ON THE ESTIMATION OF AVERAGE LENGTH OF CHAINS IN THE COMMUNICATION-PATTERN

By YASUSHI TAGA AND TATSUZO SUZUKI

(Received Nov. 25, 1957)

1. Introduction

It is an interesting problem to represent several patterns which denote analytically one-dimensional relations between the elements in a discreate group. The estimation of some parameters which characterize these patterns, based on the random sampling method, is also an interesting and important problem, but it proves sometimes difficult.

In this paper, we shall treat mainly the problem about the pattern of personnal communication, and we consider that this method is applicable also to analyse some other patterns such as sociometric patterns or patterns showed in branching processes etc., and about these problems we shall discuss in the near future.

We wish to thank Dr. C. Hayashi and K. Ishii (The Institute fo Statistical Mathematics) for many helpful suggestions.

2. The representation of patterns and parameters

For simplicity we shall treat the model in the case where the organization of the group of elements is one-dimensional, in the sense that the organization for only one activity of the group is considered.

We take up a group of elements, π, which contains N units. In a graph, elements of π are represented by N points. A connection from element i to element j is represented by a directed line from i to j; the absence of such a connection by no line from i to j. In the case i and j have any connection, we consider there exists *a directed connection from i to j* (or from j to i). Further, we call element i a starting-point when a directed line starts from i and no directed line reaches i, and j a terminal-point when a directed line terminates at j and no one starts from j. And if there are two directed connections from i to j and from j to k, then we consider there exists *a directive relation through i, j and k*, and we call such a connection *a length 2-directive relation* (or sub-chain) from i to k, and a directed connection from i to j might be taken for a length 1-directive relation from i to j.

In general, if there exist successive directed connections from i to k through each of l–1 distinct elements, $j_1, j_2, \cdots, j_{l-1}$ ($j_1, \cdots, j_{l-1} \in \pi$), then we call it *a length l-directive relation* (or sub-chain) from i to k.

Especially, when there exists a l-directive relation from i to k such that element i is a starting-point and element k is a terminal-point, then we call such a l-directive relation *a (length) l-chain from i to k*. Clearly, a directed connection from i to k such that i is a starting-point and k is a terminal-point is a (length) 1-chain from i to k.

The equivalent matrix representation is described as follows. Let e_{ij} be a function of element i and element j ($i, j=1, 2, \cdots, N, i \neq j$) such that $e_{ij}=1$ if a directed connection exists from element i to element j, and $e_{ij}=0$ otherwise, and $e_{ii}=0$. Clearly, $e_{ij}=1$ or 0 according as a directed connection from i to j exists or not. Then the pattern of directed connection (or length 1-directive relation) in the group is represented by $N \times N$ matrix, that is

$$E_1=(e_{ij})$$

where e_{ij}'s are the above-mentioned.

If for i, j and k the relations

$$e_{ij}=1 \quad \text{and} \quad e_{jk}=1 \qquad (\text{for any } j,\ j \neq i \neq k)$$

hold, then there exists a directive relation between i and k through j and, therefore, the actual number $d_{ik}^{(2)}$ of the length 2-directive relations from i to k in the pattern is represented by $d_{ik}^{(2)}=\sum_j e_{ij}e_{jk}$ ($j=1, 2, \cdots, N$ $j \neq i \neq k$). Thus the pattern of length 2-directive relation in the group is represented by an $N \times N$ matrix, that is

$$E_2=(E_1)^2=\left(\sum_j e_{ij}e_{jk}\right)=(d_{ik}^{(2)})$$

In general, the pattern of (length) l-directive relation in the group is also represented by an $N \times N$-matrix,

$$E_l=(E_1)^l=(d_{ik}^{(l)})$$

where $d_{ik}^{(l)}$ is the actual number of (length) l-directive relations from i to k in the pattern.

For simplification, let us take the following assumption.

[Assumption 1] Any chain is non-reflexive, that is, for any set of elements, $i, j_1, j_2, \cdots, j_{l-1}$,

$$e_{ij}, e_{j_1 j_2} \cdots e_{j_{l-1} i} = 0 \quad \text{(for all } i \text{ in the group).}$$

In the graph, there exists no chain illustrated in Fig. 1.

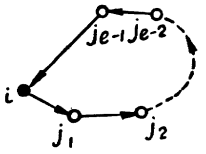

Fig. 1.

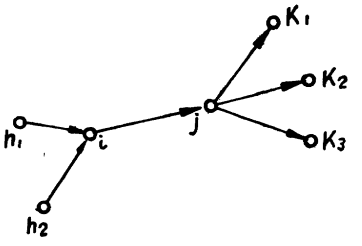

Fig. 2.

Now, we define the counting method of number of directive relations. In the general case, several chains might have the same directive relation as a common part, then we should like to count it just as times as the number of chains which contain it.

For example, in the case illustrated in Fig. 2, we consider there exist six (length) 3-chains, each of which starts from elements h_1 or h_2, and reachs k_1 or k_2 or k_3; i.e. they are following six 3-chains, namely

$$h_1 \to i \to j \to k_1, \quad h_1 \to i \to j \to k_2, \quad h_1 \to i \to j \to k_3,$$
$$h_2 \to i \to j \to k_1, \quad h_2 \to i \to j \to k_2 \quad \text{and} \quad h_2 \to i \to j \to k_3.$$

Thus the (length) 1-directive relation from i to j is counted exactly six times.

In general, let f_i be the total number of chains which reach i, regarding i as a terminal point, and let b_j be the total number of chains regarding j as a starting point. In particular, we put $f_i=1$ for a starting point i, and $b_j=1$ for a terminal point j. Then we must count every directive relation from i to j just $f_i \times b_j$ times if there exists such a relation. Clearly, $f_i \times b_j$ is the total number of chains which have the directive relation from i to j as a common part. Here we put

$$e_{ij}^{(l)} = f_i d_{ij}^{(l)} b_j \ .$$

Let $\mathfrak{A}$ be the row vector with elements a_i, where $a_i=1$ if the element i is a starting-point, and $a_i=0$ otherwise, and $\mathfrak{C}$ be the column vector with elements c_j, where $c_j=1$ if the element j is a terminal-point, and $c_j=0$ otherwise, clearly, $a_i = \prod_h (1-e_{hi})$, and $c_j = \prod_k (1-e_{jk})$. And let T_l be the total number of l-chains, then clearly

$$T_l=\mathfrak{A}E_l\mathfrak{C}\ . \tag{2.1}$$

Moreover, let S_l be the total number of l-directive relations, counting in multiplicity as mentioned above and finally, let $\mathfrak{F}$ and $\mathfrak{B}$ be the row and column vector in N-dimension with elements f_i and b_j, respectively, then

$$S_l=\mathfrak{F}E_l\mathfrak{B}\ . \tag{2.2}$$

Clearly, the relation $S_l=\sum_{i,j} f_i d_{ij}^{(l)} b_j$ holds.

Now the following relation between $\mathfrak{A}$ and $\mathfrak{F}$ can be easily obtained,

$$\mathfrak{F}(I-E_1)=\mathfrak{A} \tag{2.3}$$

where I is the unit matrix in N-dimension.

And also the relation between $\mathfrak{C}$ and $\mathfrak{B}$

$$(I-E_1)\mathfrak{B}=\mathfrak{C} \tag{2.4}$$

holds. Because the equation

$$f_j-\sum_{i=1}^{N} f_i e_{ij}=1 \text{ or } 0$$

holds according as element j is a starting point or not, and the equation

$$b_j-\sum_{i=1}^{N} e_{ji} b_i=1 \text{ or } 0$$

holds according as element j is a terminal point or not, therefore, from the equations (2.1)$\sim$(2.4) we have

$$\begin{aligned} T_l&=\mathfrak{F}(I-E_1)E_l(I-E_1)\mathfrak{B} \\ &=\mathfrak{F}E_l\mathfrak{B}-2\mathfrak{F}E_{l+1}\mathfrak{B}+\mathfrak{F}E_{l+2}\mathfrak{B} \\ &=S_l-2S_{l+1}+S_{l+2} \end{aligned} \tag{2.5}$$

and hence

$$S_l=\sum_{\rho=1}^{N-l}\rho T_{l+\rho-1}\ . \tag{2.6}$$

Moreover, we define the average length of chains, $\overline{T}$ by

$$\overline{T}=\frac{\sum_{l=0}^{N-1} lT_l}{\sum_{l=0}^{N-1} T_l}$$

then from (2.5) and (2.6) we have

$$\overline{T}=\frac{S_1}{S_0-S_1}\ . \tag{2.7}$$

($\overline{T}$ might be taken for the power of the original information given in the population π.)

Secondly, we consider some special cases.

Case 1. Let the following assumption be satisfied.

[Assumption 2] For any two chains, there exists no common elements except their starting-points or terminal-points, that is, if

$$e_{ij_1}e_{j_1j_2}\cdots e_{j_{l-1}k}=1$$
$$e_{i'j_1'}e_{j_1'j_2'}\cdots e_{j_{m-1}'k'}=1\ ,$$

then always

$$e_{ij_1}e_{j_1j_2}\cdots e_{j_{p-1}j_p'}e_{j_p'j_{p+1}}\cdots e_{j_{l-1}k}=0\ .$$

Then all chains in the group have no circular-part, and any two chains are not connected except at the starting-point or at the terminal-point, so the problem is reduced to a simple case; i.e.

$f_i=1$ if i is not a terminal-point
$b_j=1$ if j is not a starting-point,

so the number of l-directive relations from i to k is

$$e_{ik}^{(l)}=\sum_{j_1\cdots j_{l-1}}e_{ij_1}e_{j_1j_2}\cdots e_{j_{l-2}j_{l-1}}e_{j_{l-1}k}$$

and let

$$S_i^{(l)}=\sum_{k=1}^{N}e_{ik}^{(l)}$$

be the number of l-directive relations starting from the element i, then

$$S_l=\sum_{i=1}^{N}S_i^{(l)}\ .$$

And other result are the same.

Case 2. Our general case, however, is somewhat complicated in the actual investigation, so we shall set the following assumption which is slightly weaker than Assumption 2.

[Assumption 2'] Any two chains do not join into one except at the terminal-point, that is,

$f_i=1$ (for any i except the terminal-point).

In the actual survey, considering the time-lag, the assumption may be practical.

3. The estimation of T_i and $\overline{T}$

In this section, we shall preliminary consider the procedure in the actual survey. Suppose we take up a simple random sample of n elements drawn from a universe of N elements by equal probability sampling without replacement, then in the first step, element i in the propability sample will be asked for the connections to all elements in the universe, and we define x_{ij} same as e_{ij}, where x_{ij} is a random variable which represents the relation between i and j ($i=1, \cdots, n$, $j=1, 2, \cdots, N$, $i \neq j$), and $x_{ij}=1$, if there exists a directed connection from i to j, and $x_{ij}=0$ otherwise.

Consider the situation

$$x_{ij}=1 \qquad \text{(for any } j \text{ in the universe)}$$

then in the second step, we shall ask for such all j's the relation to all other elements in the universe, and define $x_{jk}=1$ or 0 according as the directed connection from j to k exists or not, and continue these procedures until the successive connections reach the (all) terminal points.

On the contrary, we may go back along the successive connections from i to the (all) starting points. Thus, we are able to know the multiplicities f_i (or b_k) about the sample point i. In general, we define

$$x_{ik}^{(l)}=f_i\Big(\sum_{j_1,\cdots,j_{l-1}} x_{ij_1}x_{j_1j_2}\cdots x_{j_{l-1}k}\Big)b_k \tag{3.1}$$

similar to $e_{ik}^{(l)}$, where $x_{ik}^{(l)}$ denotes the number of chains which have the l-relations from i to k as the common part. Thus, starting from a random sample i, we shall be able to obtain the value of $x_{ik}^{(l)}$ as the result of the survey ($i=1, \cdots, n$, $k=1, \cdots, N$, $i \neq k$).

Let

$$x_i^{(l)}=\sum_{k=1}^{N} x_{ik}^{(l)} \tag{3.2}$$

be the number of the chains, which go through i and whose lengths from i are equal to or longer than l, then according to our survey procedure, $x_i^{(l)}$ is a random sample drawn from a population of N units, where the population consists of $S_i^{(l)}$ ($i=i, \cdots, N$) and our estimation problem can

be reduced to a simple one. Now, from the result in the previous section, $\overline{T}_l$ and T are represented as a function of S_l's respectively, so in the first, we consider the estimation of S_l. It is natural to take

$$\hat{S}_l=\frac{N}{n}\sum_{i=1}^{n} x_i^{(l)} \tag{3.3}$$

as the estimate of S_l, and $\hat{S}_l$ is clearly the unbiased estimate of S_l. The variance of S_l, $\sigma_{\hat{S}_l}^2$, is

$$\sigma_{\hat{S}_l}^2=N^2\frac{N-n}{N-1}\frac{\sigma_l^2}{n} \tag{3.4}$$

where

$$\sigma_l^2=\frac{\sum_{i}^{N}(S_i^{(l)}-\overline{S}_l)^2}{N}, \qquad \overline{S}_l=\frac{\sum_{i}^{N} S_i^{(l)}}{N}.$$

And we set

$$\hat{T}_l=\hat{S}_l-2\hat{S}_{l+1}+\hat{S}_{l+2} \tag{3.5}$$

as the estimate of T_l, then

$$E(\hat{T}_l)=T_l \tag{3.6}$$

and the variance of $\hat{T}_l$, $\sigma_{\hat{T}_l}^2$ is

$$\sigma_{\hat{T}_l}^2=N^2\frac{N-n}{N-1}\frac{1}{n}[\sigma_l^2+4\sigma_{l+1}^2+\sigma_{l+2}^2-4\sigma_{l,l+1}-4\sigma_{l+1,l+2}+2\sigma_{l,l+2}] \tag{3.7}$$

where

$$\sigma_{l,l+1}=\frac{\sum_{i=1}^{N}(S_i^{(l)}-\overline{S}_l)(S_i^{(l+1)}-\overline{S}_{l+1})}{N}.$$

By (3.6) and (3.7) we estimate the distribution of lengthes of chains in the population.

Moreover, we take

$$\hat{T}=\frac{\hat{S}_1}{\hat{S}_0-\hat{S}_1}$$

as the estimate of $\overline{T}$, unless $\hat{S}_0-\hat{S}_1\neq 0$. Then

$$E(\hat{T})\neq\overline{T},$$

that is, $\hat{T}$ is not the unbiased estimate of $\overline{T}$, for $\hat{T}$ is the ratio estimate. The variance of $\hat{T}$ is approximately as follows

$$\sigma_{\hat{T}}^2 \fallingdotseq N^2 \frac{N-n}{N-1}\frac{1}{n}\left[\frac{S_0^2\sigma_1^2+S_1^2\sigma_0^2-2S_0S_1\sigma_{0\cdot 1}}{(S_0-S_1)^4}\right], \tag{3.9}$$

and the bias of $\hat{T}$, $B_{\hat{T}}$, is approximately

$$B_{\hat{T}}=E(\hat{T}-\overline{T})^2 \fallingdotseq N^2 \frac{N-n}{N-1}\frac{1}{n}\left[\frac{S_0\sigma_1^2+S_1\sigma_0^2-(S_0+S_1)\sigma_{0\cdot 1}}{(S_0-S_1)^3}\right] \tag{3.10}$$

Thus, in the case when both two assumptions (1) and (2), as previously noted, hold, we have only to perform the survey up to the first step for the purpose of estimating the average length of chains, $\overline{T}$. In the case when both two assumptions (1) and (2′), hold, it is only necessary for estimating $\overline{T}$ to pursue the directive relations backward starting from the points selected in the sample. However, in the general case without the assnmptions, the complete survey system mentioned above will only give the estimates of T_i and $\overline{T}$.

Finally, it must be noted that we can obtain the estimate of both the distribution of l-chains and the average length of chains using a random sampling method.

THE INSTITUTE OF STATISTICAL MATHEMATICS