A CORNER TEST FOR ASSOCIATION

By Paul S. Olmstead and John W. Tukey

Bell Telephone Laboratories and Princeton University

1. Summary. This paper proposes a new test (the "quadrant sum") for the association of two continuous variables. Its notable properties are:

(1) Special weight is given to extreme values of the variables.

(2) Computation is very easy.

(3) The test is non-parametric.

Significance levels (for the quadrant sum) are given to the accuracy needed for practical use. To this accuracy they are independent of sample size (see Fig. 1). The generating function of the quadrant sum is given for the null hypothesis (no association = independence). A limiting distribution is deduced and compared with the cases $2n = 4, 6, 8, 10$, and 14. Extension to higher dimensions and application to serial correlation are discussed.

2. Description of test (even number in sample). We shall describe the test as though a scatter diagram had already been drawn. The possibilities of direct computation from tabular data are indicated by the examples in sections 8 and 9.

In the scatter diagram, draw the two lines, $x = x_m$, $y = y_m$, where x_m is the median of the x-values without regard to the values of y, and y_m is the median of the y-values without regard to the values of x. Think of the four quadrants or corners thus formed as being labelled $+, -, +, -$, in order, so that the upper right and lower left quadrants are positive. Beginning at the right hand side of the diagram, count in (in order of abscissae) along the observations until forced to cross the horizontal median. Write down the number of observations met before this crossing, attaching the sign $+$ if they lay in the $+$ quadrant, and the sign $-$ if they lay in the $-$ quadrant. Repeat this process moving up from below, moving to the right from the left, and moving down from above. The quadrant sum is the algebraic sum of the four terms thus written down. This process is illustrated in Fig. 2, where the black dots represent contributions to the sum, and the dotted lines, crossings.

When there are an even number of pairs (x, y) and no ties, the medians will pass between the points. In this, the simplest case, the distribution of the quadrant sum is known for the hypothesis of no association (that is, of independence), and significance levels are given in Table 1 for the magnitude (absolute value) of the sum. It will be noticed that the sample size does not enter in any important way.

The cases of an odd number of observations and of ties are discussed in the next two sections. Simple devices make the test usable in most cases. A very great tendency toward ties, however, will make it inapplicable. This will be

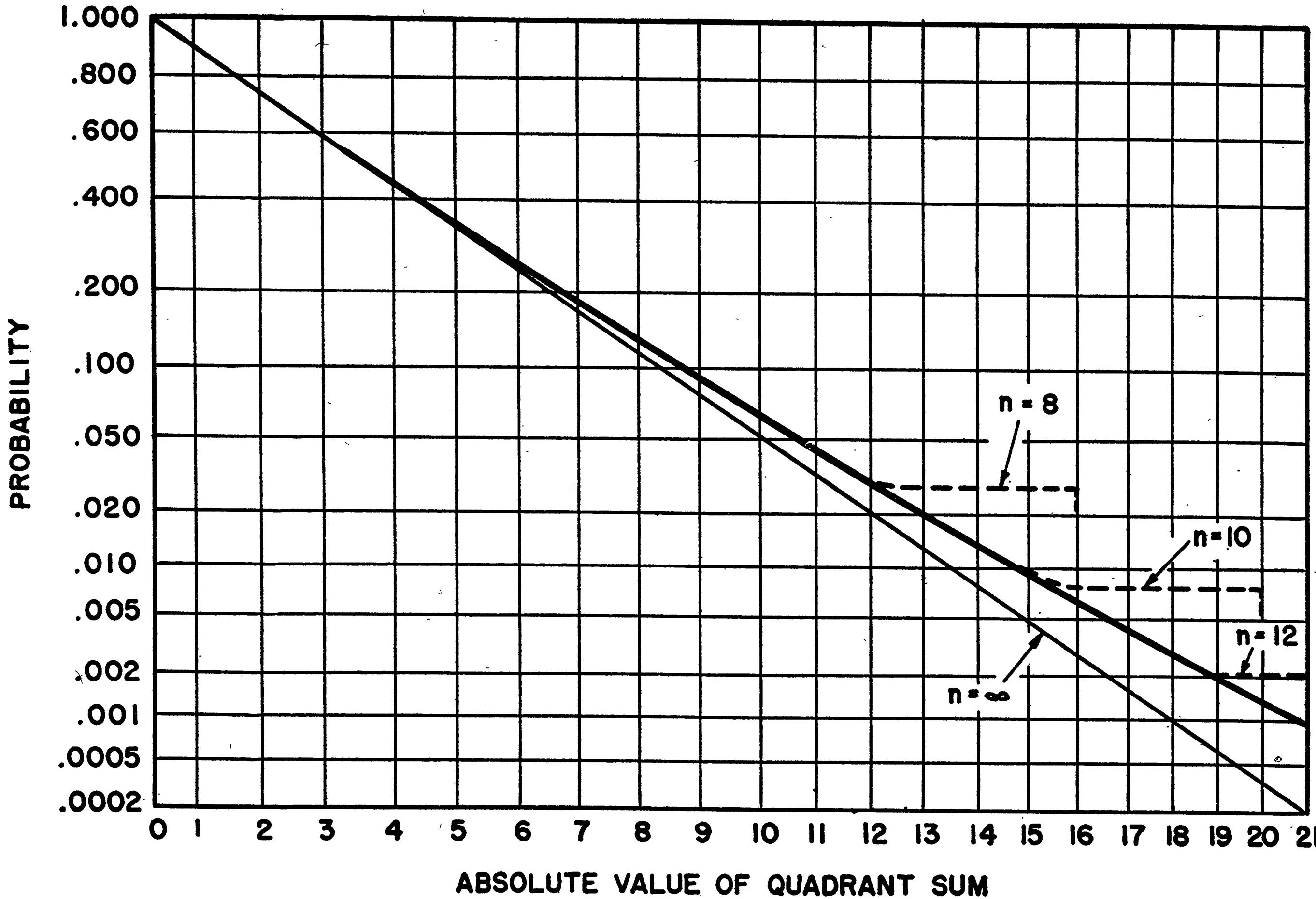

FIG. 1. Probability that an arrangement of a sample of size, N, will have a quadrant sum equal to or greater than indicated absolute value.

unimportant in most applications because of the fact that attention is being directed to the periphery.

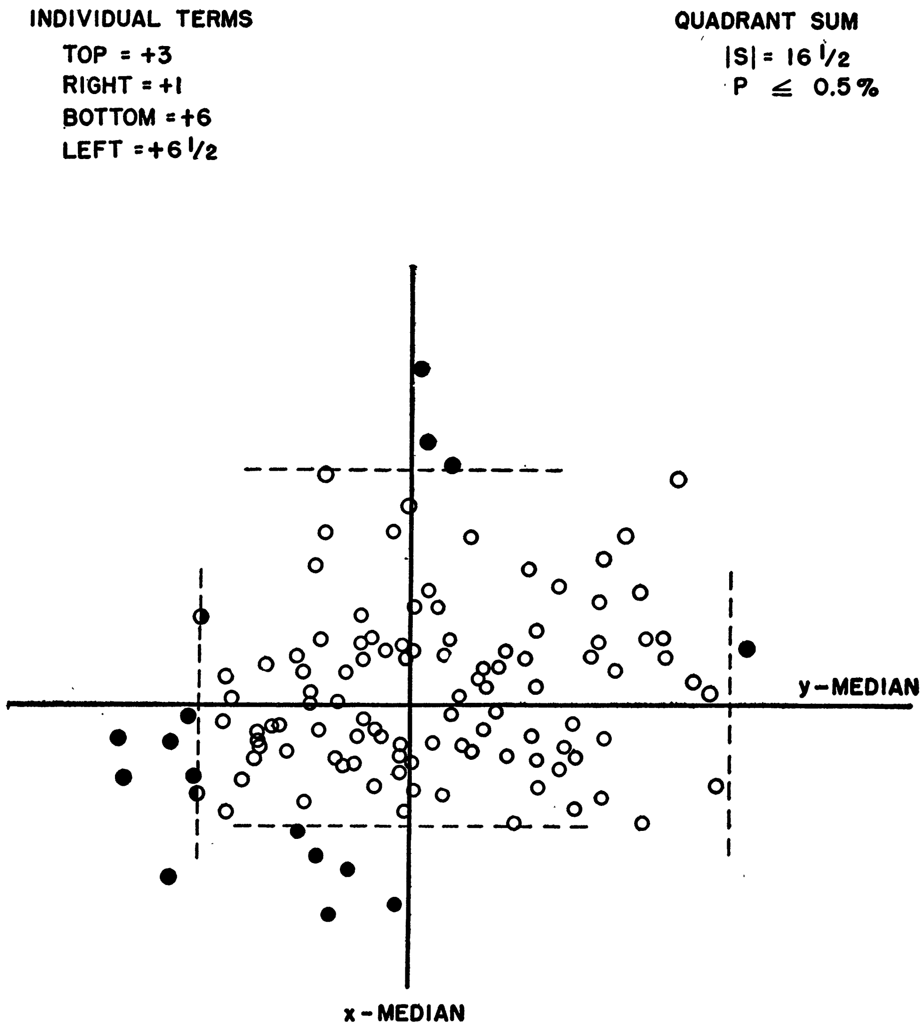

FIG. 2. Scatter diagram of 116 pairs of observations

The set of data which prompted the development of the test is shown in Fig. 2. The accompanying report described it as follows: "The various points appear to be scattered almost completely at random and give little indication of correlation." The quadrant sum is $16\frac{1}{2}$ which is significant at the 0.5% point. Intuitively, the significant association of the peripheral points is clear.

3. Description of test (odd number in sample). If the sample size is odd, then we may usually follow the process outlined above. We will have difficulty only when the counting process meets a point, one of whose coordinates is a median. In this case we employ a simple device, namely:

Given a sample of $2n + 1$ pairs, let x^* and y^* be the medians of the x-values and of the y-values, respectively. Let the pairs in which they occur be (x^*, y_k) and (x_m, y^*), respectively. Replace these two pairs by the single pair (x_m, y_k). There are now $2n$ pairs and the regular method can be applied.

The quadrant sum so obtained from an unassociated population has the same distribution as that formed directly from $2n$ pairs.

4. Description of test (treatment of ties). The behavior of the test is known when (1) there is no association, (2) the probability of a tie in x-values or y-values

TABLE 1

Working significance levels for magnitudes of quadrant sums

Significance level (Conservative)	Magnitude of quadrant sum*
10%	9
5%	11
2%	13
1%	14–15
0.5%	15–17
0.2%	17–19
0.1%	18–21

* The smaller magnitude applies for large sample size, the larger magnitude for small sample size. Magnitudes equal to or greater than twice the sample size less six should not be used.

is zero. The following approximation, which has an unknown effect on the distribution, is suggested when ties are present:

When a tied group is reached, count the number in the tied group favorable to continuing and the number unfavorable. Treat the tied group as if the number of its points preceding the crossing of the median were

$$\frac{\text{number favorable}}{1 + \text{number unfavorable}}.$$

It seems likely that this approximation is conservative.

5. Discussion. When a moderate number, say 25 to 200, of paired observations on two quantities are plotted as a scatter diagram, visual examination frequently detects what seems to be definite evidence of association between the variables. Often in such cases, the usual methods for measuring association do not find statistical significance of association. Visual judgment, particularly by engineers or scientists who may wish to take action on the basis of their findings, gives greater weight to observations near the periphery of the scatter diagram. This is not always desirable—but often it is very desirable. A quantitative test of association with such concentration on the periphery has been lacking. The quadrant sum test was developed to fill the gap. Its features of speed and non-parametricity are useful but secondary from this point of view.

When uniform attention to the whole scatter diagram is desired, the quadrant sum test is of unknown usefulness. We know little enough of the operating characteristics of the more conventional tests, such as:

1. The product moment correlation coefficient
2. The four-fold table formed by the medians
3. The biserial correlation coefficient
4. The rank correlation coefficient

and less about the operating characteristics of the present test. In this case, the quadrant sum test can only be recommended definitely for exploratory investigations of large amounts of data.

There are many situations, however, where we do not know where to concentrate our attention, and where speed and non-parametricity are cardinal virtues in a test. One example is the use of serial correlation in studying industrial processes. We may guess that here we are interested in the periphery, but neither theory nor experience can, so far, prove this. In such situations the quadrant sum is by far the fastest to use of any of the tests known to the authors, and we believe one of the most useful.

6. Elementary derivations. We can easily find the distribution of

1. An individual term of the quadrant sum
   a. For fixed sample size
   b. In the limit
2. The quadrant sum itself
   a. For fixed sample size
   b. In the limit, assuming asymptotic independence of the four terms.

This we shall do now, leaving the proof that 2a actually converges to 2b to a later section.

Consider a sample of $2n$ pairs $(x_1, y_1), \cdots, (x_{2n}, y_{2n})$ from a population in which x and y are independent. It is both clear and easily verifiable that

1. The set of $2n$ x-values, $x_1, \cdots, x_{2n}$
2. The set of $2n$ y-values, $y_1, \cdots, y_{2n}$
3. The permutation of the order of the y-values when the pairs are ordered by the x-values

which together determine the sample, are independently distributed, and that any permutation is as likely as every other. (We have assumed no ties, which is a consequence, with probability one, of the continuous cumulative distributions of x and y). Since the quadrant sum depends only on the permutation, its distribution in the absence of association does not depend on the distributions of x and y.

We must solve, then, certain purely combinatorial problems—under the hypothesis that the $2n!$ permutations of the y-values are all equally likely. It may simplify matters to assume that the values of x in the sample are $1, 2, \cdots, 2n$ and that those of y are the same. How, then, do we calculate the distribution of a single term of the quadrant sum. Let us begin with small x-values, and the pair $(1, y_1)$. If $y_1 = 1, 2, \cdots, n$, we count "one" positive, and if $y_1 = n + 1, n + 2, \cdots, 2n$, we count "one" negative. We pass on to $(2, y_2)$ and so on. How many permutations yield a count of exactly k positive values? Those in which $y_1, y_2, \cdots, y_k$ are equal to or less than n, y_{k+1} equal to or greater than $n + 1$, and the other $(2n - k - 1)y$'s are arbitrary. There are:

$$n(n - 1) \cdots (n - k + 1) \cdot (n)(2n - k - 1)!$$

such permutations, the fraction of all $(2n)!$ permutations being:

$$\frac{n(n - 1) \cdots (n - k + 1)n}{(2n)(2n - 1) \cdots (2n - k + 1)(2n - k)} \tag{1}$$

which is, then, the probability that this contribution will equal $+k$, or by symmetry, the probability that it will equal $-k$, $k \neq 0$.

For large n, this becomes merely:

$$p_k = 2^{-(|k|+1)}, \qquad k \neq 0. \tag{2}$$

In order to obtain the distribution of the quadrant sum itself, we must concern ourselves with the lack of independence of the four terms. This is indicated most clearly in the case of $2n = 2$, where the $2! = 2$ permutations yield $+1 +1 +1 +1 = 4$ and $-1 -1 -1 -1 = -4$. Here, there is complete lack of independence. We shall see later that there is effectively independence in the limit, so that it is worth while to calculate the sum of four independent terms with the limiting distribution (2) and find that it satisfies:

$$Pr(|\text{ independent sum of 4 terms }| \geq k) = \frac{9k^3 + 9k^2 + 168k + 208}{216 \cdot 2^k}, k > 0. \tag{3}$$

The details will be omitted.

A simple device, reminiscent of Wald's [3, 1943] establishment of the two-dimensional tolerance limits enables us to avoid difficulties with lack of independence and compute the exact distribution of the quadrant sum for any n. We decompose the permutation of the y-values into the following parts, which together specify the permutation:

(a) The number, j, of pairs in the upper right quadrant.
(b) The set of j values of x between $n + 1$ and $2n$ corresponding to pairs in the upper right quadrant.

(c) The set of j values of y between $n + 1$ and $2n$ corresponding to points in the upper right quadrant.
(d) The set of j values of x between 1 and n corresponding to pairs in the lower left quadrant. (Note that the use of medians ensures that the lower left and upper right quadrants contain the same number of points.)
(e) The set of j values of y between 1 and n corresponding to pairs in the lower left quadrant.
(f) The permutation of j objects defined by the pairs in the upper right quadrant.
(g) The permutation of $n - j$ objects defined by the pairs in the upper left quadrant.
(h) and (i) the permutations from the remaining quadrants.

It is easily verified that: (1) given j, items (b) to (i) can be assigned at will, (2) each assignment of (a) to (i) corresponds to one and only one permutation, (3) the quadrant sum depends only on items (b) to (e). In fact, the right hand term depends on item (b), the upper term on item (c), the left hand term on item (d) and the lower term on item (e). While j remains fixed, the terms behave independently.

For fixed j, what is the distribution of a single term? If a set of j x-values gives the term $+k$, it must contain the k largest x-values and not contain the next. There are:

$$\binom{n-k-1}{n-j-1}$$

such sets. The generating function for a single term, is, then:

$$\sum_{k=1}^{j}\binom{n-k-1}{n-j-1}x^{k}+\sum_{k=1}^{n-j}\binom{n-k-1}{j-1}x^{-k}. \tag{4}$$

Since the terms are independent for fixed j, and there are $(j!)^2((n-j)!)^2$ ways to supply the permutations forming items (f) to (i), the generating function for the quadrant sum, S_n, is:

$$G_n(x)=\sum_{j=0}^{n}\frac{(j!)^2((n-j)!)^2}{(2n)!}\left[\sum_{k=1}^{j}\binom{n-k-1}{n-j-1}x^{k}+\sum_{k=1}^{n-j}\binom{n-k-1}{j-1}x^{-k}\right]^4. \tag{5}$$

The exact probability of equalling or exceeding each value of S_n has been computed for $2n = 2, 4, 6, 8, 10$, and 14. Table 2 gives these probabilities and Fig. 3 shows the values of

$$\frac{m}{5}+\log_{10} Pr(\,|\text{quadrant sum}|\geq m)$$

this particular function being chosen for its relative constancy. The maximum value of the quadrant sum is $4n$, and for values of k less than $4n - 6$, there

TABLE 2

Probability of a Sum of Absolute Value Equal to or Greater than k when a Sample of 2n is Drawn from an Unassociated Population

$\|k\|$ \ $2n$	2	4	6	8	10	12	14	∞*
0	1.0000	1.0000	1.0000	1.0000	1.0000		1.0000	1.000000
1	1.0000	0.7500	0.9333	0.9036	0.9106		0.9115	0.912037
2	1.0000	0.7500	0.7556	0.7544	0.7567		0.7580	0.754630
3	1.0000	0.4167	0.6000	0.6000	0.6008		0.6039	0.599537
4	1.0000	0.4167	0.4667	0.4619	0.4662		0.4690	0.462963
5	0.0000	0.3333	0.3111	0.3508	0.3519		0.3547	0.346933
6	0.0000	0.3333	0.2222	0.2619	0.2589		0.2611	0.252025
7	0.0000	0.3333	0.1556	0.1821	0.1867		0.1876	0.177662
8	0.0000	0.3333	0.1111	0.1258	0.1333		0.1322	0.121817
9	0.0000	0.0000	0.1000	0.0839	0.0928		0.0918	0.081471
10	0.0000	0.0000	0.1000	0.0554	0.0642		0.0632	0.053295
11	0.0000	0.0000	0.1000	0.0375	0.0436		0.0432	0.034189
12	0.0000	0.0000	0.1000	0.0304	0.0290		0.0296	0.021557
13	0.0000	0.0000	0.0000	0.0286	0.0190		0.0202	0.013386
14	0.0000	0.0000	0.0000	0.0286	0.0127		0.0139	0.008200
15	0.0000	0.0000	0.0000	0.0286	0.0095		0.0096	0.004963
16	0.0000	0.0000	0.0000	0.0286	0.0083		0.0066	0.002972
17	0.0000	0.0000	0.0000	0.0000	0.0079		0.0045	0.001762
18	0.0000	0.0000	0.0000	0.0000	0.0079		0.0031	0.001036
19	0.0000	0.0000	0.0000	0.0000	0.0079		0.0021	0.000604
20	0.0000	0.0000	0.0000	0.0000	0.0079		0.0014	0.000350
21	0.0000	0.0000	0.0000	0.0000	0.0000		0.0010	0.000201
22	0.0000	0.0000	0.0000	0.0000	0.0000		0.0008	0.000115
23	0.0000	0.0000	0.0000	0.0000	0.0000		0.0006	0.000065
24	0.0000	0.0000	0.0000	0.0000	0.0000		0.0006	0.000036
25	0.0000	0.0000	0.0000	0.0000	0.0000		0.0006	0.000020
26	0.0000	0.0000	0.0000	0.0000	0.0000		0.0006	0.000011
27	0.0000	0.0000	0.0000	0.0000	0.0000		0.0006	0.000006
28	0.0000	0.0000	0.0000	0.0000	0.0000		0.0006	0.000003
29	0.0000	0.0000	0.0000	0.0000	0.0000		0.0000	0.000002
30	0.0000	0.0000	0.0000	0.0000	0.0000		0.0000	0.000001
31 or over	0.0000	0.0000	0.0000	0.0000	0.0000		0.0000	0.000000
Variance of k	16	24	$26\frac{2}{5}$	$26\frac{6}{7}$	$26\frac{16}{21}$	$26\frac{6}{11}$	$26\frac{140}{429}$	24

* Probability for $2n = \infty$, $k > 0$, is given by

$$\frac{9k^3 + 9k^2 + 168k + 208}{216 \cdot 2^k}.$$

is quite good agreement between the curves for finite n and formula (3) at the practically significant percentage points. The situation for very small

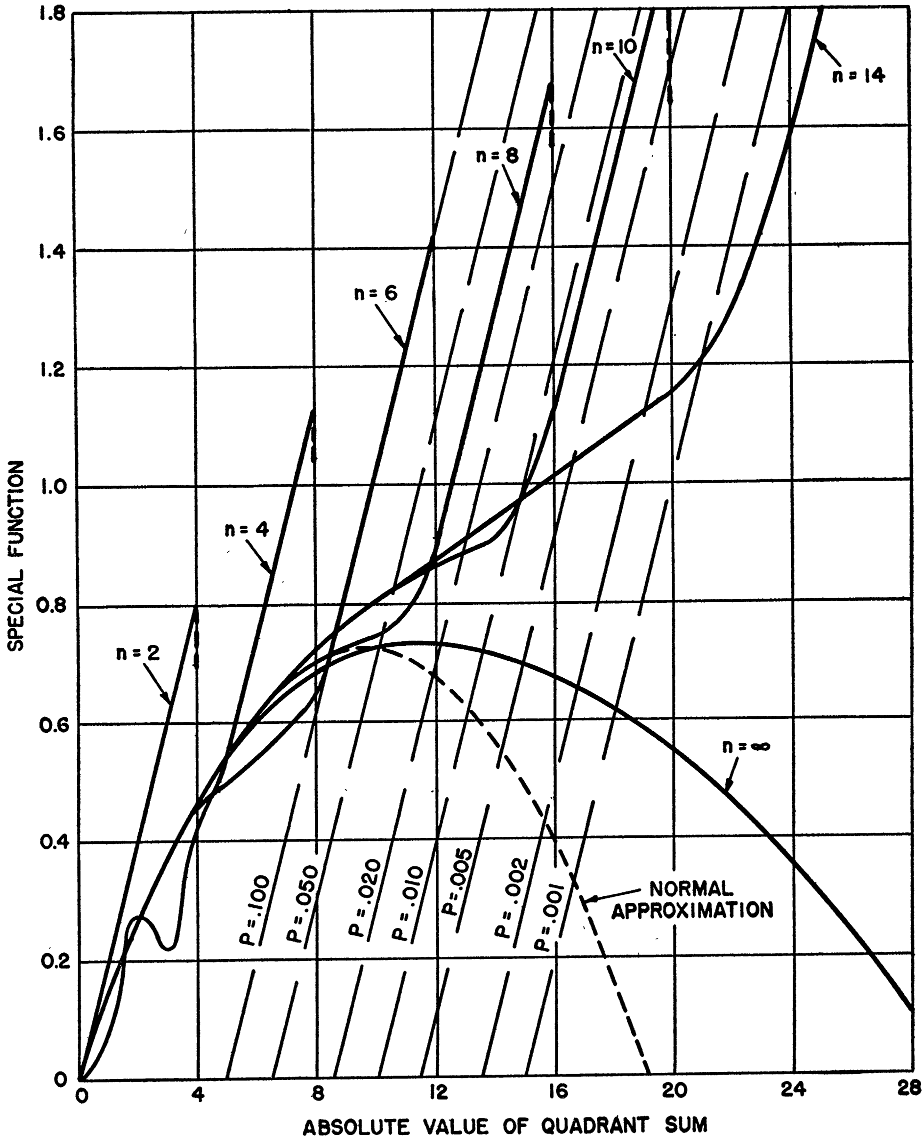

FIG. 3. Comparative relationships for finite and infinite sample sizes and normal approximation to the infinite sample size

probabilities suggests a careful consideration of the limiting behavior of the quadrant sum distribution (see section 10).

The device for samples of $2n + 1$ deserves a word of justification. If there is no association, the $2n + 1$ y-values are randomly paired with the $2n + 1$ x-values, and, in particular, the y-value paired with the x-median is randomly selected. If we pair it with the (randomly selected) x-value which was paired with the y-median we still have random pairing. The pairing of the $2n$ pairs is random, although neither the x-values nor the y-values make up a sample. The randomness of pairing is all that has been used in the discussions of this section.

7. Extension to higher dimensions. The same ideas that underlie the quadrant sum test for two variables may be extended *in several ways* to give tests for various types of association among three or more variables. Only one three-variable case will be discussed here, leaving further extension to the reader.

Given three variables, x, y, and z, and a sample of matched observations on these, it is clearly possible to use the simple quadrant sum test for two variables to investigate association between x and y separately, between y and z separately, and between z and x separately. If the Pearson coefficient of correlation were being computed and were found to be close to zero for each of these pairs, it would be assumed that there was no detectable association through the second moments. In a trivariate normal or Gaussian distribution, where the first and second moments determine the whole distribution, if there is independence between the separate pairs of variables, there is no possibility of a three-way association. It is of some interest, however, to notice that a corner sum test can be devised that will measure the effect of such triple association in case it does exist.

Consider the octants into which the three median planes for x, y, and z, respectively, divide the three dimensional scatter diagram and label the octants alternately plus and minus, in the manner suggested by Fig. 4. More precisely, an octant is counted as plus if an odd number, that is three or one, of the variables are greater than the medians of the sample, and the remaining octants are labelled minus. It is clear that we may repeat the process of coming in along each axis passing from observation to observation as long as they remain in a region of fixed sign, and writing down as a contribution to the final or octant sum the number of such consecutive elements and the sign of the region in which they were found. There will be six terms rather than four, as was the case for the test based on quadrants, and so a new set of significance levels will be required. Table 3, following, lists the situation for a very large sample.

The situation has been sketched for the case of $2n$ triples. If there are $2n + 1$ triples, then we may have trouble with the medians again. However, a similar device works, except that we must agree on a last variable in order to form the synthetic triples uniquely. For example, consider the triples $(m, 3, 5)$, $(9, m, 1)$, $(12, 4, m)$, where m denotes the median. Taking the order in which the variables are written, we get $(12, 3, 5)$ and $(9, 4, 1)$ as the synthetic triples. Other

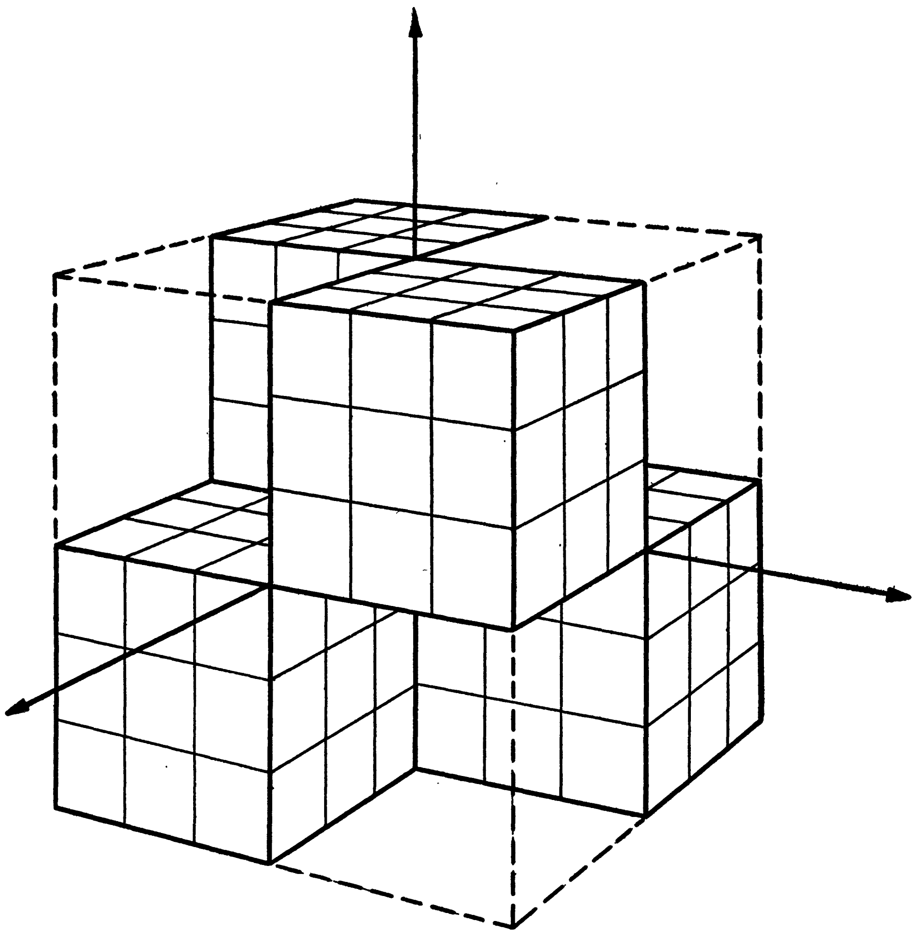

FIG. 4. Octant schematic—solid sections taken as positive

TABLE 3

Working significance levels for the magnitudes of the octant sum

Significance Level	Magnitude of Octant Sum*
10%	11
5%	13
2%	15
1%	16
0.5%	18
0.2%	20
0.1%	21

* Computed for large samples only and based on normal approximation, see section 11 for discussion of this and higher dimensional cases.

orders would yield (9, 3, 5) and (12, 4, 1) or (9, 3, 1) and (12, 4, 5). This slight dissymmetry is not pleasing but should give no difficulty.

8. Nongraphical example. The following example of 78 successive observations of four variables shows how this test may be applied without plotting and how simple the computation still remains. The data concern a metallurgical

TABLE 4

Excerpt from Tippett's Table

Time T*	Fuel F*	Material M*	Articles A*	Duration D*
1 −	246 +	1457 −	1895 +	168.5 +
2 −	196 −	2078 +	2121 +	152 +
3 −	192 −	1278 −	1437 −	153 +
4 −	202 +	1398 −	1497 −	145 −
5 −	206 +	1944 +	1592 +	153 +
6 −	218 +	1464 −	1506 −	147.5 −
7 −	155 −	1541 +	1762 +	152 +
8 −	201 +	1502 +	1818 +	144.5 −
9 −	211 +	1950 +	1144 −	151.5 +
10 −	236 +	1768 +	1654 +	151.5 +
etc. to				
78 +	185 −	1536 +	1442 −	152 +
Median 39.5	Median 199	Median 1474	Median 1588	Median 149.5

* Location of observation relative to column median; + = above; − = below.

Tippett's correlations (based on lightly rounded data)

$$r_{FM} = +0.243$$
$$r_{FA} = +0.266$$
$$r_{MA} = +0.681$$
$$r_{FM.A} = +0.088$$
$$r_{FMA.} = +0.141.$$

problem in mass production and are taken from L. H. C. Tippett, Table XXII, page 63 [2]. An excerpt from the data is given in Table 4 together with Tippett's calculated correlations. This table also shows the preliminary marking of each individual measurement as above (+) for its variable, below (−), or on the median (0). From this table we see, for example, that increasing T contributes a term -3 to the quadrant sum for T and D. It is often desirable to prepare auxiliary tables to assist in computing the components of the quadrant and hyperquadrant sums. Such a table is Table 5 for low values of Fuel ($F-$) arranged in consecutive ascending numerical order. The entries on this table for the five columns headed F, T, M, A, and D are directly comparable to the entries in Table 4. For example, $F = 155$ is $-$ with respect to the fuel median and $T = 7$, $-$; $M = 1541$, $+$; $A = 1762$, $+$; $D = 152$, $+$. The double, triple, quadruple and quintuple headed columns contain simply the algebraic multiplication of the signs in the appropriate T, M, A, or D columns. Thus, TM for $F = 155$ is $-$, MAD is $+$, and $TMAD$ is $-$. The contribution to each quadrant or hyperquadrant sum is simply the count of the consecutive like signs from the top of a column. For column AD, we have 7 consecutive $+$ signs and since the contribution is to FAD and F is $-$, the contribution in this case to the octant sum is -7. The results from the ten tables of which Table 5

TABLE 5

Sample Table for One Component of Quadrant and Hyperquadrant Sums. Low Values of Fuel (F−)

Fuel *F*	*T*	*M*	*A*	*D*	*TM*	*TA*	*TD*	*MA*	*MD*	*AD*	*TMA*	*TMD*	*TAD*	*MAD*	*TMAD*
98 −	+	−	−	−	−	−	−	+	+	+	+	+	+	−	−
135 −	+	−	−	−	−	−	−	+	+	+	+	+	+	−	−
140 −	−	−	−	−	+	+	+	+	+	+	−	−	−	−	+
146 −	−	−	−	−	+	+	+	+	+	+	−	−	−	−	+
147 −	+	+	−	−	+	−	−	−	−	+	−	−	+	+	+
149 −	−	+	−	−	−	+	+	−	−	+	+	+	−	+	−
151 −	+	−	−	−	−	−	−	+	+	+	+	+	+	−	−
153 −	+	−	+	−	−	+	−	−	+	−	−	+	−	+	+
155 −	−	+	+	+	−	−	−	+	+	+	−	−	−	+	−

Contributions to Sums

FT	*FM*	*FA*	*FD*	*FTM*	*FTA*	*FTD*	*FMA*	*FMD*	*FAD*	*FTMA*	*FTMD*	*FTAD*	*FMAD*	*FTMAD*
−2	+4	+7	+8	+2	+2	+2	−4	−4	−7	−2	−2	−2	+4	+2

is a sample are then carried to the summary computation shown in Table 6. The contribution from Table 5 is shown on line F−. The totals are computed and their probabilities of occurrence determined.

9. Serial example. The following example, a sample of 144 observations of the thickness of inlay for relay springs cut consecutively from a single sheet of material, allows us to compare the resolution of the present test with that of the serial product-moment correlation. The data are from Shewhart [1, 1941, Table 1] and the serial correlations from lag 1 to lag 22 are from recent calculations by Miss Dorothy T. Angell. The procedure for calculating the serial quadrant sums is similar to that for obtaining the sums for section 8. A table is prepared to show the observed consecutive order of the numerical values and each is identified as above (+), below (−), or on the median (0). This gives a

TABLE 6

Summary Computation Table for Quadrant and Hyperquadrant Sums

From Table	*TF*	*TM*	*TA*	*TD*	*FM*	*FA*	*FD*	*MA*	*MD*	*AD*	*TFM*	*TFA*	*TFD*	*TMA*	*TMD*	*TAD*	*FMA*	*FMD*	*FAD*	*MAD*	*TFMA*	*TFMD*	*TFAD*	*TMAD*	*FMAD*	*TFMAD*
T +	−6	+1	−2	+3							−1	+2	−3	−1	+1	−2					+1	−1	−1	+2		+1
T −	−1	+1	−2	−3							+2	−1	−1	+1	+1	−2					+3	+2	+1	−1		+4
F +	+4				+4	+1	+1				+5	+1	+1				+1	+1	+3		+1	+1	+3		+3	+3
F −	−2				+4	+7	+8				+2	+2	+2				−4	−4	−7		−2	−2	−2		+4	+2
M +		+1			+2			+5	+13		+1			+1	+1		+2	+2		+5	+1	+1		+1	+2	+1
M −		−2			+6			+25	+3		+2			+2	+2		−6	−3		−3	−2	−2		−2	+3	+2
A +			−2			−1		+3		+7		+1		−2		−2	−1		−1	+3	+1		+1	−2	−1	+1
A −			−1			+5		+1		+4		+1		+3		+1	−1		−4	−1	−3		−1	−3	+1	+3
D +				+1			+2		+1	+2			+1		+3	+1		+1	+2	−1		−2	−1	−2	+1	+2
D −				−1			+3		+1	+7			+1		+2	+1		−1	−3	+1		+2	+1	+2	+1	+3
Totals Quadrant Sums	−5	+1	−7	0	+16	+12	+14	+34	+18	+20																
Octant Sums											+11	+6	+1	+4	+10	−3	−9	−4	−10	+4						
Hexadecant Sums																					0	−1	+1	−5	+14	
Dotriacontant Sums																										+22
Probability (%) ≤	36	92	19	100	0.2	2	1	0.1	0.1	0.1	9	37	94	57	12	68	16	57	12	57	100	95	95	52	6	0.6
Significant at 5%					*	*	*	*	*	*																*
Significant at 1%					*		*	*	*	*																*
Significant at 0.2%					*			*	*	*																

table similar to one of the elements, say Fuel, in Table 4. Four computation tables similar to Table 5 are required, one for the equivalent of moving from the right, one from below, one from the left, and one from the top of a lag correlation scatter diagram. One table from each direction will take care of all lags. In the first, the marginal entries are the observed values listed in descending numerical order. Opposite these are recorded from the previous table the signs associated with observations for each lag with respect to each entry. The second table would record the signs relating to the lags from the observed values arranged in ascending order. The third table would record the signs relating to leads from the observed values arranged in ascending order and the fourth, the signs relating to leads from the observed values arranged in descending order. The sign of the contribution from each group is the algebraic product of the sign of the run and the sign of the marginal entries. The length of run is determined in the same way as in Table 5. Table 7 illustrates the procedure

TABLE 7

Relation of Lagged Observations to Median (+ = above, − = below) for Smallest Observations in Ascending Order

Thickness	Lag																									
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
2	−	−	+	−	−	+	−	+	−	+	−	+	+	+	+	+	+	+	+	+	+	+	+	+	−	−
3	−	−	+	−	+	−	−	−	−	+	−	−	+	−	+	−	+	−	+	+	+	+	+	+	+	+
8	−	−	−	+	−	−	−	−	+	+	−	+	+	−	+	+	−	−	+	+	−	−	−	−	−	−
10	−	+	−	−	−	−	+	+	−	+	+	−	+	+	−	−	+	+	−	−	−	−	−	−	+	−
13	−	−	−	−	−	+	+	+	−	+	+	+	+	(+)	(+)	(+)	(+)	(+)	(+)	(+)	(+)	(+)	(+)	(+)	(+)	(+)
17	−	−	−	+	−	−	+	−	+	−	+	−	+	+	+	+	+	+	+	+	+	+	+	+	+	−
17	−	+	−	+	+	+	+	+	+	−	−	+	+	−	−	−	−	−	−	+	−	−	+	−	+	+
18	−	−	−	−	+	+	−	+	+	−	+	+	−	−	+	+	−	−	−	−	−	−	+	−	+	−
*		+3	−2	+2	+1	−1	+3	−1	+2	−5	+3	−1	−7	−1	−3	−1	−2	−1	−3	−3	−2	−2	−2	−2	+1	+1

* Contribution to Serial Quadrant Sum.

of determining the contribution from lags associated with the observations arranged in ascending order.

Two serial quadrant sums may be computed—a circular serial quadrant sum or a noncircular serial quadrant sum. Circular items arise from considering that the beginning of the set of observations is a continuation of the end in the same way that this assumption is made in computing circular serial correlation coefficients. In Table 7, circular items are shown in parentheses and are omitted in calculating noncircular sums. In the particular table shown, the count of the run lengths was identical for both types of sum, but in other cases this may not be the case. Since the serial quadrant sum is relatively insensitive to sample size, the noncircular serial quadrant sum has for all practical purposes the same distribution as the circular quadrant sum. The correspondence in this case between the serial correlation coefficient for each lag up to 22 and the respective values of the two types of serial quadrant sums is shown in Fig. 5.

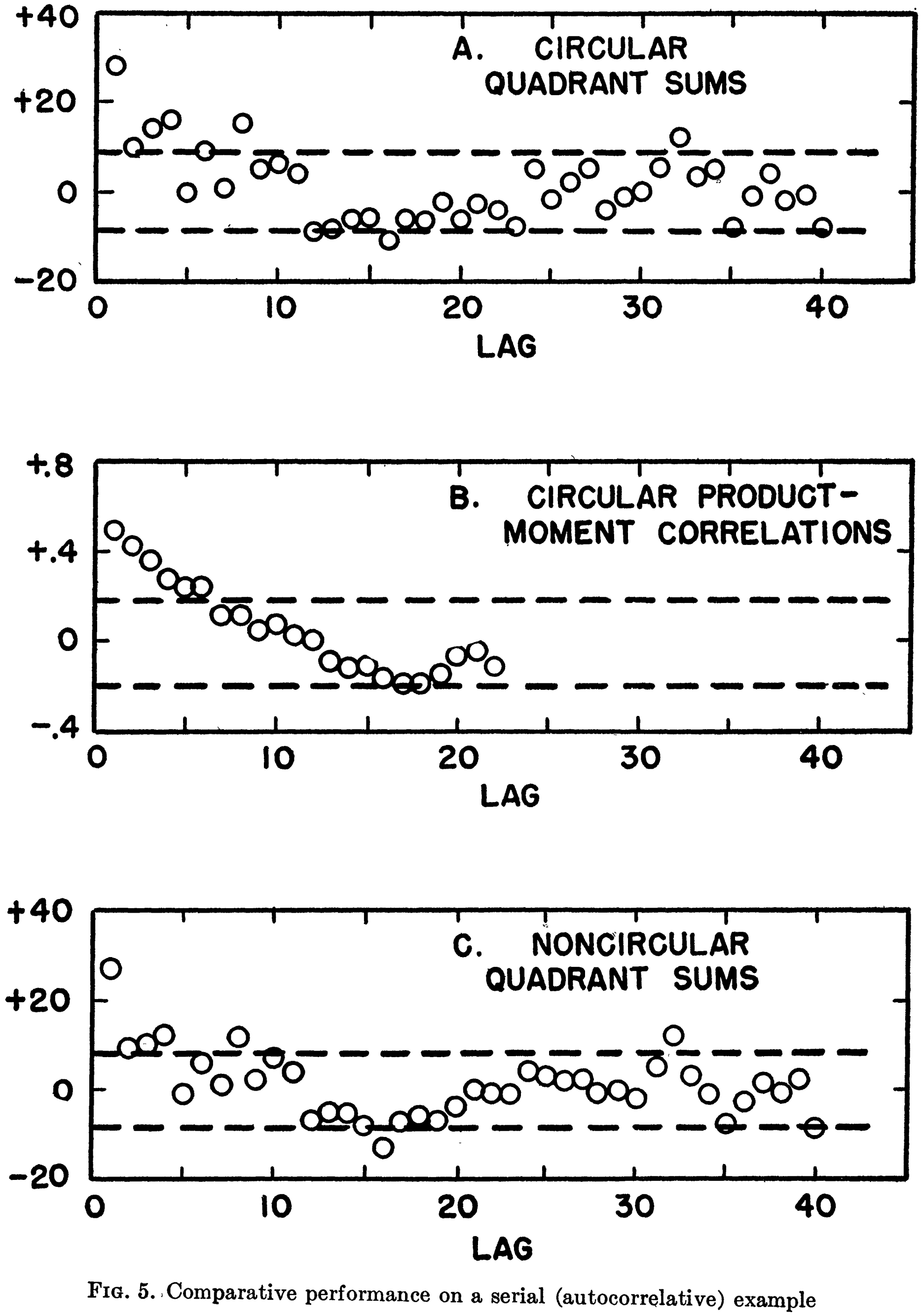

FIG. 5. Comparative performance on a serial (autocorrelative) example

10. Convergence to the limiting distribution. We shall consider several chance sums. One of these is S, which has the limiting distribution discussed in section 6. Another is S'_k, which is the sum of four independent terms, each distributed according to the limiting distribution curtailed at $\pm k$. Its generating function is

$$G_k(x) = \left(\sum_{i=1}^{k} 2^{-(i+1)} x^i + \sum_{i=1}^{k} 2^{-(i+1)} x^{-i}\right)^4.$$

The total probability assigned to $S'_k = -k, -(k-1), \cdots, k$, is less than unity, so that there is nonzero probability that S'_k is not defined. The third is S_n, the quadrant sum itself, whose generating function is (5), and the fourth is the result of the same sort of curtailment applied to S_n. It will be denoted by $S_{n,k}$ and its generating function is

$$G_{n,k}(x) = \sum_j \frac{(j!)^2((n-j)!)^2}{(2n)!}\left(\sum_{i=1}^{k}\binom{n-i-1}{n-j-1}x^i + \sum_{i=1}^{k}\binom{n-i-1}{j-1}x^{-i}\right)^4.$$

This again corresponds to a total probability less than unity.

It is clear that

$$\Pr(S_{n,k} = m) \leq \Pr(S_n = m)$$

and

$$\Pr(S'_k = m) \leq \Pr(S = m).$$

We shall soon show that

$$\lim_{n\to\infty} \Pr(S_{n,k} = m) = \Pr(S'_k = m) \tag{6}$$

and this will imply that

$$\lim_{n\to\infty} \Pr(S_n = m) = \Pr(S = m)$$

which is the desired result. The implication runs as follows: given ϵ, we can choose k so large that

$$\Pr(S'_k \text{ defined}] \geq 1 - \epsilon/3$$

whence

$$|\Pr(S'_k = m) - \Pr(S = m)| \leq \epsilon/3$$

and then choose n so large that

$$|\Pr(S_{n,k} = m) - \Pr(S_k = m)| \leq \epsilon/(24k + 6)$$

$$\text{for } m = -4k, -4k + 1, \cdots, 4k$$

whence

$$\Pr(S_{n,k} \text{ defined}) \geq 1 - \epsilon/3 - \frac{8k+1}{24k+6}\epsilon \leq 1 - \frac{16k+3}{24k+6}\epsilon$$

and hence

$$| \Pr(S_{n,k} = m) - \Pr(S_n = m) | \leq \frac{16k + 2}{24k + 6} \epsilon$$

this inequality holding automatically for $| m | > 4k$. Hence,

$$\begin{aligned} | \Pr(S_n = m) - \Pr(S = m) | \\ &\leq | \Pr(S_n = m) - \Pr(S_{n,k} = m) | + | \Pr(S_{n,k} = m) - \Pr(S'_k = m) | \\ &+ | \Pr(S'_k = m) - \Pr(S = m) | \leq \frac{16k + 2}{24k + 6} \epsilon + \frac{1}{24k + 6} \epsilon + \tfrac{1}{3}\epsilon < \epsilon \end{aligned}$$

This method is clearly of general application in such problems.

We turn now to the proof of (6). The expression for $G_{n,k}(x)$ shows that we may consider it the result of the following process: the integer j is a chance quantity with the distribution

$$\Pr(j = j_0) = \frac{(n!)^2}{(2n)!} \binom{n}{j_0}^2.$$

For fixed j, $G_{n,k}$ is the average over j of

$$G_{n,k,j}(x) = \left[\sum_{i=1}^{k} \frac{\binom{n - i - 1}{n - j - 1}}{\binom{n}{j}} x^i + \sum_{i=1}^{k} \frac{\binom{n - i - 1}{j - 1}}{\binom{n}{j}} x^{-i} \right]^4.$$

The first of these relations shows that j/n converges stochastically to $\frac{1}{2}$ as n approaches infinity. The second shows, since

$$\frac{\binom{n - i - 1}{n - j - 1}}{\binom{n}{j}} = \frac{(n - i - 1)!(n - j)!j!}{(n - j - 1)!(j - i)!n!} = \frac{(n - j)(j)(j - 1) \cdots (j - i + 1)}{n(n - 1)(n - 2) \cdots (n - i)},$$

$$\frac{\binom{n - i - 1}{j - 1}}{\binom{n}{j}} = \frac{(n - i - 1)!(n - j)!j!}{(n - j - i)!(j - 1)!n!}$$

$$= \frac{(n - j)(n - j - 1) \cdots (n - j - i + 1)j}{n(n - 1) \cdots (n - i)}$$

and both of these converge stochastically to $2^{-(i+1)}$ as n approaches infinity, that $G_{n,k,j}(x)$ converges stochastically to $G_k(x)$. Since these curtailed generating functions involve only powers of x in the finite range between $-4k$ and $+4k$, the limiting relation (6) follows at once.

11. Effectiveness of normal approximation. Fig. 3 shows the relation between the asymptotic distribution of the quadrant sum for large n and a normal distribution with variance 24, i.e., the same variance as that of the asymptotic distribution. The normal approximation is calculated from

$$\Pr(|S_n| \geq m) \approx \Pr\left(x \geq \frac{m - \frac{1}{2}}{\sqrt{24}}\right)$$

where x is normally distributed with zero mean and unit variance. The asymptotic and normal curves agree surprisingly well out to the 5% point, and an error of a full unit in the significance level first occurs beyond the 0.5% point.

Since the asymptotic distributions for the quadrant, octant, hexadecant, dotriacontant,—, sums become more and more normal, the normal approximation will be even better for higher dimensions. In r dimensions, this approximation consists in treating

$$\frac{|S_n| - \frac{1}{2}}{\sqrt{12r}}$$

as the absolute value of a standard deviate. This should be quite adequate for large samples and $r \geq 4$.

12. Unsolved problems. The central unsolved problem in connection with the quadrant sum is:

(1) What is the operating characteristic?

This has as a corollary the more general question:

(2) How can the operating characteristic of a nonparametric test be described so as to be useful to the users of the test?

There are, of course, minor problems which are much more easily soluble. A few, listed in order of practical importance, are:

(3) What is the effect on the significance levels of the use of lagged values of x as values of y?

(4) What are the exact distributions for moderate n in three or more dimensions?

(5) Do the analogous limiting distributions hold for three or more dimensions?

(6) What is a better approximation to the limiting distribution for moderate n?

To encourage others to solve some of these, we close with the assurance that they have our good wishes.

REFERENCES

[1] W. A. Shewhart, "Contribution of statistics to the science of engineering," *University of Pennsylvania Bicentennial Conference*, Volume on Fluid Mechanics and Statistical Methods in Engineering, University of Penn. Press, 1941, pp. 97–124. Also Monograph B-1319, *Bell Tel. System Tech. Pub.*, 1941.

[2] L. H. C. Tippett, *Statistical Methods in Industry*, pamphlet published by Iron and Steel Research Council of the British Iron and Steel Federation, 1943.

[3] A. Wald, "An extension of Wilks' method for setting tolerance limits," *Annals of Math. Stat.*, Vol. 14 (1943), pp. 45–55.