Bounds on the Power of Linear Rank Tests for Scale Parameters

RONALD L. WASSERSTEIN and JOHN E. BOYER, JR.*

We show that the power functions of a class of nonparametric tests for the equality of two scale parameters do not approach 1 as the ratio of the parameters approaches infinity. The class of tests, known as linear rank tests, is shown to have a fundamental flaw when applied to scale parameters, resulting in low power when the sample sizes are small.

KEY WORDS: Ansari–Bradley test; Fligner–Killeen test; Nonparametric tests; Siegel–Tukey test; Symmetric score functions.

1. INTRODUCTION

Linear rank tests are an important aspect of the class of techniques and methods known as nonparametric statistics. Many linear rank test procedures for testing location and scale parameters in the two-sample problem have been developed, and some have been found to be quite useful and powerful. In the scale parameter case, the most commonly recommended nonparametric methods are linear rank procedures.

The development of linear rank tests for scale followed naturally and intuitively from those designed for location. It seemed quite reasonable that a method so successful in the location parameter case should have a scale parameter analog. Developers of such tests thoroughly analyzed various asymptotic properties (such as asymptotic relative efficiency), and some (e.g., Klotz 1962) observed that the small-sample power of such tests seemed to be low.

We will show that it is the nature of linear rank tests for scale that leads to this problem. As a consequence, it may well be said that linear rank tests for scale are inappropriate.

2. LINEAR RANK TESTS FOR SCALE PARAMETERS

Let $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$ be independent random samples from continuous cdf's $F_X$ and $F_Y$, respectively. Let $\theta_X$ and $\theta_Y$ be the scale parameters associated with $F_X$ and $F_Y$, respectively, and assume that the two distributions have common median $\mu$. Let $N = m + n$.

Let $Z^{(1)}, \ldots, Z^{(N)}$ represent the combined ordered sample of the $X$'s and $Y$'s. Define an indicator $z_i$ as

$$z_i = 1 \quad \text{if } Z^{(i)} \text{ is an } X \text{ observation}$$

$$z_i = 0 \quad \text{if } Z^{(i)} \text{ is a } Y \text{ observation}$$

$(i = 1, \ldots, N)$. Let $z = (z_1, \ldots, z_N)'$, and denote the set of all possible values of $z$ by $\mathcal{Z}$.

Let $T = T(x_1, \ldots, x_m; y_1, \ldots, y_n)'$ be a linear rank test statistic for dispersion, having the form

$$T = a'z = \sum_{i=1}^{N} a_i z_i,$$

where $a = (a_1, \ldots, a_N)'$ is a vector of scores. $T$ is the sum of some function of the ranks of the $X$ observations in the combined ordered sample.

Consider the hypothesis $H_0: \theta = 1$ versus $H_1: \theta > 1$, where $\theta = \theta_X/\theta_Y$. Let the rejection region $R$ for this test be defined as $R = \{z \in \mathcal{Z}: T(z) < t_0\}$, where $t_0$ is chosen so that the test has size $\alpha$. (Hence small values of $T$ cause the null hypothesis to be rejected.) Let $A$ be the complement of $R$. Of course, depending on the values of $\alpha$, the rejection region may alternatively be defined to include large values of $T$ rather than small ones.

3. BOUNDS ON THE POWER OF A TEST

Consider first the common type of location hypothesis that is tested using a linear rank procedure (such as the Wilcoxon rank sum test). We wish to test $H_0: \mu_1 = \mu_2$ versus $H_1: \mu_1 \neq \mu_2$, where $\mu_i$ represents the median of the $i$th population $(i = 1, 2)$. We would certainly expect a reasonable test of this hypothesis to have the following property: when the null hypothesis is “very false,” that is, when $|\mu_1 - \mu_2|$ is very large, the probability of rejecting $H_0$ should be close to 1.

Similarly, in the scale hypothesis test under consideration here, $H_0: \theta = 1$ versus $H_1: \theta > 1$, large values of $\theta$ should be expected to lead to rejection of the null hypothesis with high probability. In fact, it should be the case that, for a given sample size,

$$\lim_{\theta \to \infty} \Pr(\text{reject } H_0 \mid \theta) = 1.$$  

If not, the test is said to be nonresolving for that sample size. This terminology originates from Jogdeo (1966).

That a statistical test should be resolving seems to be a fundamental requirement. Hodges and Lehmann (1962) investigated the location problem under the assumption that the samples are taken from normal distributions. Their results provide limits on the probabilities of the possible rankings and ensure that the usual rank tests for location are resolving. We shall now show that commonly employed linear rank tests for scale lack this property.

We begin by considering what happens to the set of possible values of $z$ as $\theta$ goes to infinity. (We provide an intuitive argument here; the mathematical argument is straightforward.) As $\theta$ goes to infinity, the spread of the $Y$'s

---

*Ronald L. Wasserstein is Assistant Professor, Department of Mathematics and Statistics, Washburn University, Topeka, KS 66621. John E. Boyer, Jr., is Associate Professor, Department of Statistics and Statistical Laboratory, Kansas State University, Manhattan, KS 66506.
becomes infinitesimally small relative to that of the X's. Thus the Y's may be thought of as being massed at a single point (the common location parameter), whereas the X's are dispersed on either side of this point. That is, we may think of \( \theta \) approaching infinity as though \( \theta_k = 1 \) and \( \theta_p = 0 \).

Hence, in the limit, there are \( m + 1 \) possible arrangements of the ones and zeros in \( z \). Let \( W = \{w_0, \ldots, w_m\} \) be this set of possible values of \( z \) as \( \theta \) goes to infinity, where the subscript indicates the number of ones to the right of the zeros (equivalently, the number of X's that are greater than all of the Y's (which are “massed at the median”)). For example, if \( m = n = 4 \), we have

\[
\begin{align*}
w_0' &= \{1 1 1 1 0 0 0 0\}, \\
w_1' &= \{1 1 1 0 0 0 0 1\}, \\
w_2' &= \{1 1 0 0 0 0 1 1\}, \\
w_3' &= \{1 0 0 0 0 1 1 1\}, \\
w_4' &= \{0 0 0 0 1 1 1 1\}.
\end{align*}
\]

In the limit, these are the only values of the \( z \) that should occur. Since by “in the limit” we mean that the null hypothesis is “very false,” each of these \( z \) values should result in rejecting the null hypothesis. If any of these \( w \) values leads us to accept \( H_0 \), the power of the test will be limited by an amount equal to the probability of the occurrence of the particular \( w \) value. This is the essence of the following.

**Lemma 1.** Let \( W = \{w_0, \ldots, w_m\} \) be the set of possible values of \( z \) as \( \theta \to \infty \). Then the probability of observing element \( w \), is given by

\[
\Pr(w_k) = \frac{m!}{(i!)(m - i)!}(.5)^m,
\]

which is, of course, the binomial pdf with \( p = .5 \).

The foregoing expression then represents the probability that exactly \( i \) of the \( m \) X's are positive, where the distribution of the X's has median 0.

**Theorem 1.** A linear rank test for \( H_0 : \theta = 1 \) versus \( H_1 : \theta > 1 \) is nonresolving for a given sample size iff \( W \cap A \) is not empty, where \( A \) is the acceptance region.

**Proof.** Since each of the \( w \) obviously has nonzero probability by the expression in Lemma 1, the result follows trivially. In fact, it follows from the foregoing argument that

\[
\lim_{\theta \to \infty} \Pr(\text{reject } H_0 \mid \theta) = \Pr(W \cap A).
\]

Thus a test will be nonresolving if any of its limiting cases (the \( w \) vectors) result in acceptance of the null hypothesis. We now describe a property of scores that results in such tests. Many of the common linear rank tests for scale have scores of this type.

**Definition.** Assume that \( N \) is even. Consider a set of scores \( a_1, \ldots, a_N \) that has the following properties: (1a) \( a_1 \leq a_2 \leq \cdots \leq a_{N/2}, a_{N/2} \leq \cdots \leq a_N \), or (1b) \( a_1 \geq a_2 \geq \cdots \geq a_{N/2}, a_{N/2} \geq \cdots \geq a_N \); and (2) \( a_1, a_2 = a_2N-1, \ldots, a_{N-2} = a_{N-1} \). We define such scores to be symmetrically unimodal (if the scores only satisfy either property 1a or property 1b, we call them unimodal). Clearly, tests satisfying property (1a) will reject the null hypothesis for small values of \( T \), whereas those with property (1b) will reject for large values of \( T \).

Note in Table 1 that the scores of Ansari and Bradley (1960), Capon (1961), Klotz (1962), and Mood (1954) are symmetrically unimodal. The Siegel–Tukey (1960) scores are unimodal but not symmetric, though they are nearly so. Clearly, it is reasonable to attempt to test for dispersion based on this type of score, since if the X's are more dispersed than the Y's, the X's will tend to be in the “tails” of the combined ordered sample. Symmetrically unimodal scores lead to nonresolving linear rank tests, as the next theorem shows.

**Theorem 2.** A linear rank test having size \( 0 < \alpha < .5 \) for \( H_0 : \theta = 1 \) versus \( H_1 : \theta > 1 \) based on symmetrically unimodal scores is nonresolving for \( m = n \).

**Proof.** It suffices to show that the set \( W \cap A \) cannot be empty. Consider the element \( w_{(m/2)} \) of \( W \) (where \( \lfloor x \rfloor \) indicates the greatest integer less than or equal to \( x \)). Suppose without loss of generality that the score function has property (1a). For this type of score function, \( w_{(m/2)} \in R \), since this case (the X’s in the tails) is clearly a case that must result in rejecting \( H_0 \), because it provides the smallest possible value of the test statistic. A test of the form discussed here that does not reject for the smallest value of the test statistic is a size 0 (not a size \( \alpha \)) test. Let \( a \) be a vector denot-

---

**Table 1. Scores Used in Some Linear Rank Tests for Scale**

<table>
<thead>
<tr>
<th>Test</th>
<th>Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ansari–Bradley</td>
<td>( N ) even: 1, 2, ..., ( N/2, N/2, \ldots, 2, 1 ) ( N ) odd: 1, 2, ..., ( (N - 1)/2, (N + 1)/2, \ldots, 2, 1 )</td>
</tr>
<tr>
<td>Capon</td>
<td>( E[U_{(i)}] ), where ( U_{(i)} ) is the ( i )th smallest order statistic from a sample of ( N ) standard normal variables</td>
</tr>
<tr>
<td>Klotz</td>
<td>( i^2 \cdot (i((N + 1))/2)^2, i = 1, 2, \ldots, N; \Phi ) is the cumulative standard normal</td>
</tr>
<tr>
<td>Mood</td>
<td>( i - (N + 1)/2 ), ( i = 1, 2, \ldots, N )</td>
</tr>
<tr>
<td>Siegel–Tukey</td>
<td>( N ) even: ( a_i = 2i, i ) even, ( 1 &lt; i \leq N/2 ) ( = 2i - 1, i ) odd, ( 1 \leq i \leq N/2 ) ( = 2(N - i) + 2, i ) even, ( N/2 &lt; i \leq N ) ( = 2(N - i) + 1, i ) odd, ( N/2 &lt; i \leq N ) ( N ) odd: drop the median and assign as for ( N ) even</td>
</tr>
</tbody>
</table>
ing the scores associated with the linear rank test. Thus $a'w_{[m]} \leq t_0$.

Consider, on the other hand, the element $w_0$. We have

$$a'w_0 = \sum_{i=1}^{n} a_i = (1/2) \sum_{i=1}^{n} a_i.$$ 

That is, in this case the value of the linear rank statistic computed on the X's equals its value when computed on the Y's. If $\alpha < .5$, then $w_0$ must be in A. (Suppose that $w_0 \in R$, and the test was made using a statistic based on the Y's instead of the X's. Then the null hypothesis would again be rejected, leading to the contradictory result that $\theta_1 < \theta_i$ and $\theta_i > \theta_j$.)

As a result, we have $a'w_{[m]} \leq t_0 < a'w_0$. Hence a linear rank test for scale must reject $H_0$ for at least one value of $w$ and accept $H_0$ for at least one value of $w$.

We have shown that balanced ($m = n$) linear rank tests based on symmetric unimodal scores are nonresolving because certain outcomes that will occur with nonzero probability result in accepting the null hypothesis, regardless of the value of $\theta$. Thus, for a given $\alpha$ level, the power can never exceed a specified value. We show some of these values in the next section, and we make it clear that the result often holds in unbalanced samples as well. We have found this result to be true in general even when sample sizes are not equal and for tests, including the Siegel–Tukey, that do not satisfy (2) of the definition of symmetric unimodality.

4. BOUNDS ON THE POWER OF CERTAIN LINEAR RANK TESTS FOR SCALE

In this section we give an example of calculating the limiting power of the Siegel–Tukey test, and we examine the limiting power of several tests over a variety of sample sizes.

### Table 2. Limiting Power of Certain Linear Rank Tests for Scale at $\alpha = .05$

<table>
<thead>
<tr>
<th>m/n</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>.750</td>
<td>.750</td>
<td>.750</td>
<td>.750</td>
<td>.750</td>
<td>.750</td>
<td>.750</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.375</td>
<td>.375</td>
<td>.375</td>
<td>.375</td>
<td>.375</td>
<td>.375</td>
<td>.375</td>
<td>.375</td>
<td>.375</td>
<td>.375</td>
<td>.375</td>
<td>.375</td>
<td>.375</td>
<td>.375</td>
</tr>
<tr>
<td>5</td>
<td>.625</td>
<td>.625</td>
<td>.625</td>
<td>.625</td>
<td>.625</td>
<td>.625</td>
<td>.625</td>
<td>.625</td>
<td>.625</td>
<td>.625</td>
<td>.625</td>
<td>.625</td>
<td>.625</td>
<td>.625</td>
</tr>
<tr>
<td>7</td>
<td>.547</td>
<td>.547</td>
<td>.547</td>
<td>.547</td>
<td>.547</td>
<td>.547</td>
<td>.547</td>
<td>.547</td>
<td>.547</td>
<td>.547</td>
<td>.547</td>
<td>.547</td>
<td>.547</td>
<td>.547</td>
</tr>
<tr>
<td>8</td>
<td>.711</td>
<td>.711</td>
<td>.711</td>
<td>.711</td>
<td>.711</td>
<td>.711</td>
<td>.711</td>
<td>.711</td>
<td>.711</td>
<td>.711</td>
<td>.711</td>
<td>.711</td>
<td>.711</td>
<td>.711</td>
</tr>
<tr>
<td>9</td>
<td>.492</td>
<td>.492</td>
<td>.492</td>
<td>.492</td>
<td>.492</td>
<td>.492</td>
<td>.492</td>
<td>.492</td>
<td>.492</td>
<td>.492</td>
<td>.492</td>
<td>.492</td>
<td>.492</td>
<td>.492</td>
</tr>
<tr>
<td>10</td>
<td>.656</td>
<td>.656</td>
<td>.656</td>
<td>.656</td>
<td>.656</td>
<td>.656</td>
<td>.656</td>
<td>.656</td>
<td>.656</td>
<td>.656</td>
<td>.656</td>
<td>.656</td>
<td>.656</td>
<td>.656</td>
</tr>
<tr>
<td>13</td>
<td>.733</td>
<td>.733</td>
<td>.733</td>
<td>.733</td>
<td>.733</td>
<td>.733</td>
<td>.733</td>
<td>.733</td>
<td>.733</td>
<td>.733</td>
<td>.733</td>
<td>.733</td>
<td>.733</td>
<td>.733</td>
</tr>
<tr>
<td>15</td>
<td>.698</td>
<td>.698</td>
<td>.698</td>
<td>.698</td>
<td>.698</td>
<td>.698</td>
<td>.698</td>
<td>.698</td>
<td>.698</td>
<td>.698</td>
<td>.698</td>
<td>.698</td>
<td>.698</td>
<td>.698</td>
</tr>
<tr>
<td>20</td>
<td>.737</td>
<td>.855</td>
<td>.885</td>
<td>.959</td>
<td>.959</td>
<td>.959</td>
<td>.959</td>
<td>.988</td>
<td>.988</td>
<td>.988</td>
<td>.997</td>
<td>.997</td>
<td>.997</td>
<td>.997</td>
</tr>
</tbody>
</table>

NOTE: The top value is for Kolz, the middle value is for Siegel–Tukey, and the bottom value is for Ansari–Bradley. (The index $m$ is the number of observations of $X$, the variable with the larger scale parameter; $n$ is the number for $Y$, the variable with the smaller parameter.)

Let $m = n = 7$. The Siegel–Tukey set of scores is $a' = (1.458912131411107632)$, and the set $w$ consists of:

- $w_0 = \{11111110000000\}$, $a'w_0 = 52$,
- $w_1 = \{11111100000000\}$, $a'w_1 = 41$,
- $w_2 = \{11111000000011\}$, $a'w_2 = 32$,
- $w_3 = \{11100000001111\}$, $a'w_3 = 29$,
- $w_4 = \{11000000001111\}$, $a'w_4 = 28$,
- $w_5 = \{10000000111111\}$, $a'w_5 = 33$,
- $w_6 = \{10000000111111\}$, $a'w_6 = 40$,
- $w_7 = \{00000001111111\}$, $a'w_7 = 53$.

Since the critical value (Siegel and Tukey 1960) of the test is .05, cases $w_0, w_1, w_6,$ and $w_7$ result in acceptance of $H_0$. Thus the limiting power of the test is

$$1 - Pr\{w_0, w_1, w_6, w_7\} = 1 - 2[.5^7 + 7(.5^7)] = .875$$

Table 2 gives values of the limiting power for $\alpha = .05$ of the Klotz (1962), Siegel–Tukey (1960), and Ansari–Bradley (1960) tests, which are probably the most frequently applied nonparametric tests for scale. All tests considered in this article are nominally at $\alpha = .05$, and the convention is to use the test that has the closest achievable level below .05. We see from the table that (a) for $m = n$, the limiting power approaches 1 as $m$ gets large; (b) some tests are resolving for very unequal sample sizes; (c) for fixed $m$, the limiting power increases with $n$, but the same is not always true for increased $m$ with fixed $n$ (this asymmetry is not surprising in light of the fact that the alternative is one-tailed); (d) the values of the limiting power for the three tests nearly always agree. When they differ, it is due to a single pair of elements ($w$ and $w_{m-1}$) that are in the acceptance region in some of the tests and in the rejection region in the others.

### 5. CONCLUSIONS

Linear rank tests of scale have low power for small sample sizes because a fundamental property of these tests is that they are nonresolving. It is the nature of the test statistics themselves, not the failure of the population medians to be identical, that causes the difficulties presented here.

In practice, when a statistician encounters a set of data with all of the $X$’s larger (or smaller) than all of the $Y$’s, there is likely to be a great hesitancy to use methods that assume equal population locations, despite the fact that this phenomenon occurs with fairly high probability for small sample sizes.

A test proposed by Fligner and Killeen (1976) overcomes some of these problems. In their test the two samples are presumed to have come from populations having some common but unknown location parameter $\mu$. The quantity $\mu$ is estimated from the sample values, with the resulting value denoted by $M$. Tests for equality of scale are then performed by ranking the $|X_i - M|$’s and the $|Y_i - M|$’s.

The Fligner–Killeen test is distribution free and has been compared with the other test statistics discussed previously (Wasserstein 1987). The difficulties described here were found to be present but to a much smaller degree for the Fligner–Killeen test than for the statistics studied in this article.

Of course, if one were truly suspicious about the assumption of equal population medians, one would be led to adjusting the two samples separately for location. This was discussed by Raghavachari (1965). After such alignment the most extreme $w_i$’s (all of the $X$’s on one side of all of the $Y$’s) would not occur and the next most extreme values would be quite unlikely to occur. Unfortunately, the distribution-free character of the test statistic is also lost.

[Received May 1989. Revised May 1990.]

### REFERENCES


