5

ON A DISTANCE FUNCTION FOR ORDERED LISTS

LAURENT SIKLOSSY

DEPARTMENT OF INFORMATION AND COMPUTER SCIENCE

UNIVERSITY OF CALIFORNIA, IRVINE

IRVINE, CALIFORNIA 92664

TECHNICAL REPORT NO. 5, OCTOBER 1970

Abstract.

Given two ordered lists of the same elements, we define their distance as the sum for each element of the absolute value of the difference of each element's position in the two lists. Various properties of this distance function are exhibited. In particular, a given list is "far", on the average, from a random list of the same elements.

1. Introduction.

In numerous cases researchers are led to compare ordered lists of the same elements. In particular, a mathematical measure of the closeness of such lists is desired. Intuitive requirements for such a distance are that two identical lists be a null distance apart and that a list and its reverse be far apart. An additional requirement may be that a given list be usually "far" from some random ordered list of the same elements. This requirement insures that the closeness of two lists is a statistically significant property.

In chapter 2 of Mathematical models in the social sciences, Kemeny and Snell show that the distance function that we shall discuss is the only metric (up to scale factor) for preferences that allow equality of choice and that satisfy the mid-point property. Our approach is in the other direction; we postulate a distance and derive the fact that it is a metric, together with additional properties. This report is consequently complementary to the work of Kemeny and Snell.

In all that follows, lists are assumed finite, ordered and containing pairwise different elements. Unless otherwise specified, lists will contain N elements, $N \geq 1$. Elements are ordered from left to right as 1st, 2nd, ..., Nth.

2. Elementary description.

Given two lists, we obtain their distance by summing over the elements the absolute value of the difference of the position in the two lists of each element.

Example: Consider the lists $L_1 = (A,C,B,D)$ and $L_2 = (C,D,B,A)$.

Element	Position in 1st list	Position in 2nd list	Contribution to distance
A	1	4	3
C	2	1	1
B	3	3	0
D	4	2	2

The distance between the two lists is 6.

In general and more precisely, given two lists $L = (L_1,L_2,\ldots,L_N)$ and $K = (K_1,K_2,\ldots,K_N)$ of the same elements let P be a mapping of the N-tuple L onto (1,2,...,N) defined by its projections $P(L_j) = j$. We define the distance between L and K as the distance between $P(L) = (1,2,\ldots,N)$ and $P(K) = (P(K_1),P(K_2),\ldots,P(K_N))$;

$$d(L,K) = d(P(L),P(K)) = \sum_{j=1}^{N} |j-P(K_j)|. \qquad (1)$$

The special role of the identity permutation $I = (1,2,\ldots,N)$ is needed.

d((A B C),(B A C)) = 2 = d((1 2 3),(2 1 3)). However, if we let P((A B C)) = (1 3 2) then P((B A C)) = (3 1 2) but $|1-3|+|3-1|+|2-2|=4$.

We shall from now on consider only lists as permutations of $I = (1,2,\ldots,N)$. For brevity we shall call the distance function sd.

$$sd\ (K,L) = \sum_{j=1}^{N} |j - (K^{-1}\ L)_j| \qquad (2)$$

where K^{-1} is the inverse permutation of K and $(KL)_j \equiv K_{L_j}$.

Note that by a change of variables we obtain:

$$sd\ (K,L) = \sum_{j=1}^{N} |L^{-1}{}_j - K^{-1}{}_j| \qquad (3)$$

3. Metric.

Given permutations $K = (K_1,\ldots,K_N)$, $L = (L_1,\ldots,L_N)$, $M = (M_1,\ldots,M_N)$ of $(1,2,\ldots,N)$:

a) $sd(K,L) = 0$ if and only if $K = L$. (4)

b) $sd\ (K,L) = sd\ (L,K)$. (5)

Proof. Eq. (3) is symmetric in K and L.

c) $sd\ (K,L) \le sd\ (K,M) + sd\ (M,L)$ (6)

Proof. Using (3) and a property of the absolute value, we have:

$\sum_{j=1}^{N} |K^{-1}{}_j - L^{-1}{}_j| \le \sum_{K_j=1}^{N} |K^{-1}{}_j - M^{-1}{}_j| + \sum_{j=1}^{N} |M^{-1}{}_j - L^{-1}{}_j|$, which is the desired result.

Equations (4), (5) and (6) show that sd is a metric.

Another interesting property is:

d) $sd\ (I,K) = sd\ (I,K^{-1})$ (7)

Proof.

$$sd(I,K) = \sum_{j=1}^{N} |j-K_j| = \sum_{K_j=1}^{N} |K^{-1}{}_j - j| = sd(I,K^{-1}).$$

If we consider lists K and L over different sets of elements, we could define a distance from K to L as the restriction of K and L to the elements of $K \cap L$ (require $K \cap L \neq \emptyset$). However properties a) and c) above fail to hold.

Example: $K = (1,2)$, $L = (2,1)$, $M = (1)$.

Then $d(K,M) = sd((1),(1)) = 0$ but $K \neq M$ and $d(K,L) = 2$ while $d(K,M) = d(L,M) = 0$.

4. Evenness.

Theorem. sd is even.

Proof. Given the permutations K and L, we can assume $L = I = (1,2,\ldots,N)$. We show that if sd (K,L) is even, so is sd (K,L') where L' is obtained from L by a transposition. We proceed by case analysis.

a) $sd(I,I) = 0$ is even.

b) Assume sd (I,K) is even.

$$sd(I,K) = \sum_{i=1}^{N} |i-K_i|$$

In K permute K_i and K_j to obtain K', $sd(I,K') = sd(I,K) + \alpha$

where $\alpha = \{ -|j-K_j| - |i-K_i| + |j-K_i| + |i-K_j| \}$

Case 1. $i,j \leq K_i, K_j$ or $K_i, K_j \leq i,j$.

$$\alpha = 0 \qquad (8)$$

Case 2. $j \leq K_j \leq i \leq K_i$ or $K_i \leq i \leq K_j \leq j$.
$|\alpha| = 2|i-K_j|$

Case 3. $i \leq K_i \leq j \leq K_j$ or $K_j \leq j \leq K_i \leq i$.
$|\alpha| = 2|i-K_i|$

Case 4. $i \leq K_j \leq j \leq K_i$ or $K_i \leq j \leq K_j \leq i$.
$|\alpha| = 2|j-K_j|$

Case 5. $j \leq K_i \leq i \leq K_j$ or $K_j \leq i \leq K_i \leq j$.
$|\alpha| = 2|i-K_i|$

Case 6. $j \leq K_i, K_j \leq i$ or $i \leq K_i, K_j \leq j$.
$|\alpha| = 2|K_i-K_j|$

Case 7. $K_i \leq i,j \leq K_j$ or $K_j \leq i,j \leq K_i$.
$|\alpha| = 2|i-j|$

Since any permutation can be obtained by a finite number of transpositions successively applied to I, the proof is complete.

5. Maxima.

Theorem. sd (K,I) is a maximum if K is the reversed permutation of I.

Proof. We assume K = (N,N-1,...,2,1) and proceed by induction over N.

a) The theorem holds for N = 1 and N = 2 by inspection of all possible cases.

b) With N > 2, assume the theorem holds for (N-2).

$$sd(I,K)= \sum_{i=1}^{N} |i-((N+1)-i)| = |1-N + \sum_{i=2}^{N-1} |i-((N+1)-i)| + |N-1| \qquad (9)$$

The second term in (9) equals $\sum_{i=1}^{N-2} |i-((N-1)-i)|$ and is the maximum distance between (1,2,...,N-2) and (N-2,...,2,1).

The two other terms in (9) both equal (N-1) and are the maximum possible contribution of a single summand of the expression for sd (I,K).

Theorem. The maximum value of sd (I,K) is

a) $\frac{1}{2}(N^2-1)$ for N odd and b) $\frac{1}{2}N^2$ for N even.

Proof. The calculations are for I and its reverse.

for N even, $sd_{max} = 2\sum_{i=1}^{N/2} (N+1-2i) = \frac{N^2}{2}$

for N odd, $sd_{max} = 2\sum_{i=1}^{\frac{N+1}{2}} (N+1-2i) = \frac{1}{2}(N^2-1)$.

Theorem. A necessary and sufficient condition for $K = (K_1,\ldots,K_N)$ to be a maximum distance from I is that:

a) for N even: $(K_1,K_2,\ldots,K_{\frac{N}{2}})$ be a permutation of $(1,2,\ldots,\frac{N}{2})$.

and $(K_{\frac{N}{2}+1},\ldots,K_{N-1},K_N)$ be a permutation of $(1,2,\ldots,\frac{N}{2})$.

b) for N odd: $(K_1,K_2,\ldots,K_J)$ be a permutation of $(N-J+1,\ldots,N-1,N)$

and $(K_{J+1)}\ldots,K_{N-1},K_N)$ be a permutation of $(1,2,\ldots,J)$

where J has the same value $\frac{N-1}{2}$ or $\frac{N+1}{2}$ in all four expressions.

Proof. We give the proof for N even. The proof for N odd is a slight variation.

Sufficiency. $(N,N-1,\ldots,\frac{N}{2}+1,\frac{N}{2}-1,\ldots,2,1)$ is at a maximum distance from I. We proceed by induction. Let $K = (K_1,\ldots,K_N)$ be at a maximum distance from I and satisfy condition a) above. Transpose K_i and K_j in K such that

$i,j>\frac{N+1}{2}$ and $K_i,K_j<\frac{N+1}{2}$ or $i,j<\frac{N+1}{2}$ and $K_i,K_j>\frac{N+1}{2}$. The resultant permutation K' satisfies a) above and by equation (9) $sd(I,K)=sd(I,K')$.

All the above transpositions will generate exactly all the permutations satisfying a).

Necessity. Assume a permutation K is at a maximum distance from I in such a way that a) above is not satisfied. Then there are at least a pair of indices i and j such that:

$$i,K_i<\frac{N+1}{2}<j,K_j$$

Let K' be obtained from K by transposing K_i and K_j.

$sd(I,K')=sd(I,K)+\alpha$, where

$$\alpha=\{-|i-K_i|-|j-K_j|+|i-K_j|+|j-K_i|\}$$

Case analysis:

$i \geq K_i, j \geq K_j$: $\alpha = 2\,(K_j - i) > 0.$

$i \leq K_i, j \geq K_j$: $\alpha = 2\,(K_j - K_i) > 0.$

$i \leq K_i, j \leq K_j$: $\alpha = 2\,(j - K_i) > 0.$

$i \geq K_i, j \leq K_j$: $\alpha = 2\,(j - i) > 0.$

It follows that $sd\,(I,K') > sd_{max}$, a contradiction.

Theorem.

The number of maxima is:

a) for N even: $\left(\left(\frac{N}{2}\right)!\right)^2$

b) for N odd: $N\left(\left(\frac{N-1}{2}\right)!\right)^2$

Proof. We use the previous theorem. a) follows immediately.

For b): let N=2k+1. The contributions for $J = \frac{N+1}{2}$ and $J = \frac{N-1}{2}$ are each $k!\,(k+1)!$ but we counted twice the permutations where $K_{\frac{N+1}{2}} = \frac{N+1}{2}$

Since there are $(k!)^2$ such cases, we obtain the desired result:

$$2k!\,(k+1)! - (k!)^2 = (2k+1)\,(k!)^2.$$

6. Average.

Theorem. The sum, over all $N!$ permutations, of the distances to I is $\frac{1}{3}\,N!\,(N^2-1)$.

Proof. Let P_j denote a permutation. The sum is:

$$\text{sum} = \sum_{j=1}^{N!} \sum_{i=1}^{N} |i - (P_j)_i| = \sum_{i=1}^{N} (N-1)! \sum_{j=1}^{N} |i-j| =$$

$$(N-1)! \sum_{i=1}^{N} \left(\sum_{s=1}^{i-1} s + \sum_{s=1}^{N-i} s \right) = \frac{1}{3}\,N!\,(N^2-1).$$

Corollary.

Assuming all permutations equally probable, the average distance between two permutations is $\frac{1}{3}(N^2-1)$.

Note. Half the maximum possible distance is $\sim\frac{N^2}{4}$

so that for N>1 the average distance is larger than half the maximum distance: our distribution is therefore skewed towards the right.

7. Comparison with rank order correlation.

Rank order correlation has been used to test the statistical proximity of ordered lists. In our notation, the Spearman coefficient of a permutation K is:

$$r_s = 1 - \frac{6 \sum_{j=1}^{N} \left((j-K_j)^2 \right)}{N^3 - N}.$$

The computational difficulty is larger since an extra squaring is required. Moreover, the sum $\sum (j-K_j)^2$ grows much faster than $\sum |j-K_j|$; in fact the maxima are $\frac{N^2}{2} + o(N)$ and $\frac{N^3}{3} + o(N)$ so that a problem of integer overflow may occur.

It is to be noticed moreover that for $r(I,K) = \sum_{j=1}^{N} (j-K_j)^2$, the triangle inequality does not hold.

Example:

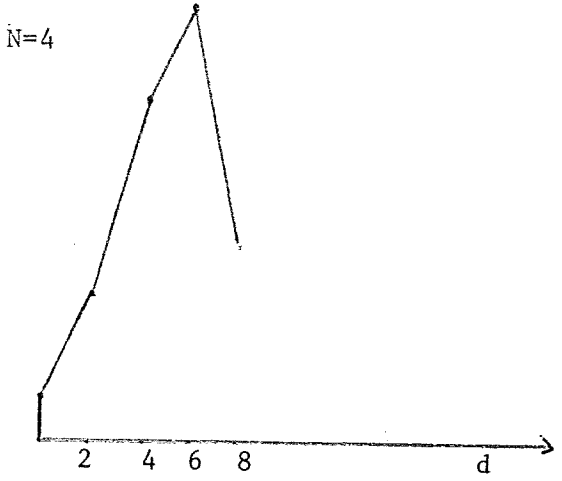
$r((1,2,3),(3,1,2))=6 > r((1,2,3),(1,3,2))+r((1,3,2),(3,1,2))=2+2=4$.*

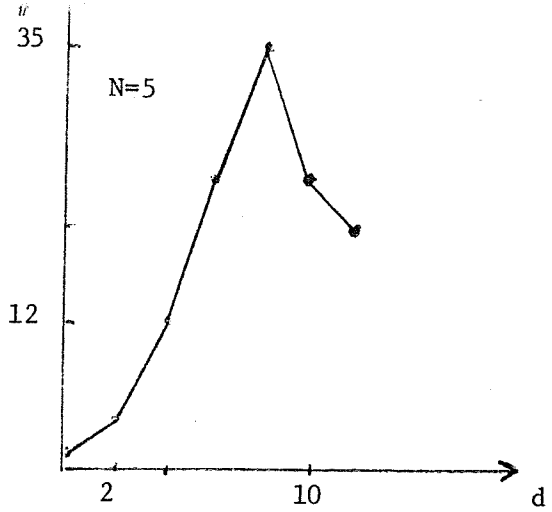
This proves that r is not a metric.

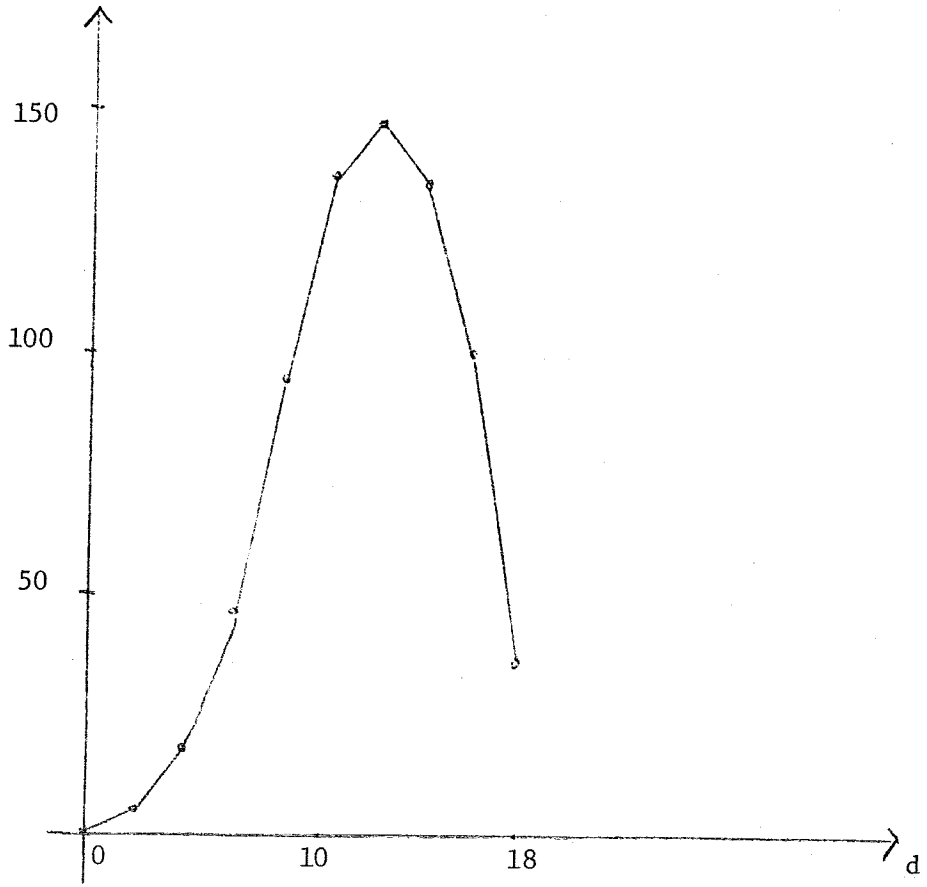
These combined reasons, ease of calculation and metric properties, make sd a more desirable function to work with.

---

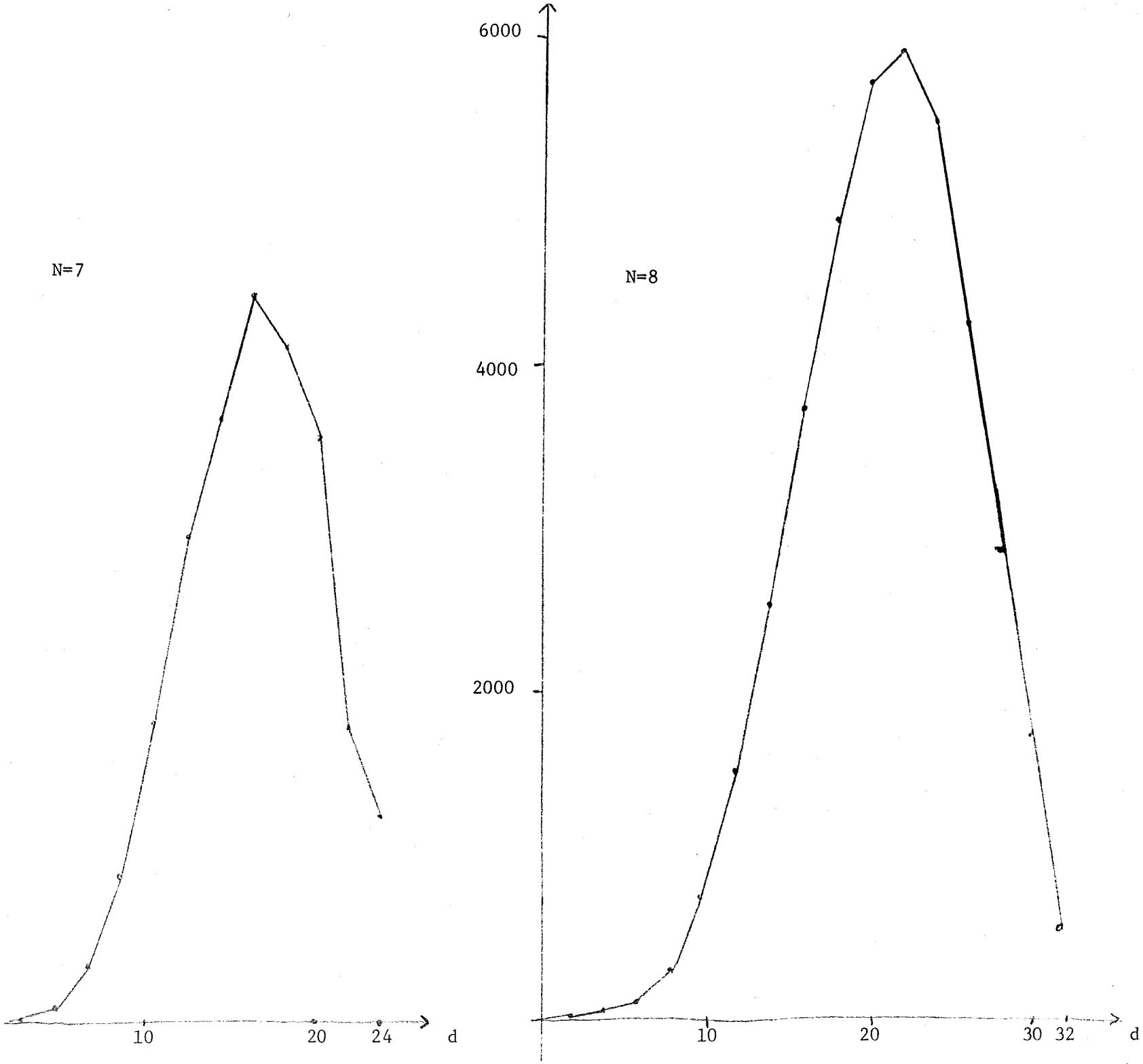
* Note that $r((1,3,2),(3,1,2)) = r((1,2,3),(2,1,3)) = 2$.







Plots of Table 1.


N=7
N=8
6000
4000
2000
10
20
24
d
10
20
30
32
d

N=	2	3	4	5	6	7	8
total number elements	2	6	24	120	720	5040	40320
average distance	1	$2\frac{2}{3}$	5	8	$11\frac{2}{3}$	16	21
number of elements at distance d							
d=0	1	1	1	1	1	1	1
d=2	1	2	3	4	5	6	7
d=4		3	7	12	18	25	33
d=6			9	24	46	76	115
d=8			4	35	93	187	327
d=10				24	137	366	765
d=12				20	148	591	1523
d=14					136	744	2553
d=16					100	884	3696
d=18					36	832	4852
d=20						716	5708
d=22						360	5892
d=24						252	5452
d=26							4212
d=28							2844
d=30							1764
d=32							576

Table 1. Distribution of the distances for permutations of length N, $2 \leq N \leq 8$.