Joint Confidence Regions for Location and Scale Differences

by

Dana Quade

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1499

January 1986
JOINT CONFIDENCE REGIONS FOR LOCATION AND SCALE DIFFERENCES

by Dana Quade

Department of Biostatistics, University of North Carolina
at Chapel Hill

On leave at Department of Biostatistics, German Cancer Research Center,
Heidelberg

Introduction

Suppose two random variables X and Y are related by the location-scale difference model, in which Y is distributed as \( yX + \delta \) for some constants \((\delta, \delta)\) in the half-plane where \( \delta \geq 0 \); that is, \( F(z) = G(\delta z + \delta) \) for \( -\infty < z < \infty \), where \( F \) and \( G \) are the distribution functions of \( X \) and \( Y \). Then consider constructing a confidence chart, showing confidence regions for \((\delta, \delta)\) within the half-plane. Given any two ordered random samples \( X_1 \leq \ldots \leq X_n \) and \( Y_1 \leq \ldots \leq Y_n \). Confidence procedures for location and scale differences separately have been discussed extensively in the literature; see for example Bauer (1972) and Noether (1978). Procedures for location and scale differences simultaneously have received scant attention, however, though they are briefly treated by Switzer (1976) as a special case of his more general approach.

It is possible in principle to draw a confidence chart to correspond with any two-sample test of homogeneity. Let

\[ H(\delta_0, \delta_0): (\delta, \delta) = (\delta_0, \delta_0) \]

be the hypothesis that the parameter \((\delta, \delta)\) takes on the prespecified value \((\delta_0, \delta_0)\). Then \( H(\delta_0, \delta_0) \) can be tested by constructing observations \( Z_i = \gamma X_i + \delta \) for \( i = 1, \ldots, m \) and then testing the homogeneity of the Ys and Zs. But it is by now well established (NOTE 1) that the confidence set for any parameter is the union of all those hypothetical values which would be accepted by a test, with confidence level \( 1 - \alpha \) if the test is at level \( \alpha \).

If \((\delta, \delta)\) would be accepted at a given value of \( \alpha \), then it would also be accepted at smaller values; and if rejected at given \( \alpha \), then also at larger values. Thus the confidence regions at different levels of \( \alpha \) form a nested series, those with less confidence inside those with more. As the confidence level decreases to 0, (or as \( \alpha \) increases to 1), the confidence region may converge to a single point, which can then be taken as the point estimate of \((\delta, \delta)\); otherwise, some point which lies in all non-vanishing confidence regions can be chosen arbitrarily.

A completely general method for ascertaining confidence regions is trial-and-error, as follows. Divide the half-plane into a set of mutually exclusive and exhaustive areas, choose one value of \((\delta, \delta)\) within each area, calculate whether it is acceptable or not, and classify the entire area accordingly; then the areas classified as
acceptable jointly constitute an approximation to the confidence region. The approximation can be refined to any desired extent by dividing the areas into subareas and calculating anew. Although, of course, this trial-and-error process may be extremely tedious.

This article has two main purposes: the first is to develop precise direct methods corresponding to various tests. It must be admitted from the outset, however, that these methods are of practical use only with small samples; trial-and-error is still best for large samples, no matter how much computation it may require.

The other main purpose is didactic. One can gain insight into the nature of an inferential procedure, and its behavior when a location-scale model holds, by constructing the confidence chart associated with it and studying the typical characteristics of the corresponding confidence region. This can be enlightening even if the procedure is not designed to deal with differences in location and scale simultaneously, and it is particularly helpful for assessing its relative merits.

A confidence chart based on the following (fictitious) data is presented for each procedure discussed, and the 50% confidence region is indicated:

\[
\begin{align*}
X &= -4, -1, 0, 1, 2, 4 \quad (m = 6) \\
Y &= 0, 1, 3, 4, 5, 6, 9 \quad (n = 7).
\end{align*}
\]
2. Rank Tests

Consider constructing the confidence chart corresponding to a test based on the ranks. The rank of $Z_i$ within the combined sample of $Y$s and $Z$s is

$$R_k(\delta, \sigma) = W_k + B_k(\delta, \sigma)$$

where

$$W_k = \frac{m+1}{2} + \sum_{j=1}^{m} \text{sgn} \left( Z_k - Z_j \right) = \frac{m+1}{2} + \sum_{j=1}^{m} \text{sgn} \left( X_k - X_j \right)$$

is the within-sample rank of $Z_k$ (or of $X_k$), and

$$B_k(\delta, \sigma) = \frac{n}{2} + \sum_{i=1}^{m} \text{sgn} \left( Z_k - Y_i \right) = \frac{n}{2} + \sum_{i=1}^{m} \text{sgn} \left( \delta X_k + \sigma - Y_i \right)$$

is the number of observations $Y_i$ such that $Y_i < \delta X_k + \sigma$. Similarly, the rank of $Y_k$ is

$$Q_k(\delta, \sigma) = V_k + A_k(\delta, \sigma)$$

where

$$V_k = \frac{n+1}{2} + \sum_{j=1}^{n} \text{sgn} \left( Y_k - Y_j \right)$$

is the within-sample rank of $Y_k$ and

$$A_k(\delta, \sigma) = \frac{m}{2} + \sum_{i=1}^{m} \text{sgn} \left( Y_k - Z_i \right) = \frac{m}{2} + \sum_{i=1}^{m} \text{sgn} \left( Y_k - \delta X_i - \sigma \right)$$

is the number of observations $X_i$ such that $\delta X_i + \sigma < Y_k$.

It can be seen that as $\delta$ and $\sigma$ vary within the half-plane the ranks change, and hence a test based on them may change, only where one of the quantities $(\delta X_i + \sigma - Y_i)$ changes sign: that is, on one of the mn lines $\sigma = Y_i - \delta X_i$ for $i = 1, \ldots, m$ and $j = 1, \ldots, n$. These lines divide the half-plane into small convex polygons which may be thought of as tiles forming a mosaic. The mosaic for the example is shown as Figure 1. As $\delta$ and $\sigma$ vary within any one tile the ranks of the observations do not change, and hence the value of any rank statistic cannot change: for values $(\delta, \sigma)$ in different tiles the ranks are different, and a rank test statistic may (but may not) be different. Thus the confidence region corresponding to the rank test consists of a collection of tiles, and no subdivision of the tiles is necessary for complete precision. These results generalize those obtained by Bauer (1972) for the location-difference and scale-difference problems separately; conversely, they specialize the approach of Switzer (1975) (NOTE 2).

One point not yet discussed is the treatment of ties, typically a source of difficulty for ranking methods. Within-sample ties, of course, are no problem, since they may be broken arbitrarily with no effect on the results. But in the context of the location-scale difference model, between-sample ties are of no concern either. This is because such ties occur only for values of $(\delta, \sigma)$ on the lines of the mosaic. The confidence region is conservative -- that is, it contains the true $(\delta, \sigma)$ with probability no less than the nominal $(1-\alpha)$ -- if it is construed as including its boundary, and conversely the region is liberal if its boundary is excluded: see Noether (1978). But all lines jointly constitute only a one-dimensional subset of the two-dimensional chart.
Note that the mn lines form m bundles, each containing n parallel lines, with one bundle corresponding to each value \( X_i \), and \(-X_i\) as the common slope. Alternatively, they also form n pencils, each containing m concurrent lines, with one pencil corresponding to each value \( Y_i \), and \((0,Y_i)\) as the point of concurrency. The top tile is at the upper left, between the \( \delta \)-axis and the line \( \delta = Y_i - \delta X_i \); the bottom tile is at the lower left, between the \( \delta \)-axis and the line \( \delta = Y_i - \delta X_i \). Any tile in which the values of \( \gamma \) are bounded away from both 0 and \( \infty \) -- that is, any tile of finite area which does not touch the \( \delta \)-axis, not even by corner -- will be called bounded; the others are unbounded. In laying out the scale of the mosaic for given data it is useful to know that the bounded tiles extend on the right to

\[
\gamma_{\text{max}} = \frac{Y_n - Y_1}{\max_i (X_i - X_{i-1})}
\]

and at top [bottom] to

\[
\delta_{\text{max}} [\delta_{\text{min}}] = \max_i [\min_i] \frac{Y_1 X_i - Y_n X_{i-1}}{X_i - X_{i-1}}.
\]

The chart tends to be better proportional if the \( X \)s are centered near zero, but clearly they can be translated as will without affecting its topology.

Each tile corresponds to some pattern of \( Y \)s and \( Z \)s. To ascertain the pattern for a particular tile, choose any value of \( (\gamma, \delta) \) within it, construct \( Z_i = \delta X_i + \delta \) for \( i=1, \ldots, m \), and form the ordered array of these \( Z \)s combined with the \( Y \)s. For example, consider the large tile at the center of Figure 1, which covers the point \( \gamma, \delta \)=(5,2). From \( X=-4,-1,0,1,2,4 \) construct \( Z=-18,-3,2,7,12,22 \) and thence the ordered array

\[
\begin{array}{ccccccccc}
-18 & 3 & 4 & 5 & 6 & 7 & 12 & 22
\end{array}
\]

which has the pattern ZZYYZYYYYZZ as noted on Figure 1. The tile which includes the point \((1,0)\) corresponds to the pattern of the given data, since then \( Z_1 = X_1 \); for the example this point lies on a boundary between tiles, indicating the fact that there are between-sample ties. The top tile corresponds to the pattern

\[
\underbrace{Y \ldots Y}_n \underbrace{Z \ldots Z}_m
\]

in which all the \( Z \)s are larger than the largest \( Y \), and the bottom tile corresponds to the opposite pattern

\[
\underbrace{Z \ldots Z}_m \underbrace{Y \ldots Y}_n
\]

in which all the \( Z \)s are smaller than the smallest \( Y \). These and other representative patterns are noted on Figure 1.

An alternative way of ascertaining the patterns is by recursion from one tile to the next. As the line \( \delta = Y_i - \delta X_i \) is crossed, the \( i \)-th \( Z \) and \( i \)-th \( Y \) switch positions. Consider for example the tile just below the one that covers \((5,2)\). In passing between these tiles one crosses the line \( \delta = Y_2 - \delta X_2 \). Switching the third \( Z \) with the second \( Y \) in the pattern ZZYYZYYYYZZ ascertained earlier yields ZZYYZYYYYZZ, as shown on Figure 1. (Note that the two tiles to be switched are in adjacent positions, which will always be the case.) This result could of course be verified by constructing the \( Z \)s for some point \((\gamma, \delta)\) in the tile.
The mosaic may also be regarded as a dual form of the pair chart (Quade, 1972). Since each line of the mosaic corresponds to a pair of observations, one X and Y, it also corresponds to one square of the chart. The bundles correspond to the columns of the pair chart, and the pencils to the rows. The tiles correspond to patterns, or paths on the chart: a tile is above a line of the mosaic if and only if the path which corresponds to it is above the square which corresponds to the line.

How many tiles are there? It is well-known that k lines divide the plane into at most \((k^2+k+2)/2\) parts. If any u of the lines are parallel, however, this total is reduced by \(u(u-1)/2\); and similarly if \((u+1)\) lines are concurrent. Now the confidence chart has \((mn+1)\) lines, counting the \(\delta\)-axis. They would divide the whole plane into \((mn^2+3mn+4)/2\) tiles, but the \(m\) bundles of \(n\) parallel lines reduce this by \(n(n-1)/2\) each, and the \(n\) pencils of \((m+1)\) concurrent lines (counting the axis) reduce this by \(m(m-1)/2\) each. After dividing the result by two because only half the plane is being considered, simple calculations indicate that there should be

\[
G(m,n) = \frac{(mn^2-mn(m+n-5)+4)}{4}
\]

tiles in all. But this may still be only an upper bound: if two \(X\)s are equal, then two bundles coincide; and if two \(Y\)s are equal, then two pencils coincide. Thus \(m\) and \(n\) in the formula for \(G\) must be interpreted as the numbers of distinct \(X\)s and \(Y\)s. The total is still further reduced if there are concurrents of 3 or more lines at points with \(\delta > 0\): this is equivalent to having 3 or more pairs of observations \((X_i, Y_j)\) lie on a line of positive slope in the \((X, Y)\) plane. In the example, \(m=6\) distinct \(X\)s and \(n=7\) distinct \(Y\)s yield \(G(6,7)=358\): but there are so many extra concurrents that the actual number of tiles is only 230. Note that if \(\min(m,n) > 2\) then \(G(m,n)\) is smaller than the number of conceivable patterns, namely \((m+n)!/\min!\). This is because certain patterns are incompatible. For example, if \(m=n=3\) then \(6!/3!3! = 20\) patterns are conceivable, but at most \(G(3,3)=19\) can occur: whatever \(X\)s and \(Y\)s may be given, only one the two patterns \(ZYYZYZ\) and \(YZZYYZ\) can be produced, no matter how \(\delta\) and \(\delta\) are chosen (NOTE 3).
3. Examples of Rank Tests

This section explains how to construct confidence charts corresponding to specific rank tests, and describes their characteristic features. Each test is illustrated by a confidence chart showing the best 50% confidence region (its exact confidence coefficient may considerably exceed 50%) based on the example presented earlier. Where a test has both one- and two-sided versions, discussion is limited to the latter.

A. Quantile Tests

Let the data be summarized by a 2xc contingency table, where the two rows represent the two samples and the c columns represent a categorization determined by prespecified quantities of the combined data. Suppose Ck observations are called for in the k-th column, where $C_1+...+C_c = m+n$. Then the only lines relevant to the test statistic are those where $i+j-1 = C_k$, $(C_1+C_2),...,$ or $(C_1+C_2+...+C_{C_k})$: the remaining lines may be omitted. The relevant lines form supertiles within which the contingency table, and hence any test statistic calculated from it, remains constant. Ascertainment of confidence regions is thus greatly simplified.

i) The median test is based on the contingency table in which the observations are classified as above or below the median of the combined sample: if m+n is odd the one observation equal to the median may be placed in a column by itself. The statistic S, equal to the number of Zs below the median minus the number above, may be used as the test criterion. The relevant lines are then those where $i+j-1-(m+n)/2 < 1$: there are $m,n$ of them if m+n is even, or twice as many if m+n is odd. Since these lines do not cross within the half-plane (NOTE 4), the confidence region is a smooth band, bounded by two of them, and stretching across the width of the chart. Figure 2A shows the relevant lines for the example, and the corresponding contingency tables for some of the supertiles: the shaded area, where $|S| \leq 1$, has exact confidence 58.3% (1000/1716) (NOTE 5). Note that the one-dimensional confidence set for $\delta$ given any fixed $\delta$ is a finite interval: this is reasonable, since the test is sensitive to location differences but generally not to scale. Contrariwise, the one-dimensional confidence set for $\delta$ given a fixed $\delta$ may not be finite, though it will always be a finite interval if the observations are all positive (NOTE 6).

ii) The outer quartiles test of Westenberg (1948) is based on the 2x3 table in which the first and last columns each have (approximately) one quarter of the observations: specifically, they may each be given $k = [m+n+1]/4$ observations, where the notation $[u]$ indicates the largest integer not greater than u. Then the 2x3 table is collapsed by combining the first and last columns, thus contrasting the observations between the quartiles with the others, and Fisher's exact p-value is calculated. The relevant lines are those where $i+j-1 = k$ or $m+n-k$: they number at most 2k, and determine at most $(k+1)^2$ supertiles (NOTE 7). Such a test is aimed at differences in scale only, presuming no difference in location: but it may be of interest to learn how it behaves if in fact there is a difference in location, with or without an accompanying difference in scale: does it reject, or not? That is, does its confidence region include extreme value of $\delta$? Figure 2B shows the relevant lines for the example, and the corresponding contingency tables for some of the larger supertiles: the shaded area has exact confidence 71.4% (1225/1716) (NOTE 8). It may be seen that this confidence region is an irregular vertical band. Given a fixed $\delta$, the one-dimensional confidence set for $\delta$ is an interval, bounded unless $\alpha$ is very small; but given $\delta$, the confidence set for $\delta$ may extend to infinity in both directions.
iii) The tertiles test is more appropriate for the location-scale difference model than the two considered so far, since it is aimed at both location and scale differences simultaneously. It is based on the 2x3 contingency table in which observations are classified as above, below, or between the two tertiles of the combined data; specifically, to allow for m+n not being a multiple of 3, put \( l = [(m+n+1)/3] \) observations at each end. The relevant lines are then those where \( i+j-1 = l \) or \( m+n-l \); there are at most 2l such lines, and they determine at most \((l+1)(3l+5)/4\) supertertiles -- fewer if m and n are greatly unequal (NOTE 9). The chi-squared statistic calculated from the contingency table may be used as the test criterion; for large sample sizes its null-hypothesis distribution is approximately \( \chi^2 \) with 2 degrees of freedom, and for small sample sizes exact P-values can easily be obtained. Since this test is sensitive to both location and scale alternatives, its confidence region does not extend in any direction without limit, at least if the sample sizes are not too small. In particular, a bounded region exists for some \( \alpha \) if \( \min(3m,3n,m+n) \geq 12 \) (NOTE 10). Thus, for the example, the five bounded supertertiles jointly constitute the region where \( \chi^2 \leq 2.14 \), with exact confidence 67.6% (NOTE 11) (1160/1716), as shown in Figure 2C. Boundedness is of course a desirable property for any confidence region; see Salvia (1980) for discussion of this point in a one-sample context.

B. Recursive Tests

A recursive test is one such that its test statistic, say \( T \), changes by an amount \( t(i,j) \) when the line \( \delta = Y_j - \delta X_i \) is crossed from above to below. Any line with change function \( t(i,j) \) equal to zero is, of course, irrelevant. Such a test requires ascertaining the statistic directly in only one tile, say the top tile; then the statistic can be calculated recursively for all the others, perhaps checking the recursion by an independent direct ascertainment in the bottom tile.

It is not difficult to show that all linear rank tests are recursive, with change functions depending only on the sum \( (i+j) \) (NOTE 12). It happens that the median and outer quartiles tests (but not the tertiles test) are linear rank tests. In particular, for the median test

\[
S = \sum_{i=1}^{m+n} \text{sgn} \left( \frac{m+n+1}{2} - R_i (\delta, \delta) \right). 
\]

This equals \(-\min(m,n)\) in the top tile and \(+\min(m,n)\) in the bottom tile; its change function is

\[
s(i,j) = \begin{cases} 
2 & \text{if } i+j-1 = (m+n)/2 \\
1 & \text{if } i+j-1 = (m+n)/2 + 1/2 \\
0 & \text{otherwise}.
\end{cases}
\]

The outer quartiles test can be treated similarly (NOTE 13).

iv) The well-known Wilcoxon-Mann-Whitney test is based on the rank sum statistic \( W_\phi = \Sigma Q_i (\delta, \delta) \) (Wilcoxon, 1945), or equivalently on the statistic \#_\phi, the number of pairs of observations consisting of one \( Z \) and one \( Y \) such that \( Z < Y \) (Mann and Whitney, 1947), where \( W_\phi = \#_\phi = n(n+1)/2 \). In the top tile \( W_\phi = n(n+1)/2 \) and \( \#_\phi = 0 \), and as one proceeds down the chart, \( W_\phi \) and \( \#_\phi \) increase by 1 every time a line is crossed; so, after crossing all \( mn \) lines, one finds in the bottom tile that \( W_\phi = n(2m+n+1)/2 \) and \( \#_\phi = mn \). Thus the test is recursive with change function identically equal to unity. The confidence region, defined by \(|W_\phi-n(m+n+1)/2|\) or \(|\#_\phi-mn/2|\) being sufficiently small, is an irregular horizontal band, but roughly similar to that given by the median test. This is reasonable, since the Wilcoxon-Mann-Whitney test is also sensitive to location rather than scale differences. The same remarks apply with respect to the one-dimensional confidence sets (NOTE 6). Figure 2D shows values of \#_\phi for
selected tiles using the example data: the shaded area, where \(|\#_{xy} - 21| \leq 5\), has exact confidence 55.5\% (952/1716) (NOTE 14).

v) Galton’s test is also aimed at location differences. For \(m=n\), its test statistic, say \(G\), is the number of pairs \((Z_i, Y_i)\) such that \(Z_i < Y_i\); for \(m \neq n\) there is no test except in some special cases (Hodges, 1955). The test is recursive, with change function equal to 1 for those lines where \(i=j\) and 0 otherwise: \(G\) equals 0 in the top tile and \(n\) in the bottom tile. The confidence region is similar to those of the two location tests already considered, but typically much larger. In order to apply Galton’s test to the example data, one observation on \(Y\) must be discarded so as to make the sample sizes equal; \(Y_4=4\) was chosen arbitrarily. Figure 2E shows the resulting confidence chart, with only the relevant lines drawn, and \(G\) indicated for selected (super)tiles: the shaded area, where \(1 \leq G \leq 5\), has exact confidence 71.4\% (5/7) (NOTE 15).

vi) A rank-sum test for scale differences was presented by Freund and Ansari (1957; see also Ansari and Bradley, 1960); closely related (and asymptotically equivalent) tests are due to Barton and David (1958) and Siegel and Tukey (1960). These are all linear rank tests. The original Freund–Ansari–Bradley test statistic is

\[
V = \frac{m}{2} \min_{i=1} \left\{ R_i(\gamma, \delta), m+n+1 - R_i(\gamma, \delta) \right\},
\]

with change function

\[
v(i, j) = \text{sgn} \left( i+j-1 - \frac{m+n}{2} \right);
\]

note that there are irrelevant lines if \(m+n\) is even. The value of \(V\) in the top and bottom tiles is \(m(m+1)/2\) if \(m\neq n+1\); otherwise this quantity must be reduced by \((m-n)^2/4\). With large and small values of \(V\) both considered significant, the confidence region has the same general shape as for the outer quartiles test, though typically a little smaller, because the distribution of \(V\) is spread over a larger number of values. Figure 2F shows \(V\) for selected tiles, using the example data: the shaded area, where \(21 \leq V \leq 25\), has exact confidence 50.7\% (870/1716) (NOTE 16).

vii) The squared-rank test for scale differences (Mood, 1954) rejects for extreme values of the statistic

\[
M = \sum_{i=1}^{m} \left[ R_i(\gamma, \delta) - \frac{m+n+1}{2} \right]^2 = \#_{\text{XYZ}} - \#_{\text{YXY}} + \frac{m(m^2-1+3n^2)}{12},
\]

where \(#_{\text{XYZ}}\) is the number of triplets of observations consisting of 2 \(Z\)s and 1 \(Y\) which form the pattern \(\text{XYZ}\) (and \(#_{\text{XZ}}, \#_{\text{YZ}}, \#_{\text{YX}}, \text{etc.},\) are defined analogously). This is a linear rank test, with change function \(m(i, j) = m+n-2(i+j)-1\). Clearly, there are irrelevant lines if \((m+n)\) is even, and \(M = m(m^2-1+3n^2)/12\) in the top and bottom tiles. The confidence region is similar to those of the scale tests already considered, and is not illustrated here.

viii) The Crouse–Steffens test. An undesirable feature of the standard rank tests for scale, including all those mentioned so far, is that they are sensitive not only to scale differences, but also to location differences, unless the sample sizes are roughly equal: with unequal sample sizes, if there is a major location difference then they tend to declare the smaller sample more variable regardless of what true scale difference there may be. The corresponding confidence regions then tend to bend like the letter \(C\) if \(m \gg n\), but like a reversed \(C\) if \(m \ll n\). (This is not evident for the example data, where the two sample sizes are practically the same.) To overcome this problem, Crouse and
Steffens (1969) proposed a "modified Mood test" which rejects for large absolute values of the statistic

$$Q = \frac{n-1}{2} \frac{m-1}{2} Y_{ZZY} - Y_{YYZ} = \frac{n-1}{2} \frac{m-1}{2} Y_{ZY} - Y_{YZ}.$$

The behavior of the Crouse-Steffens test is independent of the relative sample sizes: when faced with an extreme location difference (or any location difference with otherwise identical symmetric populations), the test accepts the hypothesis of homogeneity, and thus avoids the misleading implication of a significant scale difference. The corresponding confidence region is always an irregular vertical band, including the top and bottom tiles, no matter what the relative sample sizes are. Note that Q is a linear function of M if m=n; Q is also closely related to, but much more easily calculated than, the statistic

$$T = \frac{n+1}{2} Y_{YYZ} + \frac{n+1}{2} Y_{ZYZ},$$

proposed by Tamura (1960). Though not a linear rank test, unless m=n, the Crouse-Steffens test is recursive: its change function is (NOTE 17)

$$q(i,j) = (m-1)\left\{ \frac{n+1}{2} - j \right\} + (n-1)\left\{ \frac{m+1}{2} - i \right\},$$

with Q=0 in the top and bottom tiles. Since q(i,j)=0 if (i,j) = (1,n) or (m,1), the two lines $\delta = Y_n - \delta X_1$ and $\delta = Y_1 - \delta X_n$ are irrelevant; and there are other irrelevant lines if m-1 and n-1 are not relatively prime. Figure 2G shows the confidence chart for the example data, with the two irrelevant lines omitted and values of Q indicated for selected tiles: the shaded area, where |Q| ≤ 44, has exact confidence 50.1% (859/1716) (NOTE 18).

ix) The Lehmnn-Sundrum test. Lehmann (1951) proposed an "omnibus" test of homogeneity, that is, one sensitive to all alternatives: it rejects for large values of

$$L = \frac{n}{2} Y_{YYZ} + \frac{n}{2} Y_{ZYZ},$$

Sundrum (1954) showed that

$$L = \frac{m}{2} Y_{YYZ} + \frac{n}{2} Y_{ZYZ} - \frac{m}{2} Y_{YYZ} + \frac{n}{2} Y_{ZYZ},$$

in which form it is much easier to calculate. The test is recursive, with change function (NOTE 17)

$$(i,j) = (n-1)(i-1) - (m-1)(j-1):$$

L = m(m-1)n(n-1)/4 in the top and bottom tiles. Note that L(i,j) = 0 if (i,j) = (1,1) or (m,n), so the two lines $\delta = Y_1 - \delta X_1$ and $\delta = Y_n - \delta X_n$ are irrelevant; and there are other irrelevant lines if m-1 and n-1 are not relatively prime. Figure 2H shows the confidence chart for the example data, with the two irrelevant lines omitted, and values of L indicated for selected tiles: the shaded area, where L ≤ 94, has exact confidence 50.1% (859/1716). It can be shown that (NOTE 19)

$$L \approx \frac{m}{2} \frac{n}{2} \left( \frac{1}{2} - \frac{1}{k} - \frac{1}{2k^2} \right)$$

in all unbounded tiles, where $k = \min(m,n)$. It follows that if the sample sizes are sufficiently large that the critical value exceeds this lower limit on L, then the confidence region must be bounded; indeed, it can even be empty (see Section 5 for further discussion of such occurrences).
C. Miscellaneous Tests

x) The runs test (Wald and Wolfowitz, 1940) rejects for a sufficiently small number of runs in the pattern of Ys and Zs (or, equivalently, their ranks). This number is R = 2 in the top and bottom tiles; R = 3 in the triangular tiles between the pencils along the left edge of the chart, and in the infinite wedges between the bundles along the other edges; R = 4 in the other unbounded tiles within the uppermost and lowermost pencils and bundles; and R = 5 in all remaining unbounded tiles. Thus R is known for all 2(mn-1) unbounded tiles, which for small m and n is a substantial proportion of the total. Note also that the value of R is odd for tiles between the lines $\delta = Y_i - X_i$ and $\delta = Y_n - X_n$, corresponding to patterns which begin and end with the same symbol, and R is even otherwise. As either of these two lines is crossed, R changes by 1, and as any other line is crossed, R changes by 0 or 2; but the amount of change depends on where the line is crossed, so the test is not recursive. Figure 2i shows the confidence chart for the example data, with values of R indicated for selected tiles: the shaded area, where $R \geq 8$, has exact confidence 50.0% (858/1716) (NOTE 14). Since R\leq5 in all unbounded tiles (bounded tiles may also have R as small as 4), the confidence region is finite whenever the sample sizes are sufficiently large that the critical value is 5 or more. This is as would be expected for any omnibus test. As Figure 2i shows, however, the region may contain "holes", which is certainly a most disconcerting phenomenon.

xi) The Kolmogorov-Smirnov test rejects for large values of the statistic

$$K = \sup_{u} | F(u) - G(u) |,$$

where $F$ [$G$] is the empirical distribution function of the Zs [Ys]. This statistic can be ascertained directly for each tile as follows:

$$mnK = \max U \left\{ \max_{U} \left[ mi - m(j-1) \right], \max_{L} \left[ mj - n(i-1) \right] \right\},$$

where U denotes the lines $\delta = Y_i - X_i$ which form the upper boundary of the tile, and L denotes the lines of the lower boundary. A confidence region based on K consists of all tiles where $mnK \leq k$ for some integer $k$: that is, where

$$\delta \geq \max \left\{ Y_i - X_i \right\} \text{ for } ni-m(i-1) > k$$

and

$$\delta \leq \min \left\{ Y_i - X_i \right\} \text{ for } mj-n(i-1) > k.$$

These formulas are not really new, but are equivalent to inserting the scale parameter $\delta$ into those derived by Noether (1976) for pure location differences. Figure 2j shows the confidence chart for the example data, with values of mnK indicated for selected tiles: the shaded area, where $mnK \leq 17$, has exact confidence 56.6% (972/1716) (NOTE 14). It is not difficult to show that

$$K \geq \min \left\{ \left[ m/2 \right]/m, \left[ n/2 \right]/n \right\},$$

in all unbounded tiles (NOTE 20), so the confidence region is bounded whenever the sample sizes are sufficiently large that this lower limit reaches the critical value. This property was again to be expected, since the statistic K provides still another omnibus test. More importantly, however, since the confidence region is an intersection of half-planes, it must be convex: that is, if $(x_1, \delta_1)$ and $(x_2, \delta_2)$ are both acceptable values for the location scale difference then so is $(a x_1 + (1-a) x_2, a \delta_1 + (1-a) \delta_2)$, for $0 < a < 1$. This clearly is a desirable property for any confidence region, but no other standard test possesses it, except the median test, whose regions are never bounded. For discussion see again Salvia (1980), who proves both properties for the one-sample test.
Kolmogorov test: they are also evident in the discussion of the two-sample test by Switzer (1976), although he did not state either explicitly.

Moses (1952) and Hollander (1963) have proposed tests for extreme reactions, intended to be sensitive to the alternative that the observations in one sample, say the Ys, will exhibit extreme reactions in either direction or both. The Moses test of order k, for k ≤ [m/2], rejects for small values of Nk, the number of Ys included in the interval between the k-th smallest and k-th largest Xs (NOTE 21). This is a simple recursive test: its change function is

\[ n_k(i,j) = \begin{cases} 
1 & \text{if } i = k \\
-1 & \text{if } i = m+1-k \\
0 & \text{otherwise}.
\end{cases} \]

so there are only 2m relevant lines; and N_k=0 in the top and bottom (super)tiles. Figure 2K shows the confidence chart for the example data, with values of N_k indicated for selected (super)tiles after omitting the irrelevant lines: the shaded area, where N_k ≥ 5, has exact confidence 65.7% (1128/1718) (NOTE 22). Hollander's test rejects for small values of

\[ H = \sum (R_i(\hat{\delta}, \hat{\theta}) - \bar{F})^2. \]

H (NOTE 23) takes on its minimum value, namely \((m^2-m)/12\), in the top and bottom tiles, and also in the triangles between the pencils along the left edge of the chart; but it is not recursive, and ascertaining its value for all the tiles is tedious. This has been done for the example data, however, and the result is shown as Figure 2L, in which the shaded area, where H ≥ 76 (or 6H ≥ 456; mH is always an integer), has exact confidence 50.0% (858/1716) (NOTE 24). Hollander (1963) remarks that neither his nor Moses' test "should be thought of as a two-sample test of dispersion", although both are "obviously sensitive to one-sided dispersion alternatives". Their confidence regions, however, are quite similar to those which would be produced by one-sided versions of dispersion (scale) tests. Further consideration of such charts may clarify the nature of the distinction.
4. Some Tests Not Based on Ranks

A. Randomization Tests

Consider performing a randomization test given \( m \) Xs and \( n \) Ys. This involves calculating some criterion, say \( C \), for each of the \( L=(m+n)!/m!n! \) hypothetical datasets obtainable by reallocating the observations to the samples. To construct the corresponding confidence region for \((\delta, \delta)\) requires replacing each \( X \) by \((\delta X + \delta)\) before performing the test. Let \( C_k(\delta, \delta) \) be the value of \( C \) obtained after this replacement from the \( k \)-th allocation, and \( C_0(\delta, \delta) \) the observed value, supposing the actual allocation is listed first. Then the \( P \)-value is \( A/L \), where \( A \) is the number of allocations for which \( C_k(\delta, \delta) \) is as extreme as \( C_0(\delta, \delta) \), or more so; without loss of generality, we may define "extreme" as "large". But for each \( k=2, \ldots, L \), the relationship that \( C_k(\delta, \delta) \) should just be larger as \( C_0(\delta, \delta) \) defines a curve in the \((\delta, \delta)\) half-plane, and these \( L-1 \) curves then form a mosaic analogous to the one formed by the \( mn \) lines \( \delta=Y_i-\delta X_i \) which correspond to rank tests. And, in a further analogy with recursive rank tests, the statistic \( A \) changes by exactly \( 1 \) whenever one of the curves is crossed.

Only one specific randomization test is developed here – that based on Student's t, or equivalently on the difference (say, \( D \)) between the mean of the Xs and the mean of the Ys. Let \( \bar{X}_k \) be the mean of those Xs which become Ys in the \( k \)-th reallocation, and \( \bar{Y}_k \) the mean of those Ys which become Xs; and let \( u_k \) be the number of observations in each sample which switch to the other. Then

\[
D_k(\delta, \delta) = \frac{m(\bar{X}_k + \delta) - u_k (\delta \bar{X}_k + \delta)}{m} + \frac{u_k (\delta \bar{X}_k + \delta)}{n} - \frac{n \bar{Y}_k - u_k \bar{Y}_k}{n}
\]

\[
= (\delta \bar{X}_k + \delta - \bar{Y}_k) - u_k \left[ \frac{1}{m} + \frac{1}{n} \right] (\delta \bar{X}_k + \delta - \bar{Y}_k),
\]

where \( \bar{X} [\bar{Y}] \) is the mean of all the Xs [Ys]. The first term of this expression is just \( D_0(\delta, \delta) \), and clearly \( D_k(\delta, \delta) \) is equal to it if and only if \( \delta = \bar{Y}_k - \delta \bar{X}_k \). Thus the curves required are actually \( L-1 \) straight lines, each having as intercept the mean of a subset of \( u \) Ys and as slope the negative of the mean of a subset of \( u \) Xs, where \( u = 1, \ldots, \min(m,n) \). (For a more general treatment, except for restriction to location differences only, see Gabriel and Hall, 1983.)

To obtain a two-sided test comparable to the location tests considered earlier, reject for the largest values of \( |A-(L+1)/2| \) (rather than using \( |D| \) as the criterion, which leads to complications) (NOTE 25). The result is illustrated for the example, where \( L=1716 \), by Figure 3. The confidence region is an irregular horizontal band, much as for the Wilcoxon-Mann-Whitney test. If \( \delta \) is given, the confidence set for \( \delta \) is the interval from the \( v \)-th smallest value of \((Y_i-\delta X_i)\) to the \( v \)-th largest, where \( v=[\alpha L/2] \).

B. Tests Based on Normal Theory

Finally, for purposes of comparison, consider confidence charts corresponding to the classical tests based on normal theory. Thus for the remainder of this section assume that \( X \), and hence \( Y \) also under the location-scale difference model, are normally distributed.

The classical test of homogeneity of the Xs and the Ys, against the alternative of
a difference in location only, uses the statistic
\[
t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{(m-1)s_X^2 + (n-1)s_Y^2}{m+n-2}} \sqrt{\frac{1}{m} + \frac{1}{n}}}
\]
where \(s_X^2 = \Sigma(X_i - \bar{X})^2/(m-1)\) and \(s_Y^2 = \Sigma(Y_j - \bar{Y})^2/(n-1)\). This is distributed as Student's t with \(m+n-2\) degrees of freedom. On substituting \(\bar{X} + \delta\) for \(X\) it is seen that the confidence region is the area between the two branches of the hyperbola
\[
\delta = \bar{Y} - \bar{X} \pm t_\alpha \sqrt{\frac{(m-1)s_X^2 + (n-1)s_Y^2}{m+n-2}} \sqrt{\frac{1}{m} + \frac{1}{n}}
\]
where \(t_\alpha\) is the two-sided critical value. In the example, \(s_X^2 = 112/15\), \(s_Y^2 = 28/3\), and with 11 degrees of freedom \(t_\alpha = .697\) for \(\alpha = .5\), yielding
\[
\delta = 4 - \bar{X}/3 \pm .697 \sqrt{52(3+2\bar{X}^2)/89}
\]
This region is the shaded area in Figure 4A, where the boundaries of the 75% and 90% regions are indicated also.

A modified t-test, aimed at the classical Behrens-Fisher problem of testing equality of means rather than homogeneity, and allowing for possible differences in scale, takes
\[
t' = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_X^2}{m} + \frac{s_Y^2}{n}}}
\]
as a Student's t (approximately). For the degrees of freedom one may use the formula given by Satterthwaite (1946):
\[
DF = \frac{\left(\frac{s_X^2}{m} + \frac{s_Y^2}{n}\right)^2}{\frac{s_X^2}{m} + \frac{s_Y^2}{n}}
\]
Substituting \(\bar{X} + \delta\) for \(X\), \(DF\) becomes a function of \(\delta\), maximized at \(m+n-2\) if \(\delta^2 = m(m-1)s_X^2/n(n-1)s_Y^2\) and decreasing to \(n-1\) as \(\delta \to 0\) or \(m-1\) as \(\delta \to \infty\). The boundaries of the confidence region are then
\[
\delta = \bar{Y} - \bar{X} \pm t'_\alpha (\delta) \sqrt{\frac{s_X^2}{m} + \frac{s_Y^2}{n}}
\]
Figure 4B illustrates this in the same way as Figure 4A. The modified region is quite similar, though a little larger (NOTE 26).

For testing against the alternative of a difference in scale, the classical test takes \(F = s_Y^2/s_X^2\) as Snedecor's F with \(m-1\) and \(n-1\) degrees of freedom. Again substituting \(\bar{X} + \delta\) for \(X\), the confidence region becomes
\[
\frac{s_Y^2}{s_X^2} F_L^* \leq \delta^2 \leq \frac{s_Y^2}{s_X^2} F_U^*
\]
where \(P\{F_L^* \leq F \leq F_U^*\} = 1-\alpha\). Figure 4C illustrates this, with the 50% region shaded and 75% and 90% regions indicated as in 4A and 4B (NOTE 27). Note that the band is bounded by vertical lines, and \(\delta\) is entirely irrelevant.
There is no simple standard procedure for testing the joint hypothesis of specified location and scale differences simultaneously. One possibility is to reject for small values of the likelihood ratio

\[ \Lambda = \left( \frac{m-1}{m} \right)^{\frac{m}{2}} \left( \frac{n-1}{n} \right)^{\frac{n}{2}} \frac{(m-1)s_x^2}{m+n} + \frac{(n-1)s_y^2}{m+n} + \frac{mn(X-Y)^2}{(m+n)^2} \]

which has approximately the uniform distribution on (0,1) if m and n are large. Substituting \(s'X+s\) for \(X\), one finds that the joint confidence region is defined by

\[ \frac{-mn}{(m+n)^2} (\bar{\delta}-Y-\bar{\delta}'X)^2 \leq \Lambda_\alpha \left[ \frac{\delta^2(m-1)s_x^2}{m} \right]^{\frac{m}{m+n}} \left[ \frac{(n-1)s_y^2}{n} \right]^{\frac{n}{m+n}} \]

\[ = \frac{\delta^2(m-1)s_x^2}{m} - \frac{(n-1)s_y^2}{n} \]

where \(\Lambda_\alpha\) is the critical value for \(\Lambda\), asymptotically equal to \(\alpha\) (NOTE 27). The boundary for \(\alpha\) not too small is roughly elliptical in shape, with center at \(\delta=s_x^2/s^2\), \(\delta=Y-\bar{\delta}'X\). Figure 4D illustrates this for the data of the example. The region seems very small, perhaps indicating that the asymptotic approximation is inadequate.

Another approach is to combine the t and F tests previously considered. Since their test statistics are independent under the joint hypothesis of homogeneity (NOTE 29), it follows that if regions with confidence \(1-\alpha_1\) and \(1-\alpha_2\), say, are constructed corresponding to them, then the intersection of these regions has confidence \((1-\alpha_1)(1-\alpha_2)\). Thus the intersection of the 75% confidence regions indicated in Figures 4A and 4C is a region with confidence 52.25%. Alternatively (Peng and Littell, 1976), Fisher's method of combining P-values yields certain advantages as a test: but it is more difficult to use for confidence regions.
SECTION 5

5. Discussion

Having considered the confidence regions related to a number of different tests, the reader may well ask how to choose among them. One criterion which comes to mind is size: surely the best procedures must be those which produce the smallest confidence regions. Defining the size of a confidence region is a tricky proposition, however. To begin with, the region depends on the parametrization of the location-scale difference model so that $Y \sim \delta X + \delta$; but this could have been expressed equally well in many other ways. For example, to the alternative parametrization $X \sim \theta + \eta$ there corresponds a chart of similar structure but in which right and left are reversed, and top and bottom; the bundles of lines become the pencils and vice versa, thus drastically changing the areas and shapes of confidence regions. For this reason attention here has been focused on boundedness: any bounded region is in a sense smaller than any unbounded one, and this property is not affected by the choice of parametrization.

Boundedness, however, is desirable only to the extent that the location-scale difference model is taken for granted. In case of doubt about the model, the set of acceptable parameters for it may more properly be called a consonance region, as defined by Easterling (1976), which indicates whether the model is consonant with the data. Then a small region may result from what Switzer (1976) calls model sensitivity: the ability to reject a false model completely. For example, let $m=n$ with $X_i=n^i$ and $Y_i=x_i$ for $i=1, \ldots, n$: then the pattern of $Y$s and $X$s has at most 8 runs for all $(\xi, \delta)$, no matter what the value of $n$, and thus for any fixed $\alpha$ the corresponding region is totally null if $n$ is sufficiently large (NOTE 30). All the omnibus tests are model sensitive in this sense. But the confidence chart corresponding to a test aimed only at location and/or scale differences always contains a point where the statistic takes on its least significant possible value for the given sample sizes. Thus, even when the model is false, that is, when $Y$ is not distributed as $\delta X + \delta$ for any $(\xi, \delta)$, location tests nevertheless produce horizontal bands across the chart in which one’s confidence is misplaced, and similarly, scale tests produce misleading vertical bands. Even tests aimed at both location and scale differences, such as the tertiles test or the likelihood ratio test, produce false confidence regions when the data do not fit the model.

One conclusion from this study of confidence charts is that, whenever a location-scale difference model is contemplated, the Kolmogorov-Smirnov test should be considered. This omnibus test can be used first to evaluate the fit of the model itself. Then, if the model is adopted, the Kolmogorov-Smirnov confidence region for its parameters compares favorably with alternatives in respect of size (vaguely defined), and in addition has the desirable property of convexity, without being overly difficult to construct. In sum, these findings make the Kolmogorov-Smirnov test extremely attractive.


Mann HB, Whitney DR (1947) A test for whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* 18:50-60.


Statistics 25:139-145.


Wald A, Wolfowitz J (1940) On a test whether two samples are from the same population. Annals of Mathematical Statistics 11:147-162.


1. The principle was already stated by Camp on pages 65-66 of his article "Some recent advances in mathematical statistics, I" in the *Annals of Mathematical Statistics* (1942) 13:62-73 (except that he says "fiducial or confidence", presumably since the terminology was rather confusing in his day). Camp refers in a general way to the original papers by Neyman, but I haven’t located them yet to check on whether the principle goes back even further. The principle seems to have been generally neglected till recently, however; modern applications stem more from Natrelia’s article "The relation between confidence intervals and tests of significance" in the *American Statistician* (February 1960, 20-22 and back cover).

2. Switzer is concerned with general "treatment-effect functions". The generality of his work, however, leaves room for a more detailed consideration of the important special case of location-scale differences.

3. Suppose (without loss of generality) that the original data have the pattern XYYXXXY. And suppose the pattern YZZYYZ is obtained for Z=\(\delta X+\delta\) with some \((x,\delta)\). Since \(Y_1<Z_2=\delta X_1+\delta<Z_2=\delta X_2+\delta<Y_2\) we have \(\delta<(Y_2-Y_1)/(X_2-X_1)<1\). But since \(Z_3<Y_3<Z_3\) we have \(\delta>(Y_3-Y_2)/(X_3-X_2)<1\). Contradiction!

4. Suppose (without loss of generality) that \(X_1 \leq \ldots \leq X_n\) and \(Y_1 \leq \ldots \leq Y_n\). Then let any two of the lines be \(\delta=Y_{i'}-\delta X_i\) and \(\delta=Y_{j'}-\delta X_i\), where \(i'\neq i\). If \(X_i=X_{i'}\), they do not cross anywhere; otherwise they cross at \(x=(Y_{i'}-Y_i)/(X_{i'}-X_i)\). Now, if \(|i+j-1-(m+n)/2|<1\) and \(|i'+j'-1-(m+n)/2|<1\) also, then \((i-j')= (i'-i)\) or \((i'-i+1)\): in any case, \(j-j'\geq 0\), so \(Y_{i'}-Y_i\leq 0\), whence \(\delta \leq 0\). That is, the lines do not cross in the half-plane where \(\delta > 0\).

5. With \(S\) as defined in the first paragraph of Section 3B, the null hypothesis distribution is as follows:

\[
\begin{array}{c|c}
S & \text{probability} \times 1716 \\
-6 \text{ or } 6 & 1 \\
-5 \text{ or } 5 & 6 \\
-4 \text{ or } 4 & 36 \\
-3 \text{ or } 3 & 90 \\
-2 \text{ or } 2 & 225 \\
-1 \text{ or } 1 & 300 \\
0 & 400 \\
\end{array}
\]

For example, the value \(S=-2\) corresponds to the contingency table

\[
\begin{array}{ccc}
\text{below} & \text{median} & \text{above} \\
X & 4 & 0 & 2 & 6 \\
Y & 2 & 1 & 4 & 7 \\
6 & 1 & 6 & 13 \\
\end{array}
\]

with probability \(6!7!6!1!8!/4!0!2!1!4!13! = 225/1716\).


7. The exact number of lines is \(2k'\) where \(k' = \min(k,m,n)\). The lines for which \(i+j=1=k\) do not cross each other, and the lines for which \(i+j-1=m+n-k\) do not cross each other either (see NOTE 4); thus the chart has a bundle of \(k'\) lines crossing another bundle of \(k'\) lines.

8. The shaded area corresponds to the tables.
with hypergeometric probabilities 700/1716 and 525/1716 respectively.

9. The exact number of lines is 2l' where l' = \min(l,m,n). There are usually fewer than \((l'+1)^2\) supertiles, because if \(m+1 \leq 2l'\) then some lines in one bundle (see NOTE 7) are parallel to lines in the other bundle, and if \(n+1 \leq 2l'\) then some lines in one bundle cross lines in the other bundle on the vertical axis.

10. I have no elegant proof of this; I did it by complete exhaustion, essentially. In looking for bounded regions, remember that a tile is unbounded if it touches the vertical axis even only by a corner.

11. The complete distribution of \(\chi^2\) for a 2x3 table with row totals (6,7) and column totals (4,5,4) is as follows:

<table>
<thead>
<tr>
<th>(840 \times \chi^2)</th>
<th>1716 (\times ) prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>104</td>
<td>360</td>
</tr>
<tr>
<td>949</td>
<td>480</td>
</tr>
<tr>
<td>1794</td>
<td>320</td>
</tr>
<tr>
<td>2301</td>
<td>240</td>
</tr>
<tr>
<td>3146</td>
<td>80</td>
</tr>
<tr>
<td>4329</td>
<td>80</td>
</tr>
<tr>
<td>4836</td>
<td>60</td>
</tr>
<tr>
<td>5681</td>
<td>40</td>
</tr>
<tr>
<td>5850</td>
<td>16</td>
</tr>
<tr>
<td>6864</td>
<td>20</td>
</tr>
<tr>
<td>7540</td>
<td>12</td>
</tr>
<tr>
<td>8385</td>
<td>8</td>
</tr>
</tbody>
</table>

12. Suppose the linear rank statistic is

\[ T = \sum_{k=1}^{m+n} a_k w_k, \]

where \(w_k = 1\) if the \(k\)-th smallest observation is a Z, but 0 if it is a Y, and \(a_1, \ldots, a_{mn}\) are the fixed scores for the test. As the line \(\delta = Y_j - \bar{X}_i\) is crossed, the rank of \(Z_i\) changes from \(i+j\) to \(i+j-1\), so the change in \(T\) is

\[ t(i,j) = a_{i+j-1} - a_{i+j}. \]

13. Let \(S\) be the number of \(Z\)s between the two outer quartiles: that is, with rank greater than \(k\) but less than \(m+n+1-k\). Then in the top and bottom tiles \(S=0\) if \(m<k, S=m+n-2k\) if \(n<k\), and \(S=m-k\) otherwise. The change function is

\[ s(i,j) = \begin{cases} -1 & \text{if } i+j-1 = k, \\ 1 & \text{if } i+j-1 = m+n-k, \\ 0 & \text{otherwise} \end{cases} \]

14. I have given no references to published tables. Many tabulations of the Wilcoxon-Mann-Whitney test, the Runs test, and the Kolmogorov-Smirnov test are available.
15. Since the null hypothesis distribution of $G$ is uniform on the integers 0, 1, ..., 7 (Hodges, 1955).


17. This is easily derived from the fact that $#_{YZ}$ and $#_{YV}$ are recursive: each equals 0 in the top and bottom tiles, and their change functions are $(m+1-2t)$ and $(2t-n-1)$, respectively.

18. Since the only published table of the Crouse-Steffens statistic known to me (the one in the original paper) gives only selected critical values rather than the complete distribution, and the only published tables of the Lehmann-Sundrum statistic are limited to $\max(m,n) \leq 5$ (Sundrum, 1954), or $m=n$ (Zajta and Pandikow, Biometrika, 1977), I made my own tabulations. The SAS programs which did this, and the resulting tables, are available on request.

19. An outline of the argument is as follows. Unbounded tiles correspond to patterns with 5 or fewer runs (see paragraph on the runs test), where if there are 5 runs then the middle one must be of length 1. Consider such a pattern:

$$
\begin{align*}
\text{Runs} & \quad Z \quad \text{----------} \quad Z \quad Y \quad \text{----------} \quad Y \quad Z \quad Y \quad \text{----------} \quad YZ \quad \text{----------} \quad Z \\
\text{Lengths} & \quad \frac{m-1}{2}+a \quad \frac{n}{2}+b \quad 1 \quad \frac{n}{2}-b \quad \frac{m-1}{2}-a
\end{align*}
$$

Direct calculation shows

$$
L = \frac{W}{2} - \frac{n(m-1)(2n-1)}{8} + \frac{n(n-1)a^2-2ab(n-1)+(m-1)b^2}{2},
$$

where $W = m(m-1)n(n-1)/4$. This is minimized by making $a$ and $b$ as small as possible: take $a=0$ if $m$ is odd, or $\pm 1/2$ if $m$ is even; take $b=0$ if $n$ is even, or $\pm 1/2$ if $n$ is odd; if $m$ is even and $n$ is odd, take $a$ and $b$ with the same sign. (If the pattern has a $Y$ in the middle instead of a $Z$, then the formula for $L$ is the same with $m$ and $n$ switched.) Write $k = \min(m,n)$ and

$$
C = \begin{cases}
\frac{2k+2d-1}{2(k+d-1)} & \text{if } k \text{ and } d \text{ are both odd,}
\\
\frac{2k+2d+1}{2(k+d)} & \text{if } k \text{ is odd and } d \text{ is even,}
\\
\frac{2(k(k+d)-(2k-1)(k+d))}{2(k-1)(k+d)} & \text{if } k \text{ is even and } d \text{ is odd,}
\\
\frac{2(k(k+d-1)-(2k^3-3d)}{2(k-1)(k+d-1)} & \text{if } k \text{ and } d \text{ are both even.}
\end{cases}
$$

In each of the four cases the smallest value of $L$ for given $k$ is obtained by making $d$ as small as possible ($d=0$ if even and $d=1$ if odd), producing

$$
L / W \geq \frac{1}{2} - \frac{2k+1}{2k^2} \quad \text{if } k \text{ is odd}
$$
The first of these two cases is the worse, and its bound is the one quoted in the text; there seemed no point in providing a more precise result. Under the null hypothesis, L/W has expectation 1/3 and variance \((m+n+1)(m+n-2)/45w = 0(1/k^2)\), so the probability of \(L/W\) exceeding the bound approaches 0 as \(k\) increases without limit.

20. For the general pattern of NOTE 19 (with a \(Z\) in the middle)

\[
mnK = \max \left\{ \frac{m-1}{2} + a, \frac{m+1}{2} - a \right\}
\]

This is minimized at \(a = 1/2\) if \(m\) is even and at \(a = 0\) or 1 if \(m\) is odd, giving \(mnK \geq [m/2]/2\). Putting a \(Y\) in the middle causes \(m\) to switch with \(n\); thus the lower bound is as stated.

21. The Moses test is usually stated in terms of ranges of ranks, but this formulation is equivalent.

22. Under the null hypothesis of homogeneity,

\[
P \left( N_k = w \right) = \binom{m-2k+w}{m-2k} \binom{n-w}{i=i} \frac{(n-w+i+k-1)}{k} \left/ \binom{m+n}{n}\right.
\]

putting \(k=1,\)

\[
P \left( N_1 = w \right) = \frac{(n-w+1)}{m-2k} \left/ \binom{m+n}{n}\right.
\]

for \(w=1,\ldots,n\). I don't know why Hollander (1963, p. 395) calls this distribution "tedious to calculate": it's certainly much simpler than that of his own extreme reactions statistic. The full distribution for \(N_1\), with \((m,n)=(6,7),\) is

| \(|N_1|\) | prob. \(x 1716\) |
|---|---|
| 0 | 8 |
| 1 | 35 |
| 2 | 90 |
| 3 | 175 |
| 4 | 280 |
| 5 | 378 |
| 6 | 420 |
| 7 | 330 |

23. It turns out that

\[
H = \frac{\left(m^2-1\right)(m+3n)}{12} - \frac{(m+1)L-(m+n)Q-(n-1)T}{m(n-1)}
\]

where \(L, Q,\) and \(T\) are the Lehmann-Sundrum, Crouse-Steffens, and Tamura statistics already discussed. I worked out this formula because I had some difficulty with Hollander's statement that \(H\) is a U-statistic (p. 397 of his article). We see that \(H\) is a U-statistic only in a weaker sense than is usually meant: for fixed \(m\) and \(n\) it is of the U-statistic form, but its expectation depends on \(m\) and \(n\). In this weak sense, Spearman's rho is a U-statistic also, but the claim always made for it is only that it is a linear combination of two U-statistics -- Kendall's tau and the grade correlation. It would have been better to say only that \(H\) is asymptotically, or asymptotically equivalent
24. Since Hollander provides only the 10%, 5% and 1% critical values, I made my own tabulation for \((m,n)=(6,7)\). The lower tail of the null hypothesis distribution is as follows:

<table>
<thead>
<tr>
<th>6 x H</th>
<th>1716 x prob.</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>105</td>
<td>8</td>
<td>.0047</td>
</tr>
<tr>
<td>140</td>
<td>14</td>
<td>.0128</td>
</tr>
<tr>
<td>161</td>
<td>14</td>
<td>.0210</td>
</tr>
<tr>
<td>168</td>
<td>7</td>
<td>.0251</td>
</tr>
<tr>
<td>177</td>
<td>6</td>
<td>.0286</td>
</tr>
<tr>
<td>185</td>
<td>12</td>
<td>.0355</td>
</tr>
<tr>
<td>200</td>
<td>12</td>
<td>.0425</td>
</tr>
<tr>
<td>204</td>
<td>12</td>
<td>.0495</td>
</tr>
<tr>
<td>209</td>
<td>12</td>
<td>.0565</td>
</tr>
<tr>
<td>224</td>
<td>10</td>
<td>.0624</td>
</tr>
<tr>
<td>225</td>
<td>6</td>
<td>.0659</td>
</tr>
<tr>
<td>233</td>
<td>12</td>
<td>.0728</td>
</tr>
<tr>
<td>236</td>
<td>12</td>
<td>.0798</td>
</tr>
<tr>
<td>240</td>
<td>10</td>
<td>.0857</td>
</tr>
<tr>
<td>245</td>
<td>10</td>
<td>.0915</td>
</tr>
<tr>
<td>249</td>
<td>16</td>
<td>.1008</td>
</tr>
<tr>
<td>252</td>
<td>5</td>
<td>.1038</td>
</tr>
</tbody>
</table>

Thus the value Hollander gives as the 10% point \((H=41.50, \text{ or } 6xH=249, \text{ with } P=.100)\) is not quite accurate. \(P\) is slightly greater than .100 for his \(H\), .101 to 3 decimal places, and the true 10% point is \(H=40.83\), or \(6xH=245\), with \(P=.0915\).

25. If \(|D|\) (or \(|1|\)) is used as the criterion, we have

\[
D_k(\delta, \gamma) = \left| (\delta \bar{X} + \delta - \overline{Y}) - \left[ \frac{1}{m} + \frac{1}{n} \right] u_k (\delta \bar{X}_k + \delta - \overline{Y}_k) \right|
\]

\[= \left| D_1(\delta, \gamma) - A_s(\delta, \gamma) \right|, \text{ say.}\]

This equals \(D_1(\delta, \gamma)\) if \(A_s(\delta, \gamma)=0\) or if \(A_s(\delta, \gamma)=2D_1(\delta, \gamma)\), that is, if

\[(1) \quad \delta = \gamma - \delta x_s\]

or

\[(2) \quad \delta = \frac{2(\overline{Y} - \delta \bar{X}) - u_k \left[ \frac{1}{m} + \frac{1}{n} \right] (\overline{Y}_k - \delta \bar{X}_k)}{2 - u_k \left[ \frac{1}{m} + \frac{1}{n} \right]}\]
Thus, the "curves" referred to in the text are pairs of straight lines, and the complete confidence chart has 2(L-1) lines rather than only (L-1). There is a simplification if m=n; then the second solution for the k-th reallocation becomes

$$
\delta = \frac{2(\bar{Y} - \bar{x}) - \frac{2u_k}{n} (\bar{y}_k - \bar{x}_k)}{2 - \frac{2u_k}{n}}
$$

$$
= \frac{n(\bar{y}_k - \bar{x}_k) - u_k (\bar{y}_k - \bar{x}_k)}{n - u_k}
$$

$$
= \left(\text{mean of } Ys \text{ not becoming } \text{ in } k-th \text{ reallocation} \right) - \gamma \left(\text{mean of } Xs \text{ not becoming } \text{ in } k-th \text{ reallocation} \right)
$$

and thus the set of second lines is the same as the set of first lines: there are only (L-1) lines, and A changes by 2 when one of them is crossed. It is interesting also to note that, if m=n, the median of the quantities (\(Y_k-\bar{x}_k\)) for any given \(\gamma\), which might be taken as a point estimate of \(\delta\) for that \(\delta\) since it lies within every confidence interval, is equal to (\(\bar{Y}-\bar{x}\)).

26. Clearly, \(t'_{\alpha}(\gamma) \leq t_{\alpha}\), since the degrees of freedom are fewer, and this tends to make the regions larger; but in the example the effect is not great, since \(t'_{\alpha}(\gamma)\) lies between the critical value .697 for 11 DF and the critical value .727 for 5 DF. The denominator of \(t'\) is larger [smaller] than that of \(t\) if \((n-m)(s^2_2-s^2_1)\) is negative [positive]; but in the example n is not very much different from m, and \(s^2_2\) not very much different from \(s^2_1\); so again the effect on the region is not great.

27. The computation of the value of \(\gamma\) for the chart is as follows:

<table>
<thead>
<tr>
<th>P</th>
<th>.050</th>
<th>.125</th>
<th>.250</th>
<th>.750</th>
<th>.875</th>
<th>.950</th>
</tr>
</thead>
<tbody>
<tr>
<td>F*</td>
<td>.202</td>
<td>.335</td>
<td>.528</td>
<td>1.785</td>
<td>2.751</td>
<td>4.387</td>
</tr>
<tr>
<td>(\gamma)</td>
<td>.253</td>
<td>.418</td>
<td>.660</td>
<td>2.232</td>
<td>3.439</td>
<td>5.484</td>
</tr>
<tr>
<td>(\gamma)</td>
<td>.503</td>
<td>.647</td>
<td>.812</td>
<td>1.494</td>
<td>1.855</td>
<td>2.342</td>
</tr>
</tbody>
</table>

28. It is, of course, more customary to say that \(-2 \log(\Lambda)\) is approximately distributed as \(\chi^2\) with 2 degrees of freedom, but it is equivalent to say that \(\Lambda\) itself is approximately uniform on (0,1), and it saves writing down "log".

29. This is shown by Perng and Littell (1976).

30. (PROOF) Define n+1 intervals, in order, as follows: \(-\infty\) to \(Y_1=1\); \(Y_1=1\) to \(Y_2=2\); \(\ldots\) \(Y_{n-1}=n-1\) to \(Y_n=n\); \(Y_n=n\) to \(+\infty\). Each interval is "occupied" by at least one Z, or else it is "empty". Let \(Z_1\) be the smallest Z in the second occupied interval, and \(Z_2\) the smallest Z in the third occupied interval. Then \(Z_{n-1} = 1\) because they are separated by the second occupied interval, which has length 1; and \(Z_{n-1} = \gamma R(n^{*1+n-1}-1)\). Now consider \(Z_{n+1}\). If it doesn't exist, then there are at most 3 occupied intervals and hence at most 7 runs. If it does exist, then

- 23 -
\[ Z_{k+1} - Z_j = \delta n^j(n^{k+1-j} - 1) = n(Z_i - Z_j - 1) \geq n; \]

so \( Z_{k+1} > Z_j + n \), but \( Z_j > 1 \) since it is in the second occupied interval. Thus \( Z_{k+1} \) is in the last interval, and there are at most 4 occupied intervals, with no \( Y \) beyond the last \( Z \); so there are at most 8 runs.
FIGURE 1: THE MOSAIC
Figure 20: Outer Quartiles

$P = 0.500 \rightarrow 71.4\%$ confidence
\[ \chi^2 \leq 2.14 \rightarrow 67.6\% \text{ confidence} \]
FIGURE 2D: WILCOXON-MANN-WHITNEY  \[ \frac{\text{#_xy}}{26} \rightarrow 55.5\% \text{ confidence} \]
Figure 2E: Galton
$(Y_4 = 4$ omitted$)$

$1 \leq G \leq 5 \rightarrow 71.4\%$ confidence
Figure 2F: Freund-Ansari-Bradley

$21 \leq 25 \rightarrow 50.7\%$ confidence
Figure 2K: moses

$N_t > 5 \rightarrow 65.7\%$ confidence
\[0.528 \leq F \leq 1.785 \quad \{ \rightarrow 50\% \text{ confidence} \}
\]

\[0.812 \leq \theta \leq 1.994\]

50\% confidence

75\%

90\%
Figure 4D: Likelihood Ratio