SA

stichting
mathematisch
centrum

MC

AFDELING MATHEMATISCHE STATISTIEK SW 17/72 JUNE

D. QUADE

SA

THE PAIR CHART

Prepublication

2e boerhaavestraat 49 amsterdam

Printed at the Mathematical Centre, 49, 2e Boerhaavestraat 49, Amsterdam.

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.), by the Municipality of Amsterdam, by the University of Amsterdam, by the Free University at Amsterdam, and by industries.

THE PAIR CHART

by

Dana Quade
University of North Carolina (temporarily at Mathematical Center, Amsterdam)

1. INTRODUCTION

Let $X_1, X_2, \ldots, X_{n_X}$ be a random sample of n_X observations on a variable X with unknown distribution function F_X, and let $Y_1, Y_2, \ldots, Y_{n_Y}$ be a random sample of n_Y observations on a variable Y with unknown distribution function F_Y. To compare such samples, and in particular to test the null hypothesis $H_0 : F_X = F_Y$, is the classic "two-sample problem". The purpose of this paper is to show how a certain diagram, which may be called a "pair chart", can give insight into the problem, and in some cases facilitate the computations required.

A pair chart is constructed as follows. Draw a rectangle of width n_X units and height n_Y units. If the smallest observation in the combined samples is an X, draw a line from the lower left corner of this rectangle one unit to the right; if it is a Y, draw the line one unit up instead. From the end of this first line draw a second line, one unit to the right if the second smallest observation is an X, and one unit up if it is a Y. Continue in the same manner for all (n_X+n_Y) observations. The (n_X+n_Y) line segments then form a path from the lower left corner of the rectangle, which may be designated as the origin (0,0), to the upper right corner (n_X,n_Y).

Thus for example suppose we have the data below:

(A)

X	18, 20, 30, 32, 36, 38, 39, 41, 51, 70	$n_X = 10$
Y	28, 43, 46, 46, 50, 56, 64, 79	$n_Y = 8$

The ordering of the combined samples is

X X Y X X X X X Y Y Y Y X Y Y X Y ,

and the path is as shown in Chart A. (Ignore the shading, to be explained later.)

A second example may be based on these data:

(B)

X	19, 25, 28, 30, 30, 36, 50, 52, 57, 67	$n_X = 10$
Y	24, 31, 33, 37, 38, 42, 49	$n_Y = 7$

.

Here the ordering of the combined samples is

X Y X X X X Y Y X Y Y Y Y X X X X ,

and the path is as shown in Chart B.

Ties within one or both of the samples, as have already been encountered, cause no difficulty whatever; but between-sample ties complicate the situation somewhat, since they make it impossible to determine the path unambiguously. Thus suppose we have data as follows:

(C)

X	1, 2, 2, 2, 3, 4, 4, 7	$n_X = 8$
Y	1, 2, 2, 3, 3, 3, 5, 9	$n_Y = 8$

.

The corresponding ordering is

(XY) (XXXYY) (XYYY) X X Y X Y ,

where the observations within each pair of parentheses are all equal. At each such tie there are several possible routes for the path, depending on how the tie is resolved; these routes cover a box, as is shown in Chart C. For some purposes the diagonals of such boxes may be used to form a unique path; in general the whole box must be considered.

The name "pair chart" is derived from the interpretation of each unit square within the rectangle as one of the $n_X n_Y$ possible pairs of observa-

Chart A

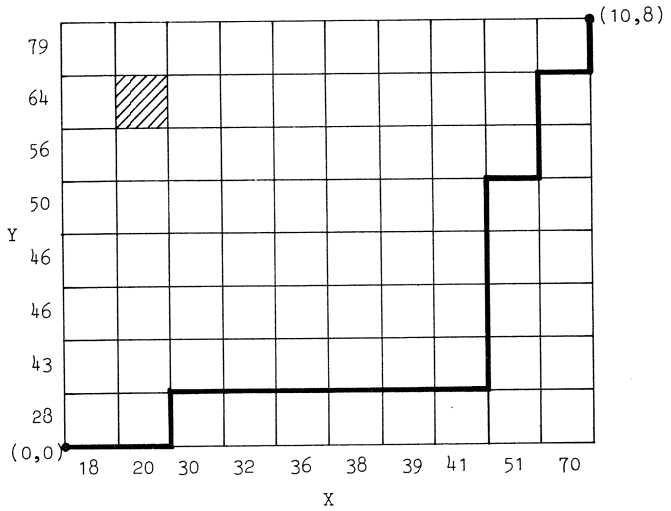


Chart B

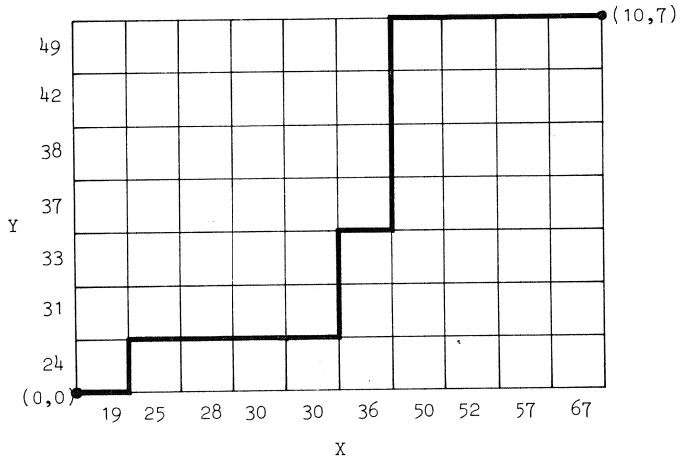
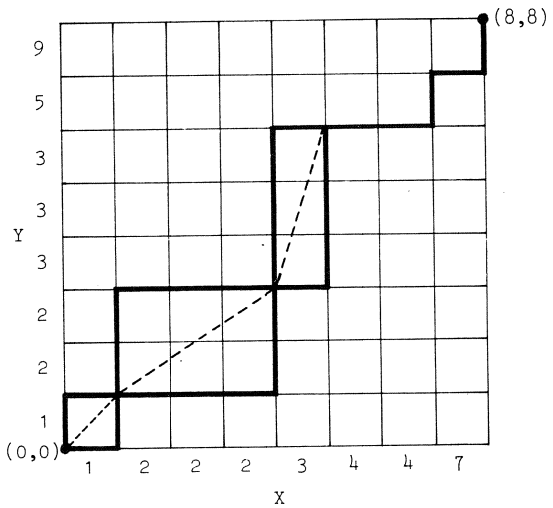


Chart C



tions (X_i, Y_j) such that the pair includes 1 X and 1 Y. Thus for example the shaded square in Chart A corresponds to the pair $X_2 = 20$, $Y_7 = 64$. The pair chart seems to have been invented by Drion [3], who used it only as a tool in studying the distribution of the Kolmogorov-Smirnov test statistic (see Section 3), although he was aware of other interpretations. Pair charts have appeared sporadically in the literature since Drion's paper, but till now they have received no unified treatment.

The applications of the pair chart are of at least three types: (i) as a descriptive representation by which the two samples can be roughly compared at a glance; (ii) as an aid in calculating or interpreting various test statistics; and (iii) as an aid in studying the distribution of such statistics, and particularly in computing their significance levels. Applications of type (i) will be discussed in Section 2. The remaining sections will treat the various test statistics, giving applications of types (ii) and (iii), of which some are new here. However, there will be no attempt to discuss the theory of the various tests, or their relative merits.

2. DESCRIPTIVE USES

Suppose the two samples are such that the X's are on the whole smaller than the Y's, and thus generally come earlier in the ordering of the combined samples. Then in constructing the pair chart most of the lines which go to the right will be drawn before those which go up, and the path will lie below the diagonal of the rectangle. This is exactly the situation of Chart A. In the opposite situation, where the X's are generally larger than the Y's, the path will lie above the diagonal of the rectangle.

On the other hand, suppose the X's are neither larger nor smaller than the Y's on the whole, but are more dispersed. Here the ordering of the combined samples will tend to show X's first, with more Y's in the middle, and then X's again at the end, so that the path starts out generally to the right, then moves up rapidly to cross the diagonal of the rectangle, and finally moves to the right again. This is the situation of Chart B. Of course, the path will cross the diagonal horizontally, from left to right, if the X's are less dispersed than the Y's.

But if the X's and Y's are well-mixed, as may be expected if $F_X \equiv F_Y$, then the whole path is likely to lie fairly close to the diagonal of the rectangle. This is exemplified by Chart C. As will be seen, various tests of the hypothesis $H_0 : F_X \equiv F_Y$ may be obtained by agreeing to reject if the path lies too far, in some appropriate sense, from this diagonal.

Further insight may be attained by considering what will happen in large samples. For this it is convenient to standardize the pair chart. We rescale it so that each point (x,y) becomes the point $(x/n_X, y/n_Y)$: then the original rectangle becomes the standard square with corners (0,0), (0,1), (1,0), (1,1), and what originally were unit squares representing the pairs now become rectangles of width $1/n_X$ and height $1/n_Y$ each. Such a standardized pair chart was discussed by Wilk & Gnanadesikan [15], who called it a "percent plot" or "P-P plot". In the limit as n_X and n_Y become infinite the path of the standardized pair chart will approach the "relative distribution function", which is the locus of points $(F_X(z), F_Y(z))$ for $-\infty < z < \infty$. This coincides with the diagonal $x = y$ of the standard square if and only if $F_X \equiv F_Y$.

Figure 1 illustrates the effects which differences in mean and variance between X and Y may be expected to have on pair charts. It shows the relative distribution functions of some nonstandard normal variables with respect to a standard normal variable. Figure 2 illustrates the effects of differences in distributional form, by showing the relative distribution functions of nine nonnormal variables (summarized in Table 1) with respect to normal variables. The tertiles have been made to agree in each case, so that the paths all pass through four equally-spaced points on the diagonal: namely, (0,0), $(^1/3, ^1/3)$, $(^2/3, ^2/3)$, and (1,1). When the nonnormal variable has a symmetric distribution, the curve also passes through the point $(^1/2, ^1/2)$. A little study of these two Figures should give a good basis for interpreting the different patterns found in pair charts of sample data.

Figure 1

Relative distribution functions of some nonstandard normal variables with respect to a standard normal variable ($\mu_X = 0$, $\sigma_X = 1$)

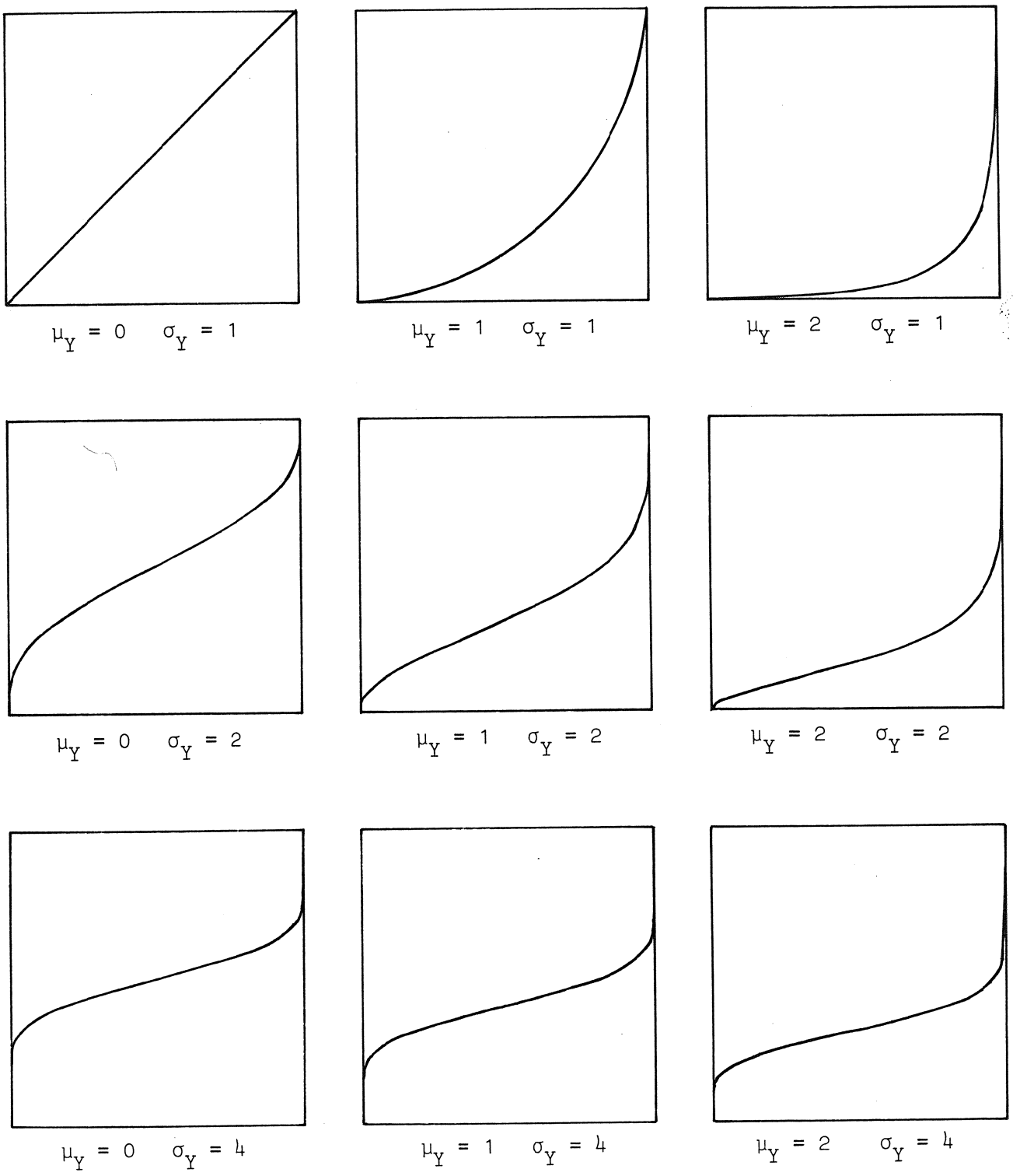


Figure 2

Relative distribution functions of some nonnormal variables with respect to normal variables having the same tertiles

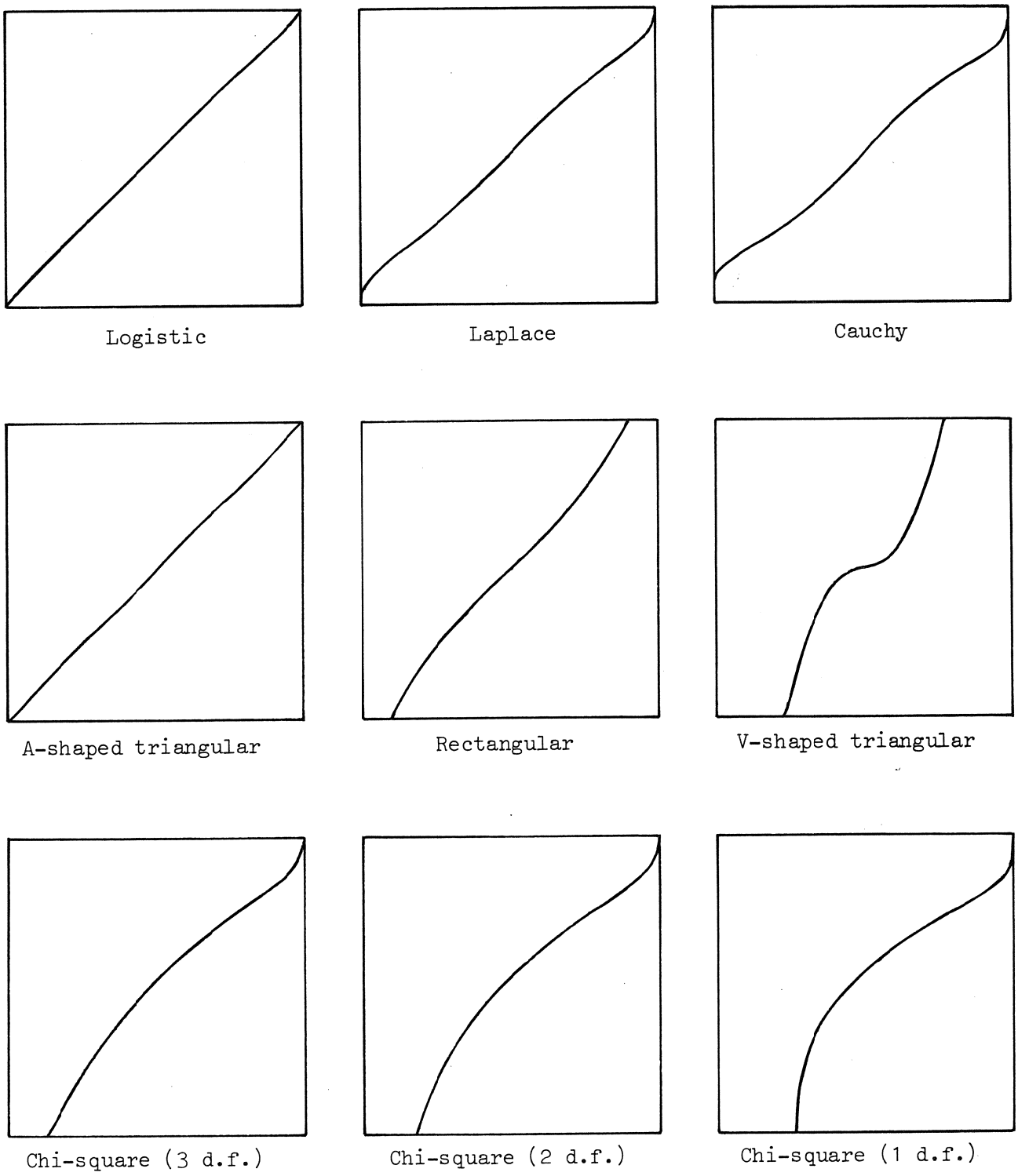


Table 1

	Name of nonnormal variable	Density function, f(x)	Tertiles
1.	Logistic	$e^{-x}(1+e^{-x})^{-2}$, $-\infty < x < \infty$	-.6931, + .6931
2.	Laplace	$\frac{1}{2}e^{-\lvert x\rvert}$, $-\infty < x < \infty$	-.4055, + .4055
3.	Cauchy	$(1+x^2)^{-1}$, $-\infty < x < \infty$	-.5774, + .5774
4.	A-shaped triangular	$1 - \lvert x\rvert$, $-1 < x < 1$	-.1835, + .1835
5.	Rectangular	$\frac{1}{2}$, $-1 < x < 1$	-.3333, + .3333
6.	V-shaped triangular	$\lvert x\rvert$, $-1 < x < 1$	-.5774, + .5774
7.	$\chi^2(3)$	$e^{-x/2}\sqrt{\frac{\pi x}{2}}$, $0 < x < \infty$	1.5680, 3.4047
8.	$\chi^2(2)$	$\frac{1}{2}e^{-x/2}$, $0 < x < \infty$	.8109, 2.1972
9.	$\chi^2(1)$	$e^{-x/2} / \sqrt{2\pi x}$, $0 < x < \infty$	.1855, .9359

3. THE KOLMOGOROV-SMIRNOV TESTS

Define the "empirical distribution functions" $F_X^{(n_X)}(z)$ and $F_Y^{(n_Y)}(z)$ at each value of z, $-\infty < z < \infty$, by the following relationships:

$$n_X F_X^{(n_X)}(z) = \text{number of observations } X_i \text{ such that } X_i \leq z$$

$$n_Y F_Y^{(n_Y)}(z) = \text{number of observations } Y_j \text{ such that } Y_j \leq z.$$

Then the one-sided Kolmogorov-Smirnov tests reject $H_0 : F_X \equiv F_Y$ for large values of the statistics

$$D_X = \sup_{-\infty < z < \infty} \left[F_X^{(n_X)}(z) - F_Y^{(n_Y)}(z) \right]$$

and

$$D_Y = \sup_{-\infty < z < \infty} \left[F_Y^{(n_Y)}(z) - F_X^{(n_X)}(z) \right],$$

and the two-sided test rejects for large values of

$$D = \max (D_X, D_Y).$$

The computational method usually proposed for these tests - see for example Siegel[9] - involves actually drawing the two empirical distribution functions and then inspecting the vertical distance between them. This may, of course, involve considerable labor. A more convenient computational method, first made explicit by Hodges [4], is as follows: let (X_X, Y_X) be the point on the path farthest below the diagonal of the rectangle (if two or more points are equally far below it, any one of them may be used), and let (X_Y, Y_Y) be the point farthest above the diagonal; then

$$D_X = \left| \frac{X_X}{n_X} - \frac{Y_X}{n_Y} \right|, \quad D_Y = \left| \frac{X_Y}{n_X} - \frac{Y_Y}{n_Y} \right|.$$

To illustrate, observe that the point on the path farthest below the diagonal in Chart A is $(X_X, Y_X) = (8,1)$, and thence calculate $D_X = (8/10 - 1/8) = .675$. The path is nowhere above the diagonal, so (X_Y, Y_Y) may be taken as $(0,0)$ or $(10,8)$, yielding $D_Y = 0$ either way; and $D = \max(.675,0) = .675$. Similarly in Chart B we have $(X_X, Y_X) = (5,1)$, so $D_X = .358$, and $(X_Y, Y_Y) = (6,8)$, so $D_Y = .400$. To understand the method, note that the vertical distance from the diagonal of the rectangle to any point (x,y) is $n_Y \, |y/n_Y - x/n_X|$. But the lattice points of the path constitute the locus of points $(n_X F_X^{(n_X)}(z), n_Y F_Y^{(n_Y)}(z))$ for $-\infty < z < \infty$. Thus the maximum value of $|F_X^{(n_X)}(z) - F_Y^{(n_Y)}(z)|$ is given by the maximum of $|y/n_Y - x/n_X|$ for lattice points (x,y) on the path. Intuitively, the Kolmogorov-Smirnov statistics measure the discrepancy of the data from H_0 in terms of the maximum distance from the path to the diagonal of the rectangle.

Where there are between-sample ties, the test statistics as defined above are obtained by letting the path follow the diagonals of the boxes. For example, in Chart C we can take (X_X,Y_X) as (4,3), (7,6) or (8,7), with $D_X = .125$, and we have $(X_Y,Y_Y) = (5,6)$, with $D_Y = .125$ also. It should be clear that a test statistic thus calculated is as small as could possibly result from any possible resolution of the ties. Thus the corresponding test becomes conservative, in that its Type I error does not exceed the value found in tables; but of course it also loses power. The maximum value of D_X can be found by letting the path follow the bottom and right edges of the boxes, which gives $(X_X,Y_X) = (4,1)$ and $D_X = .375$ for Chart C; similarly, the maximum value of D_Y can be found by using the top and left edges of the boxes, giving $(X_Y,Y_Y) = (4,6)$ and $D_Y = .250$ for Chart C.

The labor required by the standard computational method makes it tempting to perform a preliminary grouping of the data; for instance, Siegel [9] does so in both of the examples he presents. But grouping is undesirable since it tends to increase the number of between-sample ties. With the method based on the pair chart, however, the computational effort is so reduced that grouping should no longer be necessary.

Finally, Hodges [4] presents various methods which use the pair chart to calculate the level of significance corresponding to any observed value of one of these test statistics. The simplest of these is as follows. A particular path gives rise to a value of $D \geq c/n_X n_Y$ if and only if it reaches one or the other of the two parallel lines determined by $|xn_Y - yn_X| = c$. Let $H(x,y)$ be the number of possible routes from the origin to the point (x,y) which do not reach either of these lines; in particular, $H(n_X,n_Y)$ is the number of possible paths which give rise to values of D less then $c/n_X n_Y$. Now, the total number of possible paths on the pair chart is $(n_X+n_Y)\ !\ /\ n_X\ !\ n_Y\ !$, and under H_0 these are all equally likely; thus the significance level associated with $D = c/n_X n_Y$ is

$$P = 1 - \frac{H(n_X,n_Y)\ n_X\ !\ n_Y\ !}{(n_X+n_Y)\ !} \quad .$$

The values of $H(x,y)$ can be built up using the recursion formula

$$H(x,y) = H(x-1,y) + H(x,y-1)$$

with the initial condition $H(0,0) = 1$ and the boundary conditions that $H(x,y) = 0$ for $x < 0$, $y < 0$, or $|xn_Y - yn_X| \geq c$.

The procedure is illustrated in Figure 3, which correponds to Chart B, where $n_X = 10$, $n_Y = 7$, and $D = .400$, giving $c = n_X n_Y D = 28$. Next to each lattice point (x,y) between the lines $|7x - 10y| = 28$ has been written the corresponding quantity $H(x,y)$. The total number of possible paths is $17\,!\,/\,7\,!\,10\,! = 19448$, of which $H(10,7) = 11019$ produce values of $c < 28$, or $D < .400$; hence $P = 1 - 11019/19448 = .4344$.

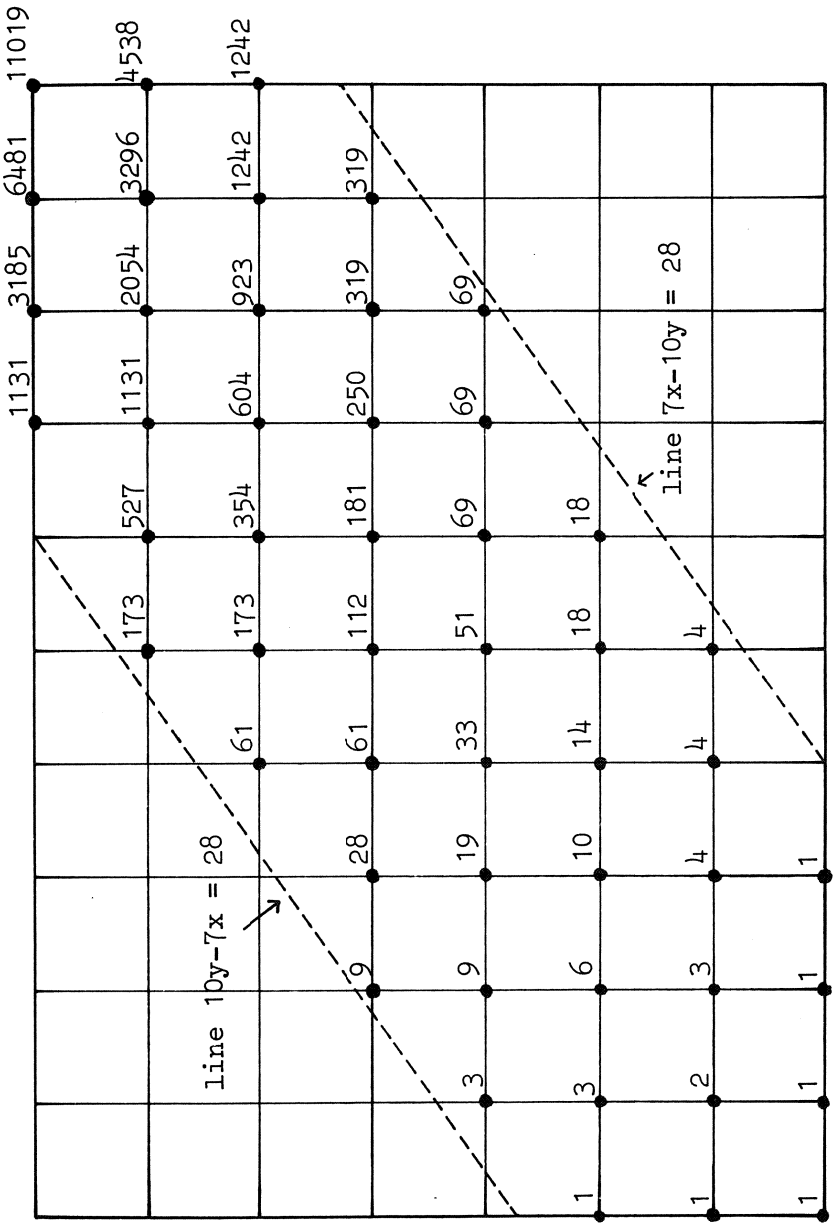

11019
4538
1242
6481
3296
1242
319
3185
2054
923
319
69
1131
1131
604
250
69
69
527
354
181
69
18
173
173
112
51
18
4
61
61
33
14
4
28
28
19
10
4
1
9
9
6
3
1
3
3
2
1
1
1
1
line 10y-7x = 28
line 7x-10y = 28

Figure 3

A short cut follows from noticing that the number of possible paths through a given point (x,y) is the number of routes from the origin to (x,y), namely H(x,y), multiplied by the number of routes from (x,y) to the point (n_X,n_Y), which by symmetry is $H(n_X-x, n_Y-y)$. Hence the recursion formula need not be carried out beyond the line $x + y = (n_X+n_Y+1) / 2$. Thus in Figure 3

$$H(10,7) = H(4,5)\ H(6,2) + H(5,4)\ H(5,3) + H(6,3)\ H(4,4)$$

$$= (61)(18) + (112)(51) + (69)(61)$$

which gives 11019 as before. See Hodges [4] for further methods more suitable for D_X and D_Y.

4. RUNS

If there are no between-sample ties, the path of the pair chart consists entirely of horizontal and vertical line segments, each corresponding to a "run" of observations from the same sample. Wald & Wolfowitz [13] proposed rejecting $H_0 : F_X \equiv F_Y$ for small observed values of R, the total number of runs; they showed, under fairly general conditions, that the expected value of R is maximized if H_0 is true. Some insight into this result may perhaps be obtained from contemplating the pair chart: if $F_X \equiv F_Y$ then the path must be close to the diagonal of the rectangle, and in order to lie as close as possible it must be made up of many short segments which continually cross and recross the diagonal; if it were made up of fewer and longer segments then it would have to lie farther away on the whole.

In charts A and B, which are without ties, there are 8 and 7 runs, respectively. But the occurrence of ties tends to produce a particularly great ambiguity in the value of R, rendering the test unsatisfactory. Thus in Chart C the number of runs might be any integer from 7 to 14 - a range of nearly 4 standard deviations under H_0 - depending on how the ties are resolved.

5. THE WILCOXON AND MANN-WHITNEY TESTS

The $n_X n_Y$ pairs (X_i, Y_j) may be classified into three groups as follows: pairs such that $X_i > Y_j$, which correspond to squares lying below the path of the pair chart (on the X-side of it); pairs such that $X_i < Y_j$, which correspond to squares lying above the path (on the Y-side); and tied pairs, such that $X_i = Y_j$, which correspond to squares lying within boxes. Thus the area U_X (or, U_Y) within the rectangle and below (or, above) the path, where the path is taken to follow the diagonals of the boxes if there are between-sample ties, is equal to the number of pairs such that $X_i > Y_j$ (or, $X_i < Y_j$), plus half the number such that $X_i = Y_j$. That is, U_X and U_Y are the familiar test statistics of Mann & Whitney [7]. This relationship was already known to Drion [3].

In Chart A we find $U_X = 18$, $U_Y = 62$; and in Chart B, $U_X = U_Y = 35$. In Chart C we have $U_X = 25 + (\frac{1}{2})(10) = 30$, $U_Y = 29 + (\frac{1}{2})(10) = 34$, where 10 is the number of tied pairs, equal to the combined areas of the boxes, and 25 and 29 are the numbers of untied pairs such that $X_i > Y_j$ and $X_i < Y_j$ respectively. Of course, the extreme values attainable with any resolution of ties could also be calculated easily, by assigning the whole area of the boxes first to U_X and then to U_Y. Note that $U_X + U_Y \equiv n_X n_Y$.

The pair chart also provides a nice illustration of the equivalence between the test statistics of Mann & Whitney [7] and Wilcoxon [14]. For simplicity, assume there are no ties, and let R_i be the rank of X_i in the combined samples. Then

$$\begin{aligned} R_i &= 1 + (\text{number of observations in combined samples less than } X_i) \\ &= 1 + (\text{number of X's less than } X_i) + (\text{number of Y's less than } X_i) \\ &= (\text{rank of } X_i \text{ in sample of X's}) + (\text{number of Y's less than } X_i), \end{aligned}$$

and if the X's have been ordered

$$R_i = i + (\text{number of Y's less than } X_i).$$

But the columns of the pair chart correpond to the ordered X's, and the number of Y's less then X_i is the number of squares below the path in the corresponding column, say B_i. Write three rows under the pair chart, the first containing the quantities i, the second B_i, and the third $R_i = i + B_i$. The sums of these rows are

$$\sum_{i=1}^{n_X} i = n_X(n_X+1) / 2,$$

$$\sum_{i=1}^{n_X} B_i = U_X \quad \text{(the Mann-Whitney statistic)},$$

and

$$\sum_{i=1}^{n_X} R_i = T_X \quad \text{(the Wilcoxon statistic)}.$$

For Chart A these rows would be

i	1	2	3	4	5	6	7	8	9	10	55
B_i	0	0	1	1	1	1	1	1	5	7	18
R_i	1	2	4	5	6	7	8	9	14	17	73

Thus is verified the equation

$$T_X = U_X + n_X(n_X+1) / 2.$$

The relationship between T_Y and U_Y could be illustrated similarly, using the rows of the pair chart.

Finally, it may be mentioned that Klotz[5] uses a pair chart as an aid in calculating the distributions of these test statistics in the presence of ties.

6. SCALE TESTS BASED ON PAIRS

For each z, $-\infty < z < \infty$, let the quantity

$$S(z) = \sum_{i=1}^{n_X} \sum_{j=1}^{n_Y} \psi(X_i, Y_j, z) ,$$

where

$$\psi(x,y,z) = \begin{cases} 1 & \text{if } y < x < z \text{ or } z < x < y \\ 1/2 & \text{if } x = z \neq y \text{ or } y = z \neq x \\ 1/4 & \text{if } x = y = z \\ 0 & \text{otherwise.} \end{cases}$$

If there are no ties, then $S(z)$ is the number of ways in which it is possible to choose an X and a Y from the data such that the X lies between the Y and the constant z. If ties are present, they lead to fractional counts as indicated in the definition of ψ. These fractions have been determined so that if Z_1 and Z_2 are independent observations on any random variable Z, continuous or not, then

$$E\{\psi(Z_1, Z_2, z)\} = \frac{1}{2} - P\{Z < z\} P\{Z > z\} - \frac{1}{4} P^2\{Z = z\}$$

It is possible to obtain $S(z)$ from the pair chart as follows. Draw the line $x = k(z)$, where $k(z)$ is the number of observations X_i such that $X_i < z$, plus half the number such that $X_i = z$. Then $S(z)$ is the area within the rectangle of the pair chart to the left of the line and below the path, plus the area to the right of the line and above the path, where again the path is taken to follow the diagonals of the boxes if there are between-sample ties. For example, consider calculating $S(3)$ in Chart C. We have 4 X's with values less than 3, and 1 X with value equal to 3, so $k(3) = 4.5$. The area to the left of the line $x = 4.5$ and below the path is 8.375, and the area to the right of the line and above the path is 6.375, so $S(3) = 8.375 + 6.375 = 14.75$.

Referring back to the interpretations of pair chart patterns as given in Section 2, it may be seen that if the X's and Y's do not differ greatly in location then S(z) will tend to be small (or, large) if the Y's are less (or, more) variable than the X's. Thus it seems reasonable to use S(z), for some suitable value of z, as a statistic for testing $H_0 : F_X \equiv F_Y$ against alternatives which imply differences of scale between the X- and Y-populations. In particular, suppose the true common median under H_0, say μ, is known. Then $S(\mu)$ is the test statistic proposed for this "scale problem" by Sukhatme [12], except that he made no suggestions for dealing with ties.

On the other hand, Ansari & Bradley [1] have proposed a test against scale differences which does not require knowledge of the true common median under H_0. Their statistic is

$$W = \sum_{i=1}^{n_X} \min (R_i, n_X+n_Y+1-R_i) ,$$

where R_i is again the rank of X_i in the combined samples. This can also be calculated from the pair chart, using the alternate formula (to which there are certain exceptions noted below)

$$W = S(m) - k(m) [n_X - k(m)] + n_X[n_X+1] / 2,$$

where m is the median of the combined samples. Thus in Chart A we have $m = 42$, with $k(m) = 8$, and $S(m) = 6 + 4 = 10$, giving

$$W = 10 - (8) (2) + (10) (11) / 2 = 49,$$

which checks with the defining formula in terms of ranks.

If (n_X+n_Y) is odd, and if in addition the median observation is an X, then a correction of .25 must be added in the alternate formula. Thus in Chart B there are $n_X+n_Y = 10 + 7 = 17$ observations in the combined sample, and their median $m = 36$ is the value of X_5, We then have $k(m) = 5.5$ and $S(m) = 5.5 + 2 = 7.5$, so

$$W = 7.5 - (5.5) (4.5) + (10) (11) / 2 + .25 = 38.$$

This unlovely complication can be avoided, however, by following the rule proposed by Siegel & Turkey [10] for their closely related test: namely, always discard the middle observation if (n_X+n_Y) is odd. Then in Chart B the median of the combined samples is m = 35, with k(m) = 5, S(m) = 4 + 0 = 4, and

$$W = 4 - (5)\ (4) + (9)\ (10) / 2 = 29.$$

Both of these results for W check with the defining formula.

If there are ties in the data, Ansari & Bradley recommend computing W by assigning to each X within any group of ties the average of the different scores which the group would receive if the ties were somehow resolved. The alternate formula, as presented above, will produce the same result unless there is a between-sample tie at the median of the combined samples: that is, unless there are both X's and Y's equal to m. In that case I can propose no exact correction to the alternate formula, but it is generally only slightly in error. In Chart C, for example, the median of the combined samples is m = 3, and there are both X's and Y's equal to this value. We have already calculated k(3) = 4.5 and S(3) = 14.75, so the alternate formula gives

$$W = 14.75 - (4.5)\ (3.5) + (8)\ (9) / 2 = 35.$$

This compares with W = 34.75 as defined by Ansari & Bradley.

7. TESTS BASED ON TRIPLETS

Define

$$N_{XYX} = \sum_{i_1<i_2} \sum_{j} \phi(X_{i_1}, X_{i_2}, Y_j)$$

and

$$N_{YXY} = \sum_{i} \sum_{j_1<j_2} \phi(Y_{j_1}, Y_{j_2}, X_i)\ ,$$

where

$$\phi(a,b,z) = \begin{cases} 1 & \text{if } a < z < b \text{ or } b < z < a \\ 1/2 & \text{if } z = a \neq b \text{ or } a \neq b = z \\ 1/3 & \text{if } a = b = z \\ 0 & \text{otherwise} \end{cases}$$

If there are no ties, then N_{XYX} (or, N_{YXY}) is the number of ways in which it is possible to choose from the data 2 X's and 1 Y (or, 2 Y's and 1 X) such that the Y lies between the X's (or, the X lies between the Y's). If ties are present, they lead to fractional counts as indicated in the definition of ϕ. These fractions have been determined so that the expected value of $\phi(Z_1,Z_2,Z_3)$ is $1/3$ if Z_1, Z_2, and Z_3 are independent observations from any distribution, continuous or not.

The quantities N_{XYX} and N_{YXY} are easily calculated with the aid of the pair chart. In the j-th row of the rectangle, let L_j be the number of squares to the left of the path, and R_j the number to the right of it; if some squares, say Q_j of them, lie within a box, count these as equally divided between L_j and R_j.

Then

$$N_{XYX} = \sum_{j=1}^{n_Y} L_j R_j - \sum_{j=1}^{n_Y} Q_j(Q_j+2) / 12 .$$

Similarly,

$$N_{YXY} = \sum_{i=1}^{n_X} B_i A_i - \sum_{i=1}^{n_X} H_i(H_i+2) / 12 ,$$

where B_i, A_i, and H_i are the numbers of squares below the path, above it, and boxed, respectively, within the i-th column. The second term in each formula is a correction for ties and can of course be ignored in untied data. Thus in Chart A, with no ties, we have

$$N_{XYX} = (2)(8) + (8)(2) + (8)(2) + (8)(2) + (8)(2) + (9)(1) + (9)(1) + (10)(0)$$
$$= 48,$$

and

$$N_{YXY} = (0)(8) + (0)(8) + (1)(7) + (1)(7) + (1)(7) + (1)(7) + (1)(7) + (1)(7) + (5)(3) + (7)(1) = 64.$$

In Chart B, where again there are no ties, we have

$$N_{XYX} = (1)(9) + (5)(5) + (5)(5) + (6)(4) + (6)(4) + (6)(4) + (6)(4) = 155,$$

and

$$N_{YXY} = (0)(7) + (1)(6) + (1)(6) + (1)(6) + (1)(6) + (3)(4) + (7)(0) + (7)(0) + (7)(0) + (7)(0) = 36.$$

In Chart C, including the correction term for ties, we have

$$\begin{aligned} N_{XYX} &= (.5)(7.5) + (2.5)(5.5) + (2.5)(5.5) + (4.5)(3.5) + (4.5)(3.5) \\ &\quad (4.5)(3.5) + (7)(1) + (8)(0) \\ &\quad - \{(1)(3) + (3)(5) + (3)(5) + (1)(3) + (1)(3) + (1)(3) + (0)(2) \\ &\quad + (0)(2)\} / 12 \\ &= 85.5 - 42/12 \\ &= 82, \end{aligned}$$

and

$$\begin{aligned} N_{YXY} &= (.5)(7.5) + (2)(6) + (2)(6) + (2)(6) + (4.5)(3.5) + (6)(2) \\ &\quad + (6)(2) + (7)(1) \\ &\quad - \{(1)(3) + (2)(4) + (2)(4) + (2)(4) + (3)(5) + (0)(2) + (0)(2) \\ &\quad + (0)(2)\} / 12 \\ &= 86.5 - 42/12 \\ &= 83. \end{aligned}$$

The quantities defined above are related to several tests of $H_0 : F_X \equiv F_Y$. In particular, it seems intuitively reasonable that N_{XYX} will tend to be large and N_{YXY} small if the X's and Y's do not differ greatly in location but the X's are more variable; or, N_{XYX} small and N_{YXY} large if the X's are less variable. This suggests basing a test against suspected differences in scale on such a statistic as $(N_{XYX}-N_{YXY})$. And, in fact, the well-known squared-rank statistic of Mood [8], usually written as

$$M = \sum_{i=1}^{n_X} (R_i - \frac{n_X+n_Y+1}{2})^2 ,$$

where R_i is the rank of X_i as in previous sections, was shown by Crouse & Steffens [2] to be expressible alternatively as

$$M = N_{XYX} - N_{YXY} + \frac{1}{12} n_X(n_X^2+3n_Y^2-1) .$$

These same authors proposed a modified test based on the statistic

$$M^* = 2 (N_{XYYX}-N_{YXXY}) ,$$

where notations such as N_{XYYX} indicate the number of ways in which it is possible to choose from the data 2 X's and 2 Y's such that after ordering they will have the indicated pattern. Then Crouse & Steffens show that

$$M^* = (n_Y-1) N_{XYX} - (n_X-1) N_{YXY} .$$

Finally, Lehmann[6] proposed a test, consistent against all alternatives, based on the statistic

$$L = (N_{XXYY}+N_{YYXX}) / N,$$

where $N = n_X(n_X-1) n_Y (n_Y-1) / 4$ is the total number of ways of choosing 2 X's and 2 Y's from the data. This was shown by Sundrum [11] to be expressible as

$$L = 1 - \{(n_Y-1) N_{XYX} + (n_X-1) N_{YXY}\} / 2N;$$

hence L also is easily computed from the pair chart.

For charts A, B, and C, the reader may verify the following:

	A	B	C
M	276.5	324	169
M^*	110	606	-7
L	629/1260	318/945	206.5/784

REFERENCES

[1] A.R. Ansari & R.A. Bradley, "Rank sum tests for dispersion", Annals of Mathematical Statistics 31 (1960), 1174-1189.

[2] C.F. Crouse & F.E. Steffens, "A distribution-free two sample test for dispersion for symmetrical distributions", South African Statistical Journal 3 (1969), 55-67.

[3] E.F. Drion, "Some distribution-free tests for the difference between two empirical cumulative distribution functions", Annals of Mathematical Statistics 23 (1952), 563-574.

[4] J.L. Hodges, Jr., "The significance probability of the Smirnov two-sample test", Arkiv för Matematik 3 (1958), 469-486.

[5] J.H. Klotz, "The Wilcoxon, ties, and the computer", Journal of the American Statistical Association 61 (1966), 772-787.

[6] E.L. Lehmann, "Consistency and unbiasedness of certain nonparametric tests", Annals of Mathematical Statistics 22 (1951), 165-179.

[7] H.S. Mann & D.R. Whitney, "On a test whether one of two random variable is stochastically larger than the other", Annals of Mathematical Statistics 18 (1947), 50-60.

[8] A.M. Mood, "On the asymptotic efficiency of certain nonparametric two-sample tests", Annals of Mathematical Statistics 25 (1954),514-522.

[9] Sidney Siegel, Nonparametric Statistics for the Behavioral Sciences. McGraw-Hill (1956).

[10] S. Siegel & J.W. Tukey, "A nonparametric sum of ranks procedure for relative spread in unpaired samples", Journal of the American Statistical Association 55 (1960), 429-445.

[11] R.M. Sundrum, "On Lehmann's two-sample test", Annals of Mathematical Statistics 25 (1954), 139-145.

[12] B.V. Sukhatme, "On certain two sample nonparametric tests for variances", Annals of Mathematical Statistics 28 (1957), 188-194.

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

[14] F. Wilcoxon, "Individual comparisons by ranking methods", Biometrics 1 (1945), 80-83.

[15] M.B. Wilk & R. Gnanadesikan, "Probability plotting methods for the analysis of data", Biometrika 55 (1968), 1-17.