ZASTOSOWANIA MATEMATYKI
APPLICATIONES MATHEMATICAE
Hugo Steinhaus Jubilee Volume
X (1969)

B. KOPOCIŃSKI and E. TRYBUSIOWA (Wrocław)

THE DISTRIBUTION OF THE DISCRETE TREE LENGTH ON A LINE

1. Assume a point process on a line in which the distances between successive points are independent, nonnegative, and identically distributed random variables. Let us join every point of this process with the nearest one: we obtain sets of linked points which are called trees (see [3], [4]). The number of links in a tree will be called discrete tree length on a line or simply — tree length. It is easy to see that tree lengths are identically distributed random variables, the distribution of which does not depend upon the distribution of the distances between successive points of the given point process.

The present note answers a question posed by S. Zubrzycki, namely, what is the distribution of the discrete tree length. More interesting but more difficult, too, are similar questions asked by Zubrzycki for a Poisson point process on the plane: what is the distribution of the tree length on the plane, and what is the distribution of the number of links emerging from a given point. These distributions, estimated empirically, have been applied to comparisons of the configuration of process points (see [5], [6], [7]).

2. Let $\{D_n;\ -\infty < n < \infty\}$ denote a sequence of discrete tree lengths on a line, and let U_n denote the Euclidean length of the interval between trees number n and $n+1$. From the assumption that point distances are independent and identically distributed follows that the sequence $\{U_n;\ -\infty < n < \infty\}$ is a stationary, transitive Markov chain. Moreover, the discrete length D_{n+1} with condition $U_n = u$ depends upon u and does not depend upon $\{U_{k-1}, D_k;\ -\infty < k < n\}$. Thus it follows (see [2], p. 233) that the correlation coefficient between D_m and D_{m+n} decreases geometrically with increasing n:

$$\begin{aligned}\varrho(D_m, D_{m+n}) &= \varrho(D_m, U_m)\varrho(U_m, U_{m+n-1})\varrho(U_{m+n-1}, D_{m+n}) \\ &= \varrho^n \mathrm{Const}, \quad |\varrho| < 1, \quad n > 0.\end{aligned}$$

Denote by P_n, $n = 1, 2, \ldots$, the probability that a given tree on a line has length n, $P_n = \Pr\{D_j = n\}$, and by p_n, $n = 0, 1, 2, \ldots$, the probability that a given distance between process points on the line belongs to a tree of length n; let p_0 denote here the probability that the distance does not belong to any tree.

Now we define the random variables $D_{j,n}$ in the following way:

$$D_{j,n} = \begin{cases} 1 & \text{if} \quad D_j = n, \\ 0 & \text{if} \quad D_j \neq n, \end{cases} \qquad \text{for} \quad n = 1, 2, \ldots$$

The random variables D_j and, for any n, the random variables $D_{j,n}$ form stationary, transitive stochastic processes, thus from the strong law of large numbers we have (see [2], p. 465)

$$p_0 = \lim_{N\to\infty} \frac{N}{N+D_1+D_2+\ldots+D_N} = \frac{1}{1+E(D_1)},$$

$$p_n = \lim_{N\to\infty} \frac{n[D_{1,n}+D_{2,n}+\ldots+D_{N,n}]}{N+D_1+D_2+\ldots+D_N} = \frac{n\Pr\{D_1 = n\}}{1+E(D_1)}.$$

Hence we obtain

$$p_0 = \frac{1}{1+\sum\limits_{i=1}^{\infty} iP_i}, \tag{1}$$

$$p_n = \frac{nP_n}{1+\sum\limits_{i=1}^{\infty} iP_i} \quad \text{for} \quad i = 1, 2, \ldots, \tag{2}$$

and from above

$$P_n = \frac{p_n}{np_0}. \tag{3}$$

We shall now calculate the probabilities p_n. For p_0 observe that the event that a given distance between process points does not belong to any tree depends only upon the length of this distance and upon the lengths of the neighbouring distances. These distances, however, are independent, any ordering of their lengths is thus equally probable. The assumption of a continuous distribution of those lengths allows us the assumption of different lengths. A given distance does not belong to any tree in two cases only, namely, when the distance of greatest length is the middle one. We have thus $p_0 = 1/3$ and from (1) we get the mean tree length of 2. Similary, we may prove that $p_1 = 2/15$.

To calculate p_n for $n > 1$ consider $n+4$ distances of different lengths. Let $K(n)$ denote the number of those permutations of these distances which result in a tree of length n. A given distance may be in any place of the tree, thus

$$p_n = \frac{nK(n)}{(n+4)!}. \tag{4}$$

We shall prove later that

$$K(n) = n(n+3)2^{n+1}, \tag{5}$$

wherefrom we get

$$p_n = \frac{n^2(n+3)2^{n+1}}{(n+4)!} \tag{6}$$

and for the distribution of the tree length

$$P_n = \frac{3p_n}{n} = \frac{3n(n+3)}{(n+4)!}\,2^{n+1}. \tag{7}$$

The probabilities p_n and P_n for different n are given in Table 1.

TABLE 1. The distributions of discrete tree length

	Probability distributions		Empirical distributions		
n	p_n	P_n	(a)	(b)	(c)
0	0,3333	—	—	—	—
1	0,1333	0,4000	131	412	389
2	0,2222	0,3333	337	335	333
3	0,1714	0,1714	192	157	179
4	0,0889	0,0667	171	54	72
5	0,0353	0,0212	97	33	21
6	0,0114	0,0057	42	7	3
7	0,0031	0,0014	13	1	2
8	0,0008	0,0003	12	1	1
9	0,0002	0,0000	4	—	—
10	0,0000	0,0000	—	—	—
11	0,0000	0,0000	1	—	—
0-11	0,9999	1,0000	1000	1000	1000

3. To prove formula (5) consider $n+4$ distances of different lengths, say of the lengths $1, 2, \ldots, n+4$. For any permutation of these distances $d_1, d_2, \ldots, d_{n+4}$ to form a tree of length n it is necessary and sufficient to fulfil the following three conditions:

(i) $d_1 < d_2$,

(ii) $d_{n+4} < d_{n+3}$,

(iii) there exists a k such that $2 \leqslant k \leqslant n+3$ and $d_2 > d_3 > \ldots$ $\ldots > d_{k-1} > d_k < d_{k+1} < \ldots < d_{n+2} < d_{n+3}$.

We have previously denoted by $K(n)$ the number of permutations satisfying conditions (i), (ii) and (iii). Given any additional condition $\mathscr{A}$

we shall denote by $K(n\,|\,\mathscr{A})$ the number of permutations which satisfy also condition $\mathscr{A}$.

Conditions (i)-(iii) imply either $d_2 = n+4$ or $d_{n+3} = n+4$. For symmetry reasons assume $d_2 = n+4$ and thus

$$(8) \qquad K(n) = 2K(n\,|\,d_2 = n+4).$$

If $d_{n+2} = n+4$ then $3 \leqslant d_{n+3} \leqslant n+3$ and hence

$$(9) \qquad K(n\,|\,d_2 = n+4) = \sum_{i=3}^{n+3} K(n\,|\,d_2 = n+4,\, d_{n+3} = i).$$

From condition (ii) we get

$$(10) \quad K(n\,|\,d_2 = n+4,\, d_{n+3} = i) = \sum_{j=1}^{i-1} K(n\,|\,d_2 = n+4,\, d_{n+3} = i,\, d_{n+4} = j).$$

If $i = 3$ we have

$$(11) \qquad K(n\,|\,d_2 = n+4,\, d_{n+3} = 3) = 2n$$

since only two cases are possible: either $d_{n+4} = 1$ with $d_{n+2} = 2$ following, or $d_{n+4} = 2$ which gives $d_{n+2} = 1$. In both cases d_1 may be chosen in n ways from among numbers $4, 5, \ldots, n+3$: the remaining numbers have to be arranged in an decreasing order on the places $3, 4, \ldots, n+1$.

For $i > 3$ let us form two sets

$$A_{i,j} = \{k\colon k = 1, 2, \ldots, i-1, \text{ and } k \neq j\},$$
$$B_{i,j} = \{k\colon k = i+1, i+2, \ldots, n+3\}.$$

Consider two possibilities. In the first let $d_1 \in A_{i,j}$. Then the element d_1 may be chosen in $i-2$ ways, and the elements of the set $B_{i,j}$ as being greater than d_{n+3}, have to be arranged in decreasing order on the places $3, 4, \ldots, n-i+5$: the remaining $i-3$ elements of set $A_{i,j} - \{d_1\}$ are to be disposed on places $n-i+6, n-i+7, \ldots, n+2$. It is easy verified that, given condition (iii), this may be done in 2^{i-4} ways.

In the second possibility $d_1 \in B_{i,j}$. Now the element d_1 may be chosen in $n-i+3$ ways, the remaining elements of the set $B_{i,j} - \{d_1\}$ are to be arranged in a decreasing order on the places $3, 4, \ldots, n-i+4$, and the remaining $i-2$ elements of $A_{i,j}$ may be disposed on places $n-i+5$, $n-i+6, \ldots, n+2$ in 2^{i-3} ways. Thus

$$(12) \quad K(n\,|\,d_2 = n+4,\, d_{n+3} = i,\, d_{n+4} = j) = (i-2)2^{i-4} + (n-i+3)2^{n-3}.$$

From (8)-(12) we obtain

$$K(n) = 4n + 2\sum_{i=4}^{n+3}\sum_{j=1}^{i-1}\left[(i-2)2^{i-4} + (n-i+3)2^{i-3}\right]$$
$$= 4n + \sum_{i=0}^{n-1}(2n-1)(i+3)2^{i+1} = n(n+3)2^{n+1}.$$

This ends the proof of formula (5).

4. The distribution of tree length may be found as a simple exercise by Monte Carlo methods. We have used in our calculations three pseudo-random number generators, namely (a) the Fibonacci generator, (b) the multiplicative generator, both of which belong to the programme library of the Odra 1003 computer, and (c) the middle-square generator used e.g. in [1].

The empirical distributions of 1000 trees obtained by using the above mentioned generators are given in the last three columns of Table 1. They cannot be compared by means of the standard tests of fit since the tree lengths on a line are dependent. It seems, however, that the empirical distribution from type (a) generator is too far from the expected distribution. It was not known a priori which of the generators was best suitable for the Monte Carlo calculations required.

References

[1] E. D. Cashwell and C. J. Everett, *A practical manual on the Monte Carlo method for random walk problems*, Pergamon, London 1959.

[2] J. L. Doob, *Stochastic processes*, J. Wiley, New York 1953.

[3] K. Florek, J. Łukaszewicz, J. Perkal, H. Steinhaus and S. Zubrzycki, *Sur la liaison et la division des points d'un ensemble fini*, Colloq. Math. 3-4 (1951). pp. 282-285.

[4] —, *Taksonomia wroclawska*, Przegl. Antropol. 27 (1951), pp. 193-211.

[5] H. Kowarzyk, H. Steinhaus and S. Szymaniec, *Arrangement of chromosomes in human cells*, I. *Associacions in methaphase figures*, Bull. Acad. Polon. Sci. Cl. VI 12 (1965), pp. 321-326.

[6] L. Zubrzycka, *O rozmieszczeniu punktów próbkowych na plaszczyźnie*, Zastosow. Matem. 5 (1960), pp. 161-171.

[7] S. Zubrzycki, *O łańcuszkach gwiezdnych*, Zastosow. Matem. 1 (1954), pp. 157-205.

DEPT. OF APPLIED MATHEMATICS, UNIVERSITY OF WROCŁAW
DEPT. OF MATHEMATICS, GRADUATE SCHOOL OF ECONOMICS, WROCŁAW

Received on 18. 11. 1967