the three scale tests are discussed, and two examples of the use of each test in solving the same problem are given. From the results obtained in the two examples, it is apparent that all the three tests tend to give equivalent answers. (DB)

A Comparison of Smme Nonparametric Tests for Scale.

Douglas A. Penfield
Rutgers University

Paper presented at the annual meeting of the American Educational Research Association on April 4, 1972

Introduction

Behavioral scientists frequently find it of interest to determine whether random samples have been drawn from populations having the same variance. For the two sample problem such a hypothesis regarding the equality of variance (scale) is tested by forming a ratio between the two sample variances and then checking for significance using the F distribution. Most researchers fail to realize that this test of hypothesis is only appropriate when the underlying distribution of scores is approximately normally distributed. Whereas z and t tests are robust to violation of the normality assumption, especially when N is large, X^2 and F tests are extremely sensitive to nonnormality of sample data. When this normality assumption cannot be met, the researcher is forced to search out a nonparametric test to investigate the hypothesis of interest.

A number of nonparametric tests have been proposed as possible alternatives to the parametric F-test. Under specified conditions each in its own way would be considered to be a "good" test. Since nonparametric tests generally require a substitution for the actual data, a primary distinction between these tests is the variable.

being substituted in place of the original scores. In this paper three different nonparametric tests for scale are being compared in order to contrast varying methods of development and usage. These three tests are the Siegel-Tukey (Siegel& Tukey, 1960) test (S-T), the Mood (Mood, 1954) test (M) and the Normal Scores (Capon, 1961) test (NS). There are a number of asymptotically equivalent forms of the normal scores test for scale. The test presented in this paper will use expected normal order statistics in the test development.

Procedures for Developing Scale Tests

Let $X_1, X_2, \ldots, X_n$ be an independent random sample from a distribution F and $Y_1, Y_2, \ldots, Y_m$ be an independent random sample from a distribution G. Combining the n+m=N scores of the two samples and ranking them in ascending order produces the array of ordered scores $V^1 < V^2 < \ldots < V^N$. If ties exist, break them at random. The hypothesis (H_0) to be tested is that the two distributions, F and G, are identical with respect to scale. The alternative hypothesis (H_1) is that the dispersion of scores is different for F and G.

Siegel-Tukey Test

This test replaces the pooled data from the two samples with a reordering of the ranks (i) from 1 to N. To illustrate the ranking procedure, note the following table when N is assumed to be an even number.

Ordered Score	V^1	V^2	V^3	V^4	...	$V^{N/2}$	...	V^{N-3}	V^{N-2}	V^{N-1}	V^N
Rank Replacement	1	4	5	8	...	N	...	7	6	3	2

At the left end of the ordered set of scores, V^1 is assigned

a rank of 1. The test development now requires a move to the extreme right end of the ordered scores where V^N is replaced by the rank 2 and V^{N-1} by the rank 3. Now move back to the left and substitute the ranks 4 for V^2 and 5 for V^3. Operating in pairs, this process is repeated until all ordered scores are replaced by their appropriate ranks. If N is an odd number, throw out the middle score. This will enable the adjacent ranks to sum to the same number and thus achieve a desired symmetry to the test. If there is no difference in scale between the populations from which the two samples are drawn, then the sum of the ranks associated with each sample should be approximately equal. On the other hand, if the score spread is not homogeneous for the two groups, then the rank sum of the sample with the greatest spread will be significantly smaller than the rank sum of the more compressed sample.

Using an indicator variable, Z_i, let $Z_i = 1$ if the i^{th} replacement score is associated with the X sample and $Z_i = 0$ if the i^{th} score is tagged to the Y sample. This indicator variable is useful in setting up the test statistic which is defined as

$$S-T = \sum_{i=1}^{N} iZ_i$$

The null distribution of the S-T test is exactly the same as that of the Wilcoxon test. Thus, Wilcoxon tables (Owen, 1962) can be used to determine the significance of S-T for N<20. Equivalent tables have been developed by Siegel and Tukey (1960). For N≥20 the distribution of S-T approximates a normal distribution with $E(S-T) = n(N+1)/2$ and $Var(S-T) = nm(N+1)/12$. The test statistic becomes

$$Z = \frac{(S-T) - E(S-T)}{\sqrt{Var(S-T)}}$$

Mood Test

The rank procedure developed by Mood is completely analogous to the parametric F-test. Replace the scores in the pooled sample by their corresponding ranks. Knowing that the mean of a set of ranks from 1 to N is (N+1)/2, determine the sum of squared rank deviations about this mean for the X sample. This needs to be done only for the X sample since, when dealing with ranked data, the sum of squared deviations for sample X plus the sum of squared deviations for sample Y adds to the constant $N(N^2-1)/12$. When the samples are of unequal size, it is customary to select the smaller of the two samples for purposes of analysis.

The indicator variable, Z_i, can once again be used to develop the test statistic. Let $Z_i = 1$ if the rank score (i) is associated with the X sample and $Z_i = 0$ otherwise. The test statistic for N<20 is

$$M = \sum_{i=1}^{N} \left(i - \frac{(N+1)}{2}\right)^2 Z_i$$

A large value of M implies that the variability of the X sample is significantly greater than the variability of the Y sample. For small values of M, one draws the opposite conclusion. A table of critical values for N<20 is not available at the present time. Nevertheless, it is possible to derive critical values for a given sample size and alpha level in a short period of time.

When N is greater than 20, values of M approximate a normal distribution with $E(M) = n(N^2-1)/12$ and $Var(M) = nm(N+1)(N^2-4)/180$. In this case the test statistic is

$$Z = \frac{M - E(M)}{\sqrt{Var\ (M)}}$$

Normal Scores Test

This test represents an attempt to reconstruct the F test using expected normal order statistics, $E(V^i)$, in place of the original scores. If V^i is the i^{th} ranked score in the combined sample of size N (since values are conditional upon N), then $E(V^i)$ is the expected value of the score in the i^{th} position assuming the score has come from a standard normal population. The term $E(V^i)$ acts as a distance measure in much the same way that Z scores express relative distance in a normal distribution. The data for the two samples is first pooled and then ranked from low to high. Ranks are then replaced by corresponding $E(V^i)$s. High ranks will have large positive $E(V^i)$s and low ranks will have large negative $E(V^i)$s Those ranks toward the middle of the distribution will have $E(V^i)$s close to zero. The distribution of the $E(V^i)$s for a given N is symmetric about zero. Tables of expected normal order statistics can be found in Owen's (1962) Handbook of Statistical Tables.

The normal scores test is completely analogous to Mood's test except for the fact that expected normal order statistics replace ranks in the test statistic formulation. As was true in the previously mentioned tests, it is customary to work with data from the smaller of the two samples. Using our indicator variable, Z_i, let $Z_i = 1$ if the $E(V^i)$ is associated with the X sample and $Z_i = 0$ if the $E(V^i)$ is linked to the Y sample. Since the mean of the $E(V^i)$s is zero, the test statistic reduces to

$$NS = \sum_{i=1}^{N} (E(V^i))^2 Z_i$$

A table of critical values for N<20 is not available, but the prob-

ability associated with a given value of NS is easily determined.

For N>20, use the large sample approximation to the normal distribution. The mean and variance of NS are given by

$$E(NS) = \frac{n}{N} \sum_{i=1}^{N} (E(V^i))^2 \qquad \text{and}$$

$$Var(NS) = \frac{nm}{N(N-1)} \sum_{i=1}^{N} (E(V^i))^4 - \frac{m}{n(N-1)} (E(NS))^2$$

Example I

An experimenter wishes to determine whether a special training program will influence the abstract reasoning scores of nine year old mentally retarded females. To test his theories he selects 12 (all that were available) nine year old girls who have IQ scores recorded between 65 and 75 on the Stanford Binet. He randomly assigns six of the children to the experimental condition and six to the control. After training the experimental group for a month, the experimenter then gives both groups an abstract reasoning test. The results are as follows:

Experimental	Control
19	20
21	22
27	23
30	23
31	25
35	26

He believes that the scores of the group receiving special training will have a greater dispersion than those of the control group. Is he justified in making this conjecture? Let the probability of a Type I error be 0.05 or less. Data pertinent for analyzing this problem by the procedures introduced previously are presented in

Table I.

(Insert Table I here)

This experiment will be analyzed using the parametric F test, the S-T test, the M test, and the NS test. The hypothesis under test (H_0) is that the dispersion (scale) is the same for both the experimental and control groups. The alternate hypothesis proposed is that score variability is greater for the experimental group.

F Test

To validly use this test, the distribution of scores must resemble a normal curve. There is no way to justify this assumption in Example I. Nevertheless, the F will be computed for comparison purposes.

$$F = \frac{S_e^2}{S_c^2} = \frac{37.77}{4.57} = 8.26$$

If the test is conducted at the 0.05 level, the decision rule would be to reject H_0 if $F \geq F_{5,5}(.95) = 5.05$. Since F = 8.26, the hypothesis (H_0) is rejected. The variance of the scores in the experimental group is significantly larger than the variance of the control group scores.

Siegel-Tukey Test

Using the rank reordered totals from Table I and assuming $Z_i = 1$ for the experimental group and $Z_i = 0$ for the control group, the test statistic is

$$S\text{-}T = \sum_{i=1}^{N} iZ_i = 24$$

Consulting tables in Owen's (1962) Handbook, a value of $S\text{-}T \leq 24$ would occur less than 1% of the time. Therefore, the hypothesis (H_0) is rejected at the 0.05 level.

Mood Test

Performing the analysis on the experimental group scores, the test statistic is

$$M = \sum_{i=1}^{N} \left(i - \frac{N+1}{2}\right)^2 Z_i = 111.5$$

For the M test there are no critical value tables to determine whether or not H_0 is to be rejected. Therefore, the exact probability of occurrence must be computed for a value of M greater than or equal to 111.5.

The total number of ways of dividing 12 subjects into two groups of six each is $C_6^{12} = 924$. Working exclusively with the experimental group, the sum of six squared deviations about the mean greater than or equal to 111.5 can occur in exactly 10 ways. Probability statements regarding possible values of M in the upper tail of the distribution are as follows:

$$P(M \geq 125.5) = 1/924 = 0.001$$

$$P(M \geq 119.5) = 5/924 = 0.005$$

$$P(M \geq 113.5) = 6/924 = 0.006$$

$$P(M \geq 111.5) = 10/924 = 0.011$$

Since the probability of obtaining a value of $M \geq 111.5$ is less than 0.05, reject H_0

Normal Scores Test

For this test the test statistic is

$$NS = \sum_{i=1}^{N} \left(E(V^i)\right)^2 Z_i = 8.099$$

As was true in the case of the M test, no critical value tables exist for NS when N is less than 20. The exact probability of obtaining an NS value greater than or equal to 8.099 is 10/924

= 0.011. This is found in exactly the same manner that the significance level was determined for the M test. Once again H_0 is rejected.

Example II

Fifty first grade boys known to have a low expectation of success on intellectual tasks and a high anxiety about performance in school were randomly assigned to either an arousal or a non-arousal condition for purposes of experimentation. The arousal group was verbally encouraged to try exceedingly hard to accomplish a specified task. The non-arousal group was told not to worry about their performance on the task, simply try to have a good time. The dependent variable of interest was the amount of time (in sec.) they would continue to attempt to solve a difficult puzzle. The results were as follows:

Arousal					Non-Arousal				
139	360	295	360	335	360	49	140	120	162
130	181	91	182	203	131	129	249	38	44
153	360	155	225	71	82	195	47	138	65
124	38	36	203	294	287	54	133	62	220
175	360	360	45	189	131	118	98	131	90

The experimenter felt that score variability would be greater for the arousal group than the non-arousal group. Test the hypothesis that there is no difference in scale between the two groups. Let $\alpha = 0.05$. Data necessary for calculating the large sample approximations to the normal distribution for the scale tests of interest are presented in Table II.

[Insert Table II here]

The experimenter is hypothesizing a directional alternate hypothesis. Therefore, all tests with the exception of the S-T test will be performed with α located in the upper tail of the distribution. In the case of the S-T test, H_o will be rejected for large negative values of the test statistic.

F Test

$$F = \frac{S_A^2}{S_{NA}^2} = \frac{12096.42}{6430.08} = 1.88$$

The hypothesis (H_o) states that there is no significance difference between the variances of the two populations from which the samples are drawn. Reject H_o if $F \geq F_{24,24}(.95) = 1.98$. Since F=1.88 is less than 1.98, fail to reject H_o. The experimenter's conjecture is not borne out. The arousal condition does not produce a greater variation among time scores than the non - arousal condition.

Siegel-Tukey Test

Since N is greater than 20, the normal approximation is appropriate for testing H_o.

$$S\text{-}T = \sum_{i=1}^{N} i\, Z_i = 585 = \text{the sum of the ranks for the arousal group.}$$

$$E(S\text{-}T) = n\frac{(N+1)}{2} = 25\ \frac{(51)}{2} = 637.5$$

$$Var(S\text{-}T) = nm\frac{(N+1)}{12} = 25(25)\frac{(51)}{12} = 2656.25$$

The test statistic is

$$Z = \frac{(S\text{-}T) - E(S\text{-}T)}{\sqrt{Var(S\text{-}T)}} = \frac{585-637.5}{\sqrt{2656.25}} = -1.02$$

For the S-T test H_0 will be rejected in favor of the arousal group when Z is a negative value less than - 1.645. Since Z = -1.02, H_0 is not rejected.

Mood Test

Replace each time score of the combined sample by its rank.

$$M = \sum_{i=1}^{N} \left(i - \frac{N+1}{2}\right)^2 z_i = 5936.25$$

= the sum of the squared rank deviations about the mean for the arousal group.

$$E(M) = n\frac{(N^2-1)}{12} = 25\ \frac{(2499)}{12} = 5206.25$$

$$\text{Var}\ (M) = \frac{nm(N+1)\ (N^2-4)}{180} = 25\,\frac{(25)(51)(2496)}{180}$$

$$= 442000$$

The test statistic is

$$Z = \frac{M-E(M)}{\sqrt{\text{Var}(M)}} = \frac{5936.25 - 5206.25}{\sqrt{442000}}$$

$$= 1.10$$

The decision rule is to reject H_0 when $Z \geq Z_{.95} = 1.645$. Once again H_0 is not rejected. The time score dispersion is not statistically different for the arousal and non-arousal populations.

Normal Scores Test

The procedures are identical to the M test except that expected normal order statistics replace ranks.

$$NS = \sum_{i=1}^{N} (E(V^i))^2 z_i = 29.625 = \text{the sum of the squared expected normal order statistics for the arousal group.}$$

$$E(NS) = \frac{n}{N} \sum_{i=1}^{N} (E(V^i))^2 = \frac{25}{50} (47.434) = 23.72$$

$$\text{Var } (NS) = \frac{nm}{N(N-1)} \sum_{i=1}^{N} (E(V^i))^4 - \frac{m}{n(N-1)} (E(NS))^2$$

$$= \frac{25(25)}{50(49)} (119.56) - \frac{25}{25(49)} (23.72)^2 = 19.02$$

Using the large sample approximation, the test statistic is

$$Z = \frac{NS - E(NS)}{\sqrt{\text{Var } (NS)}} = \frac{29.625-23.72}{\sqrt{19.02}} = 1.35$$

Since Z = 1.35 is less than 1.645, fail to reject H_o.

Conclusion

It is obvious from the results generated in the two examples that all tests tend to give equivalent answers. At least we can say the conclusions are consistent regardless of which test is used to test H_o. Obviously this is largely a function of the distribution of data for the two examples. The agreement will not always be as consistent. Klotz (1961) has compared the relative efficiency of the S-T, M and NS tests for a specified number of distributions. For scores drawn from distributions with sharp tails (exponential, rectangular, etc.), the NS test is preferred to S-T and is equally as effective as M. When the distribution of scores has heavy tails (Cauchy, etc.), use the S-T test for testing equality of scale. Naturally when data is normally distributed the F test is most powerful. Assuming normality of scores, the asymptotic relative efficiency of S-T to F is 0.61, of M to F is 0.76 and of NS to F is 1.0. Bradley (1968), Conover (1971), and Gibbons (1971) provide excellent coverage and development of the more commonly used nonparametric tests for scale.

REFERENCES

Bradley, J. V. (1968). Distribution-Free Statistical Tests. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

Capon, J. (1961). "Asumptotic Efficiency of Certain Locally Most Powerful Rank Tests," Annals of Mathematical Statistics, 32, 88-100.

Conover, W. J. (1971). Practical Nonparametric Statistics. New York: John Wiley & Sons, Inc.

Gibbons, J. D. (1971). Nonparametric Statistical Inference. New York: McGraw-Hill.

Klotz, J. H. (1962). "Nonparametric Tests for Scale, "Annals of Mathematical Statistics, 33, 495-512.

Mood, A. M. (1954). "On the Asymptotic Efficiency of Certain Nonparametric Two-Sample Tests," Annals of Mathematical Statistics 25, 514-522.

Owen, D. B. (1962). Handbook of Statistical Tables. Reading, Mass.: Addison-Wesley.

Siegel, S. and Tukey, J. W. (1960). "A Nonparametric Sum of Ranks Procedure for Relative Spread in Unpaired Samples," Journal of the American Statistical Association, 55, 429-445.

Table I

Abstract Reasoning Data for the F, Siegel-Tukey, Mood and Normal Scores Tests

Abstract Reasoning	Rank (i)	S-T Rank	$(i-\frac{N+1}{2})^2$	$E(V^i)$	$(E(V^i))^2$
Exp.					
19	1	1	30.25	-1.629	2.654
21	3	5	12.25	-0.793	0.629
27	9	7	6.25	0.537	0.288
30	10	6	12.25	0.793	0.629
31	11	3	20.25	1.116	1.245
35	12	2	30.25	1.629	2.654
Total	46	24	111.50	1.653	8.099
Cont.					
20	2	4	20.25	-1.116	1.245
22	4	8	6.25	-0.537	0.288
23	5	9	2.25	-0.312	0.097
23	6	10	.25	-0.103	0.011
25	7	11	.25	0.103	0.011
26	8	12	2.25	0.312	0.097
Total	32	54	31.50	-1.653	1.749

Table II

Arousal - Non-Arousal Data for the F, Siegel - Tukey, Mood and Normal Scores Tests

Arousal

Time (sec)	Rank (i)	S-T Rank	$(i-\frac{N+1}{2})^2$	$E(V^i)$	$(E(V^i))^2$	$(E(V^i))^4$
139	26	50	0.25	0.025	0.001	0.000
130	20	40	30.25	-0.278	0.077	0.006
153	28	46	6.25	0.125	0.016	0.000
124	18	36	56.25	-0.384	0.147	0.022
175	31	39	30.25	0.278	0.077	0.006
360	50	2	600.25	2.249	5.058	25.583
181	32	38	42.25	0.330	0.109	0.012
360	49	3	552.25	1.855	3.441	11.841
38	3	5	506.25	-1.629	2.654	7.042
360	47	7	462.25	1.464	2.143	4.594
295	43	15	306.25	1.030	1.061	1.126
91	14	28	132.25	-0.610	0.372	0.138
155	29	43	12.25	0.176	0.031	0.001
36	1	1	600.25	-2.249	5.058	25.583
360	46	10	420.25	1.331	1.772	3.138
360	45	11	380.25	1.219	1.483	2.208
182	33	35	56.25	0.384	0.147	0.022
225	39	23	182.25	0.735	0.540	0.292
203	37	27	132.25	0.610	0.372	0.138
45	5	9	420.25	-1.331	1.772	3.138
335	44	14	342.25	1.120	1.254	1.574
203	36	30	110.25	0.551	0.304	0.092
71	11	21	210.25	-0.802	0.643	0.414
294	42	18	272.25	0.949	0.901	0.811
189	34	34	72.25	0.438	0.192	0.037
Totals	763	585	5936.25	7.586	29.625	87.818

Table II (cont.)

Non-Arousal

Time (sec.)	Rank (i)	S-T Rank	$(i-\frac{N+1}{2})^2$	$E(v^i)$	$(E(v^i))^2$	$(E(v^i))^4$
360	48	6	506.25	1.629	2.654	7.042
131	23	45	6.25	-0.125	0.016	0.000
82	12	24	182.25	-0.735	0.540	0.292
287	41	19	240.25	-0.873	0.762	0.581
131	22	44	12.25	-0.176	0.031	0.001
49	7	13	342.25	-1.120	1.254	1.574
129	19	37	42.25	-0.330	0.109	0.012
195	35	31	90.25	0.494	0.244	0.060
54	8	16	306.25	-1.030	1.061	1.126
118	16	32	90.25	-0.494	0.244	0.060
140	27	47	2.25	0.075	0.006	0.000
249	40	22	210.25	0.802	0.643	0.414
47	6	12	380.25	-1.219	1.486	2.208
133	24	48	2.25	-0.075	0.006	0.000
98	15	29	110.25	-0.551	0.304	0.092
129	17	33	72.25	-0.438	0.192	0.037
38	2	4	552.25	-1.855	3.441	11.841
138	25	49	0.25	-0.025	0.006	0.000
62	9	17	272.25	-0.949	0.901	0.811
131	21	41	20.25	-0.227	0.052	0.003
162	30	42	20.25	0.227	0.052	0.003
44	4	8	462.25	-1.464	2.143	4.594
65	10	20	240.25	-0.873	0.762	0.581
220	38	26	156.25	0.671	0.450	0.203
90	13	25	156.25	-0.671	0.450	0.203
Totals	512	690	4476.25	-7.586	17.809	31.738