ON A STOCHASTIC MODEL CONCERNING THE PATTERN OF COMMUNICATION
—Diffusion of News in a Social Group—

By YASUSHI TAGA and KEIITI ISII

(Received Aug. 2, 1959)

§ 0. Introduction

We had an experimental study on the channels of communication in several villages in the Tôhoku district, Japan, a few years ago (see [1] and [2]). The aim of the study was to observe through what types of channels a certain information was communicated and what patterns of channels were formed in a social group. Of course, the information was supposed to be so simple that the deformation of its contents was negligible in such communications.

In the present report, we shall study a stochastic model in which a certain information I is communicated from an information source to persons or from a person to another in a system consisting of a social group π and an information source S. We are interested in the pattern of communication formed in our stochastic model, and in estimating the ratio of the two intensities of communications from S to a person and from a person to another. For this purpose, we shall introduce some random variables suitable for characterizing the pattern of communication, and investigate the stochastic behavior of them. But the assumption of homogeneity (both in time and in persons) of communication in the social group plays an essential role in our theory, and our method does not apply to inhomogeneous cases. We shall discuss in the near future the stochastic model for more general groups without such an assumption.

§ 1. Formulation of the problem

Consider a social group π consisting of M persons and an information source S which has a function to communicate constantly a certain information I to persons in π. Suppose that the information I starts from S and is communicated from a person to another in course of time. Following this process up to time t, we should obtain a tree-shaped pattern as is illustrated in Fig. 1. By the pattern $U(t)$ we

mean the shape of such a tree, regardless of when or to whom the information is communicated. Therefore, $U(t)$ represents only the diagrammatic pattern of the tree produced up to time t. We make some stochastic assumptions on our model, and regard $U(t)$ as a random variable representing the state of π at time t.

Suppose that in the time interval $(t, t+\Delta t)$ the transmission of I from person A to person B has probability $\lambda(A, B; t)\Delta t+o(\Delta t)$, while the transmission of I from the information source S to person A has probability $\mu(A; t)\Delta t+o(\Delta t)$, and that the probability of two or more transmissions is $o(\Delta t)$.

We shall consider in this paper a stochastic model in which $\lambda(A, B; t)$ is a constant λ and $\mu(A; t)$ is a constant μ. Further, assume that all possible communications in the group π are made mutually independently, and that the events which occur in the time interval $(t, t+\Delta t)$ are independent of the past history of the system up to time t.

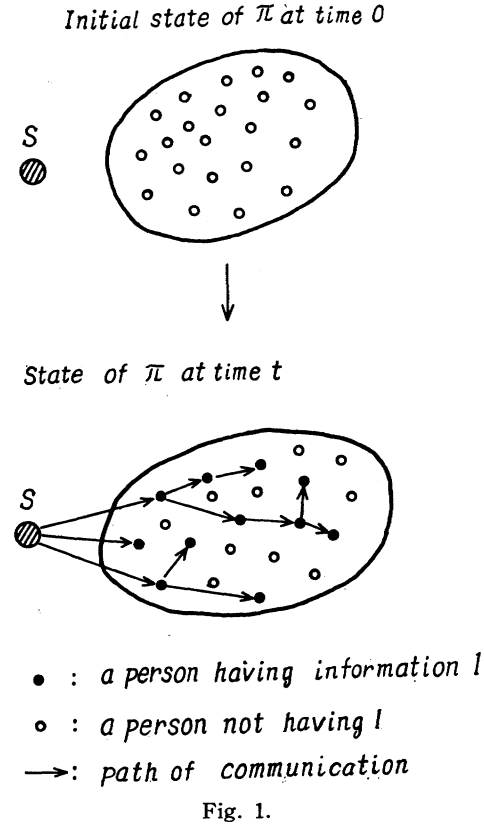

Fig. 1.

Now, the random variable $U(t)$ is so complicated that it might be difficult to treat $U(t)$ itself as a Markov process. We shall, therefore, introduce some random variables to characterize $U(t)$, say, $N(t)$, $K(t)$ and $L(t)$ which are defined as follows. Let $N(t)$ be the random variable representing the number of persons who have received the information I up to time t; $K(t)$ the number of persons who have received I but have not yet communicated it to any of others up to time t; $L(t)$ the number of persons who have received I directly from the information source S up to time t. These random variables can be determined uniquely as functions of $U(t)$. Their probability distributions will be studied in the following sections.

The list of notations often used throughout the paper is given below:

Notations

π : social group under consideration.

M : size of π, that is, the number of the members of π.

I : a certain information.

S : information source of I.

$U(t)$: random variable representing the pattern of communication in π up to time t.

$N(t)$: random variable which represents the number of persons having received I up to time t.

$K(t)$: random variable which represents the number of persons who have received I but not yet communicated it to any other person up to time t.

$L(t)$: random variable which represents the number of persons who have received I directly from S up to time t.

$A_n(t)$: state of π at time t such that $N(t)=n$.

$A_{n,k}(t)$: state of π at time t such that $N(t)=n$ and $K(t)=k$.

$A^*_{n,l}(t)$: state of π at time t such that $N(t)=n$ and $L(t)=l$.

$P_n(t)=\Pr\{N(t)=n\}$

$P_{n,k}(t)=\Pr\{N(t)=n \text{ and } K(t)=k\}$

$P^*_{n,l}(t)=\Pr\{N(t)=n \text{ and } L(t)=l\}$

$p_{n,k}=\Pr\{K(t)=k \mid N(t)=n\}$,

i.e., the conditional probability that $K(t)=k$, given $N(t)=n$.

$p^*_{n,l}=\Pr\{L(t)=l \mid N(t)=n\}$,

i.e., the conditional probability that $L(t)=l$, given $N(t)=n$.

§ 2. Probability distribution of patterns

2.1. Distribution of $N(t)$

We say that π is in state $A_n(t)$ at time t, if the random variable $N(t)$ takes value n $(0\leqq n\leqq M)$ at time t.

Then, under the assumption of our stochastic model mentioned in the preceding section, the probabilities that the transitions $A_n(t)\to A_n(t+\Delta t)$ and $A_{n-1}(t)\to A_n(t+\Delta t)$ will occur in a sufficiently small time interval $(t, t+\Delta t)$ are, respectively, $\{1-(M-n)(n\lambda+\mu)\cdot\Delta t\}+o(\Delta t)$ and $(M-n+1)\{(n-1)\lambda+\mu\}\Delta t+o(\Delta t)$. Therefore, as is well known in the

theory of birth process (see [3]), the probability $P_n(t)$ that $N(t)=n$ satisfies the equation

$$P_n(t+\Delta t)=(1-c_n\Delta t)P_n(t)+c_{n-1}\Delta t P_{n-1}(t)+o(\Delta t)\ .$$

Taking limit for $\Delta t\to 0$, we get the following differential equations:

$$P_n'(t)=-c_nP_n(t)+c_{n-1}P_{n-1}(t)\ , \qquad (1\leqq n\leqq M) \tag{2.1}$$

and

$$P_0'(t)=-\beta_0P_0(t)\ , \tag{2.2}$$

where $\alpha_n=n(M-n)\lambda$, $\beta_n=(M-n)\mu$, $c_n=\alpha_n+\beta_n$.

The solutions of these equations are easily obtained under the initial conditions $P_0(0)=1$, $P_0(n)=0$ $(1\leqq n\leqq M)$.

In fact, we have

$$P_0(t)=\exp(-\beta_0t)\ , \tag{2.3}$$

and if $c_i\neq c_j$ whenever $0\leqq i<j\leqq n$, we have

$$\begin{aligned}P_n(t)&=\exp(-c_nt)\int_0^t\exp(c_n\tau)c_{n-1}P_{n-1}(\tau)d\tau\\&=c_1c_2\cdots c_{n-1}\sum_{i=0}^{n-1}\frac{\exp(-c_it)}{(c_0-c_i)\cdots(c_{i-1}-c_i)(c_{i+1}-c_i)\cdots(c_{n-1}-c_i)}\ .\end{aligned} \tag{2.4}$$

The solutions for the case that some of c_i's are equal are easily obtained by a slight modification.

2.2. JOINT DISTRIBUTION OF $N(t)$ AND $K(t)$.

Suppose we take two-dimensional random variable $(N(t), K(t))$ as a suitable representation of the pattern $U(t)$ up to time t. We consider the state $A_{n,k}(t)$ of the group π at time t, in which $N(t)$ and $K(t)$ take values n and k, respectively $(0\leqq k\leqq n\leqq M)$. Then, the probability that the transitions $A_{n,k}(t)\to A_{n,k}(t+\Delta t)$, $A_{n-1,k}(t)\to A_{n,k}(t+\Delta t)$ and $A_{n-1,k-1}(t)\to A_{n,k}(t+\Delta t)$ occur in a sufficiently small time interval $(t, t+\Delta t)$ are, respectively, $\{1-(M-n)(n\lambda+\mu)\Delta t\}+o(\Delta t)$, $k(M-n+1)\lambda\Delta t+o(\Delta t)$, and $(M-n+1)\{(n-k)\lambda+\mu\}\Delta t+o(\Delta t)$.

Therefore, the probabilities $P_{n,k}(t)=\Pr\{N(t)=n \text{ and } K(t)=k\}$ satisfy the equation

$$\begin{aligned}P_{n,k}(t+\Delta t)&=(1-c_n\Delta t)P_{n,k}(t)+d_{n-1,k}\Delta t P_{n-1,k}(t)\\&\quad+(f_{n-1,k-1}+\beta_{n-1})\Delta t P_{n-1,k-1}(t)+o(\Delta t)\ , \qquad 1\leqq k\leqq n\leqq M\end{aligned}$$

and

$$P_{0,0}(t+\Delta t)=(1-\beta_0\Delta t)P_{0,0}(t)+o(\Delta t)\ ,$$

where $\alpha_n=\lambda n(M-n)$, $\beta_n=\mu(M-n)$, $c_n=\alpha_n+\beta_n$,

$$d_{n,k}=\lambda k(M-n)\ , \qquad f_{n,k}=\lambda(n-k)(M-n)\ .$$

From these equations, we get the following differential equations:

$$P'_{n,k}(t)=-c_nP_{n,k}(t)+d_{n-1,k}P_{n-1,k}(t) +(f_{n-1,k-1}+\beta_{n-1})P_{n-1,k-1}(t)\ , \qquad 1\leqq k\leqq n\leqq M \tag{2.5}$$

$$P'_{0,0}(t)=-\beta_0P_{0,0}(t)\ . \tag{2.6}$$

These differential equations, under the initial conditions $P_{0,0}(0)=1$ and $P_{n,k}(0)=0$ $(1\leqq k\leqq n\leqq M)$, satisfy the recursive formulae

$$P_{n,k}(t)=\exp(-c_nt)\int_0^t\exp(c_n\tau)\{d_{n-1,k}P_{n-1,k}(\tau) +(f_{n-1,k-1}+\beta_{n-1})P_{n-1,k-1}(\tau)\}d\tau\ , \qquad 1\leqq k\leqq n\leqq M \tag{2.7}$$

and

$$P_{0,0}(t)=\exp(-\beta_0t) \quad (\equiv P_0(t))\ . \tag{2.8}$$

Using these formulae, we can determine all $P_{n,k}(t)$ successively. But we are interested in the conditional probabilities of the random variable $K(t)$, given $N(t)=n$:

$$\Pr\{K(t)=k\,|\,N(t)=n\}\equiv p_{n,k}(t)\ .$$

These conditional probabilities have the remarkable property that they are independent of time t and size M of the group π. Concerning them, we have the following results.

[I] *Conditional probability*

The conditional probability

$$p_{n,k}=\Pr\{K(t)=k\,|\,N(t)=n\}$$

is independent of time t and size M of the group π, and satisfies the following recursive formula:

$$p_{n,k}=\frac{1}{n-1+\delta}\{kp_{n-1,k}+(n-k+\delta)p_{n-1,k-1}\}\ , \tag{2.9}$$

where $\delta=\mu/\lambda$, $p_{n,0}=0$ $(1\leqq n)$, $p_{n,k}=0$ $(n<k)$.

[II] *Conditional expectation*

The conditional expectation $E_n(K)=E\{K(t)\,|\,N(t)=n\}$ of the random variable $K(t)$, given $N(t)=n$, satisfies the following recursive formula:

$$E_n(K)=\frac{n-2+\delta}{n-1+\delta}E_{n-1}(K)+1\ . \tag{2.10}$$

Explicit form is given by

$$E_n(K)=\frac{n}{2}\frac{n-1+2\delta}{n-1+\delta} \qquad \text{for } n>1 \tag{2.11}$$

and

$$E_1(K)=1 .$$

[III] *Conditional variance*

The conditional variance $V_n(K)$ of $K(t)$, given $N(t)=n$, satisfies the following recursive formula

$$V_n(K)=\frac{n-3+\delta}{n-1+\delta}V_{n-1}(K) + \frac{(n+1)(n-2+\delta)\{(n-1)(n-2)+2(n-2)\delta+2\delta^2\}}{4(n-1+\delta)^2(n-2+\delta)^2}, \tag{2.12}$$

which is reduced to

$$V_n(K)=\frac{n(n-1)\{(n-1)(n-2)+4(n-2)\delta+6\delta^2\}}{12(n-1+\delta)^2(n-2+\delta)} . \tag{2.13}$$

[IV] Especially, when $\delta=0$ or $\mu=0$, namely if the information source S stops its function as soon as it once communicates I to the first person, we obtain the following results.

$$E_n(K)=\frac{n}{2} , \tag{2.10'}$$

$$V_n(K)=\frac{n}{12} . \tag{2.11'}$$

These are regarded on the other side as the asymptotic values of $E_n(K)$ and $V_n(K)$ when $n\to\infty$. As these results are independent of parameter λ and μ, they might be useful in testing the assumption of homogeneity of communication in the group π.

[V] *Normal approximation of the conditional distribution*

For sufficiently large n, conditional distribution $\{p_{n,k}\}$ of $K(t)$, given $N(t)=n$, can be asymptotically replaced by a normal distribution.

In fact, the conditional distribution of the standardized variable $(K-E_n(K))/\sqrt{V_n(K)}$, given $N(t)=n$, tends to the standard normal distribution $N(0, 1)$.

PROOF OF [I]. We shall first prove, by mathematical induction, that the conditional probabilities $p_{n,k}$ are independent of t. By equation (2.8), $p_{0,0}=P_{0,0}(t)/P_0(t)=1$ holds, and $p_{0,0}$ is independent of time t. Suppose

the conditional probabilities $p_{n-1,k}$ and $p_{n-1,k-1}$ are independent of time t. Then, substituting $P_{n-1,k}(t)=p_{n-1,k}P_{n-1}(t)$ and $P_{n-1,k-1}(t)=p_{n-1,k-1}P_{n-1}(t)$ into equation (2.7), we obtain

$$P_{n,k}(t)=\{d_{n-1,k}p_{n-1,k}+(f_{n-1,k-1}+\beta_{n-1})p_{n-1,k-1}\} \cdot \exp(-c_nt)\int_0^t \exp(c_n\tau)P_{n-1}(\tau)d\tau \ .$$

Comparing this result with equation (2.4), we can see that

$$p_{n,k}=\frac{P_{n.k}(t)}{P_n(t)}=\frac{1}{c_{n-1}}\{d_{n-1,k}p_{n-1,k}+(f_{n-1,k-1}+\beta_{n-1})p_{n-1,k-1}\} \ .$$

Therefore, $p_{n,k}$ is independent of time t, and this result leads to the following recursive formula:

$$p_{n,k}=\frac{1}{n-1+\delta}\{kp_{n-1,k}+(n-k+\delta)p_{n-1,k-1}\} \ ,$$

because $c_{n-1}=\alpha_{n-1}+\beta_{n-1}=(M-n+1)\{(n-1)\lambda+\mu\}$,

$$d_{n-1,k}=\lambda k(M-n+1), \quad \text{and} \quad f_{n-1,k-1}=\lambda(n-k)(M-n+1).$$

PROOF OF [II]. Multiplying both sides of equation (2.9) by k, and then summing up them over $k=1, 2, \cdots, n$, we obtain the equation (2.10).

By mathematical induction, equation (2.10) leads to equation (2.11).

PROOF OF [III]. As in the case of proving [II], recursive formula (2.12) of the conditional variance $V_{n-1}(K)$ is obtained. It leads to formula (2.13) by mathematical induction.

PROOF OF [V]. We shall first show that the p-th moment of $(K-E_n(K))/\sqrt{V_n(K)}$ tends to that of the standard normal distribution as $n\to\infty$ for every fixed positive integer p.

First consider the case where $\delta=0$. We have then $E_n(K)=n/2$, $V_n(K)=n/12$, and it is easily shown by induction that the distribution $\{p_{n,k}\}$ is symmetric in the sense that $p_{n,k}=p_{n,n-k}$. Hence

$$E[(K-E_n(K))^p]=0 \quad \text{for} \quad p=2r+1, \quad r=0, 1, 2, \cdots.$$

Using the formulae (2.9) and symmetric property of $\{p_{n,k}\}$, we can easily obtain

$$a_{n+1}^{(2r)}=\left(1-\frac{2r}{n}\right)a_n^{(2r)}+2\sum_{k=0}^{2r-3}\binom{2r}{k}\left(-\frac{1}{2}\right)^{2r-k}\frac{1}{n}a_n^{(k+1)} + \sum_{k=0}^{2r-2}\binom{2r}{k}\left(-\frac{1}{2}\right)^{2r-k}a_n^{(k)} \ ,$$

where

$$a_n^{(p)}=E_n[(K-E_n(K))^p] .$$

It suffices to prove that

$$a_n^{(2p)}=V_n(K)^p\cdot\alpha_{2p}(1+o(1)) ,$$

where α_{2p} is the $2p$-th moment of the standard normal distribution:

$$\alpha_{2p}=1\cdot 3\cdot 5\cdots(2p-3)(2p-1) .$$

This is clearly true for $p=1$. Now assume that this is true for $p\leqq r-1$. Then, it follows from the foregoing formula that

$$\begin{aligned}
a_{n+1}^{(2r)}&=\left(1-\frac{2r}{n}\right)a_n^{(2r)}+\frac{1}{4}\binom{2r}{2}a_n^{(2r-2)}+O(n^{r-2})\\
&=\left(1-\frac{2r}{n}\right)\left(1-\frac{2r}{n-1}\right)a_{n-1}^{(2r)}+\frac{1}{4}\binom{2r}{2}\left\{\left(1-\frac{2r}{n}\right)a_{n-1}^{(2r-2)}+a_n^{(2r-2)}\right\}\\
&\quad+O\{(n^{r-2})+(n-1)^{r-1}\}\\
&=\cdots\cdots\cdots\\
&=\left(1-\frac{2r}{n}\right)\left(1-\frac{2r}{n-1}\right)\cdots\left(1-\frac{2r}{1}\right)a_1^{(2r)}\\
&\quad+\frac{1}{4}\binom{2r}{2}\sum_{m=1}^{n-1}\left(1-\frac{2r}{n}\right)\left(1-\frac{2r}{n-1}\right)\cdots\left(1-\frac{2r}{m+1}\right)a_m^{(2r-2)}\\
&\quad+O\{n^{r-2}+(n-1)^{r-2}+\cdots+2^{r-2}+1^{r-1}\} .
\end{aligned}$$

The first term vanishes for $n\geqq 2r$, and the summands in the second term are

$$\begin{cases}
O(\sqrt{n}^{r-1}) & \text{for } m\leqq\sqrt{n} ,\\
\left(\frac{m}{n}\right)^{2r}a_m^{(2r-2)}\left(1+O\left(\frac{1}{m}\right)\right) & \text{for } m>\sqrt{n} .
\end{cases}$$

Thus we obtain

$$\begin{aligned}
\frac{a_{n+1}^{(2r)}}{\{(n+1)/12\}^r}&=\frac{1}{4}\binom{2r}{2}\frac{1}{n^{2r}}\left(\sum_{m>\sqrt{n}}m^{2r}\right)\cdot\frac{(n/12)^{r-1}}{\{(n+1)/12\}^r}\alpha_{2r-2}(1+o(1))\\
&\quad+O\left(\frac{n^{r-1}}{(n+1)^r}\right)\\
&=\frac{1}{4}\binom{2r}{2}\times 12\left[\int_0^1 x^{2r}dx\right]\alpha_{2r-2}+o(1)\\
&=3r(2r-1)\times\frac{1}{3r}\alpha_{2r-2}+o(1)\\
&=(2r-1)\alpha_{2r-2}+o(1)\\
&=\alpha_{2r}+o(1) .
\end{aligned}$$

Hence we have proved for every p

$$E_n\left[\left(\frac{K-E_n(K)}{\sqrt{V_n(K)}}\right)^p\right]\to\alpha_p \qquad (n\to\infty)\,.$$

Validity of this fact for $\delta\neq 0$ can be seen by comparing the p-th moments of the both cases $\delta=0$ and $\delta\neq 0$, or by a slight modification of the foregoing proof.

Now the probability distribution F_n of $\{K-E_n(K)\}/\sqrt{V_n(K)}$ has the finite moment of every order, which converges to that of the standard normal distribution. We can, therefore, conclude that the family of distribution $\{F_n\}$ is completely compact (see [4], p. 185), and that the limit distribution of any convergent subsequence has the same moments as the standard normal distribution. We know, however, that such a distribution must be identical with the normal distribution. Hence the family $\{F_n\}$ itself must converge to the normal distribution, for, otherwise, by compactness, $\{F_n\}$ must have another limit distribution having the same moments as the normal distribution, which is impossible. Thus the central limit convergence has been proved.

2.3. JOINT DISTRIBUTION OF $N(t)$ AND $L(t)$.

Now we shall investigate the number of persons who have received the information I directly from the information source S. This number may be considered to indicate the intensity of communication of S. Let $P^*_{n,l}(t)$ be the probability that $N(t)=n$ and $L(t)=l$, i.e., the probability that exactly n persons have the information I at time t and exactly l of them have received it from S. Then, in the same way as in section 2.2, we obtain the following formulae.

$$\begin{aligned}
&P^*_{n,l}(t+\Delta t)=(1-c_n\Delta t)P^*_{n,l}(t)+\alpha_{n-1}P^*_{n-1,l}(t)+\beta_{n-1}P^*_{n-1,l-1}(t)\Delta t+o(\Delta t) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{for}\quad 1\leqq l\leqq n, \\
&P^*_{n,n}(t+\Delta t)=(1-c_n\Delta t)P^*_{n,n}(t)+\beta_{n-1}P^*_{n-1,n-1}(t)\Delta t+o(\Delta t) \\
&P^*_{n,0}(t+\Delta t)=0 \qquad\qquad\qquad\qquad\qquad\qquad \text{for}\quad n\geqq 1, \\
&P^*_{0,0}(t+\Delta t)=(1-\beta_0\Delta t)P^*_{0,0}(t)+o(\Delta t)\,,
\end{aligned}$$

where α_i, β_i, c_i are the same quantities as defined in section 2.2.

Letting $\Delta t\to 0$, we obtain the following system of differential equations.

$$P^{*\prime}_{n,l}(t)=-c_nP^*_{n,l}(t)+\alpha_{n-1}P^*_{n-1,l}(t)+\beta_{n-1}P^*_{n-1,l-1}(t) \quad \text{for}\quad 1\leqq l\leqq n\leqq M\,, \tag{2.14}$$

$$P^{*\prime}_{0,0}(t)=-\beta_0 P^*_{0,0}(t)\ . \tag{2.15}$$

Here we have put $P^*_{n,l}(t)=0$ for $n<l$ for the sake of convenience.

Now integrating (2.15) under the initial condition $P^*_{0,0}(0)=1$, we have

$$P^*_{0,0}(t)=\exp(-\beta_0 t)\ . \tag{2.16}$$

Integrating (2.14) under the initial conditions $P^*_{n,l}(0)=0$ (for $l\geqq 1$), we obtain

$$P^*_{n,l}(t)=\exp(-c_n t)\int_0^t \exp(c_n\tau)\{\alpha_{n-1}P^*_{n-1,l}(\tau)+\beta_{n-1}P^*_{n-1,l-1}(\tau)\}\,d\tau$$
$$\text{for}\quad 1\leqq l\leqq n\leqq M\ . \tag{2.17}$$

Using (2.16) and (2.17) we can obtain inductively the explicit forms of $P^*_{n,l}(t)$, but we do not give them it here, for it is not our aim and we are interested in the conditional probability distribution of $L(t)$, given $N(t)=n$. It will be seen below that it depends on neither time t nor size M of group π.

We shall now investigate properties of conditional probability $\Pr\{L(t)=l \,|\, N(t)=n\}$.

[I] *Conditional probability*

The conditional probability $p^*_{n,l}=Pr\{L(t)=l \,|\, N(t)=n\}$ depends only on n, l and μ/λ, and contains neither time t nor size M of the group. It satisfies the following recursive formulae:

$$\begin{aligned} &p^*_{0,0}=1\ , \\ &p^*_{n,0}=0 \qquad\qquad \text{for}\quad n\geqq 1\ , \\ &p^*_{n,l}=\frac{n-1}{n-1+\delta}p^*_{n-1,l}+\frac{\delta}{n-1+\delta}p^*_{n-1,l-1}\ , \end{aligned} \tag{2.18}$$

where $\delta=\mu/\lambda$.

[II] *Conditional expectation*

The conditional expectation $E^*_n(L)=E\{L(t) \,|\, N(t)=n\}$, given $N(t)=n$, satisfies the following recursive formulae.

$$E^*_0(L)=0 \tag{2.19}$$

$$E^*_n(L)=E^*_{n-1}(L)+\frac{\delta}{n-1+\delta}\ . \tag{2.20}$$

The explicit form is

$$E^*_n(L)=\frac{\delta}{\delta}+\frac{\delta}{1+\delta}+\cdots+\frac{\delta}{n-1+\delta} \qquad \text{for}\quad n\geqq 1\ . \tag{2.21}$$

[III] *Conditional variance*

The conditional variance $V_n^*(L)$, given $N(t)=n$, is given by

$$V_n^*(L)=E_n^*(L)-\sum_{i=0}^{n-1}\left(\frac{\delta}{i+\delta}\right)^2. \tag{2.22}$$

[IV] *Normal approximation of the conditional distribution*

When n is large, the conditional distribution $\{p_{n,l}^*\}$ of $L(t)$, given $N(t)=n$, is asymptotically replaced by the normal distribution with mean $E_n^*(L)$ and variance $V_n^*(L)$.

In fact, the conditional distribution of the standardized variable $(L-E_n^*(L))/\sqrt{V_n^*(L)}$ tends to the standard normal distribution as n increases indefinitely.

PROOF OF [I]: We shall prove [I] by induction. First, notice that

$$P_0(t)=\exp(-\beta_0 t),$$

which, together with the formula (2.16), shows that

$$p_{0,0}^*=\frac{P_{0,0}^*(t)}{P_0(t)}=1.$$

It is clear that $p_{n,0}^*=0$ for $n\geqq 1$. Now take any integer n ($\geqq 1$) and assume $p_{n-1,l}^*$ to be time-free for $l=1, 2, \cdots, n-1$. Then, by virtue of (2.17), it follows that

$$\begin{aligned}P_{n,l}^*(t)&=\exp(-c_n t)\int_0^t \exp(c_n\tau)\{\alpha_{n-1}P_{n-1,l}^*(\tau)+\beta_{n-1}P_{n-1,l-1}^*(\tau)\}d\tau\\&=\exp(-c_n t)\int_0^t \exp(c_n\tau)\{\alpha_{n-1}p_{n-1,l}^*P_{n-1}(\tau)+\beta_{n-1}p_{n-1,l-1}^*P_{n-1}(\tau)\}d\tau\\&=(\alpha_{n-1}p_{n-1,l}^*+\beta_{n-1}p_{n-1,l-1}^*)\int_0^t \exp\{-c_n(t-\tau)\}P_{n-1}(\tau)d\tau.\end{aligned}$$

The recursive formula for $P_n(t)$ given in section 2.1 shows that the right-hand side is equal to

$$(\alpha_{n-1}p_{n-1,l}^*+\beta_{n-1}p_{n-1,l-1}^*)\frac{1}{c_{n-1}}P_n(t).$$

Hence we obtain

$$\begin{aligned}p_{n,l}^*=\frac{P_{n,l}^*(t)}{P_n(t)}&=\frac{\alpha_{n-1}}{c_{n-1}}p_{n-1,l}^*+\frac{\beta_{n-1}}{c_{n-1}}p_{n-1,l-1}^*\\&=\frac{(n-1)\lambda}{(n-1)\lambda+\mu}p_{n-1,l}^*+\frac{\mu}{(n-1)\lambda+\mu}p_{n-1,l-1}^*\\&=\frac{n-1}{n-1+\delta}p_{n-1,l}^*+\frac{\delta}{n-1+\delta}p_{n-1,l-1}^*.\end{aligned}$$

This shows that $p^*_{n,l}$ $(l=1, 2, \cdots, n)$, are also time-free and satisfy (2.18).

PROOF OF [II]: Multiply both sides of (2.18) by l and take the summation over $l=1, 2, \cdots, n$. Then, we can easily obtain (2.20) for $n \geqq 1$. (2.19) is obvious. The proof of (2.21) can now be performed by induction.

PROOF OF [III]: Multiply both sides of (2.18) by l^2 and take the summation over $l=1, 2, \cdots, n$. Then, by an easy calculation and by formula (2.21), we can obtain the recursive formula for $V^*_n(L)$ and we can prove (2.22) again by induction. Detailed calculations are omitted.

It is seen from the above results that the conditional standard deviation $\sqrt{V^*_n(L)}$ is less than the square-root of $E^*_n(L)$. This fact is useful for estimating the parameter δ by means of the observed values of N and L. We shall discuss this problem in the next section.

PROOF OF [IV]: Consider the infinite sequence of mutually independent random variables $X_1, X_2, \cdots, X_n, \cdots$, where X_n takes only two values 1 and 0, with probabilities $\delta/(n-1+\delta)$ and $(n-1)/(n-1+\delta)$, respectively. Let $S_n=X_1+X_2+\cdots+X_n$ be the partial sum of $X_1, X_2, \cdots X_n, \cdots$. Then clearly,

$$\Pr\{S_1=1\}=1\,,$$

$$\begin{aligned}\Pr\{S_n=l\} &= \Pr\{S_{n-1}=l\}\cdot\Pr\{X_n=0\}+\Pr\{S_{n-1}=l-1\}\cdot\Pr\{X_n=1\}\\ &= \frac{n-1}{n-1+\delta}\Pr\{S_{n-1}=l\}+\frac{\delta}{n-1+\delta}\Pr\{S_{n-1}=l-1\}\,.\end{aligned}$$

These formulae are entirely identical with the recursive formulae for $p^*_{n,l}$ given by (2.18). Hence we have $\Pr\{S_n=l\}=p^*_{n,l}$, and it suffices to show that the distribution of S_n is asymptotically normal for large n. Noticing that the variance $V(S_n)\sim\delta\cdot\log n$ for large n and that the summands $\{X_k\}$ are uniformly bounded, we can apply the central limit theorem in the bounded case (see [4] p. 277) to our problem, and the assertion follows.

§ 3. Estimation of μ/λ and test of homogeneity.

3.1. ESTIMATION OF THE PARAMETER $\delta=\mu/\lambda$.

The explicit forms of the conditional expectations and variances, given $N(t)=n$, which were obtained in the preceding section, make it possible for us to estimate the intensity ratio $\delta=\mu/\lambda$ based upon the sample values of N and L.

In section 1, we considered the pattern $U(t)$ of communication up to time t to be a random variable with parameter t, and $N(t)$, $K(t)$ and $L(t)$ to be statistics determined by this random variable. It might seem at first sight that the full knowledge of pattern $U(t)$ has an advantage over the knowledge of only $N(t)$ and $L(t)$ for estimation of δ. It, however, is not true, and we can show that the bi-variate statistic $(N(t), L(t))$ is sufficient for parameter (λ, μ). In fact, we can prove the following proposition.

The bi-variate statistic $(N(t), L(t))$ *is a sufficient statistic for the family of distributions* $\Pr(U(t)\,;\lambda, \mu)$ *with the parameter* (λ, μ).

PROOF: Since size M of the group is finite, the number of all possible patterns is finite. In particular, let $\{u_1, u_2, \cdots, u_w\}$ be the set of all possible patterns such that the number of informed persons is n and the number of persons who have received information I directly from the information source S is l. Fig. 2 illustrates the tree of a pattern in the case $n=10$ and $l=3$. The knots of the tree are indicated by $a_1, a_2, \cdots, a_n$. Let $s_1, s_2, \cdots, s_q$ be all possible orderings of $a_1, a_2, \cdots, a_n$, in which the information is carried over. Any s_i is represented by a permutation of $a_1, a_2, \cdots, a_n$, which corresponds to a process of branching of the tree. Then, it follows immediately that

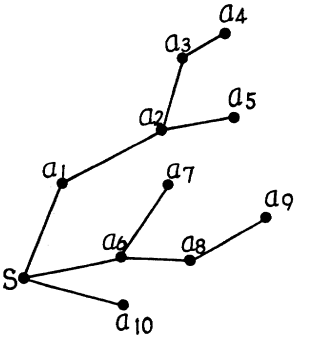

Fig. 2.

$$\Pr\{U(t)=u_i\}$$

$$= \sum_{s_1,\cdots,s_q} \int_0^t \exp(-c_0 t_1)\gamma_0 dt_1 \int_{t_1}^t \exp\{-c_1(t_2-t_1)\}\gamma_1 dt_2 \cdots$$

$$\cdot \int_{t_{n-1}}^t \exp\{-c_{n-1}(t_n-t_{n-1})\}\gamma_{n-1} \exp\{-c_n(t-t_n)\}dt_n$$

$$= \sum_{s_1,\cdots,s_q} \gamma_0\gamma_1\cdots\gamma_{n-1} \exp(-c_n t) \int_0^t dt_1 \int_{t_1}^t dt_2 \int_{t_2}^t dt_3 \cdots$$

$$\cdot \int_{t_{n-1}}^t dt_n [\exp\{(c_1-c_0)t_1\} \exp\{(c_2-c_1)t_2\} \cdots \exp\{(c_n-c_{n-1})t_n\}]\,,$$

where γ_j is equal to β_j or α_j according as the j-th knot (in the order of being produced) is directly connected with S or not, and α_j, β_j, c_j are those quantities defined in Section 2. Now, since the number of knots which are directly connected with S is exactly l, we have for any

s in $s_1, \cdots, s_q$,

$$\gamma_0\gamma_1\cdots\gamma_{n-1}=M(M-1)\cdots(M-n+1)\lambda^{n-l}\mu^l c(s) ,$$

where $c(s)$ is determined only by s and is independent of λ, μ, M and t.

Hence, putting

$$\sum_{s_1,\cdots,s_q} c(s)=k(u_i) ,$$

we obtain

$$\frac{\Pr\{U(t)=u_i\}}{\Pr\{U(t)=u_j\}}=\frac{k(u_i)}{k(u_j)} \tag{3.1}$$

for any pair of patterns u_i, u_j in $\{u_1, u_2, \cdots, u_w\}$, and we finally obtain

$$\Pr\{U(t)=u_i \,|\, N(t)=n, L(t)=l\}=\frac{k(u_i)}{\sum_{j=1}^{w} k(u_j)} .$$

The right-hand side is clearly independent of (λ, μ). This assures us that the conditional probability of the pattern u_i, given $N=n$ and $L=l$, is independent of (λ, μ). This proves the sufficiency of statistic $(N(t), L(t))$.

Now we shall proceed to estimate the parameter δ by means of the observed values of N and L.

If n is large[1], upon applying normal approximation (section 2, 3, [IV]) of the conditional probability of $L(t)$, given $N(t)=n$, we have, for any α such that $0<\alpha<1$,

$$\Pr\{|L-E_n^*(L)|<c\sqrt{V_n^*(L)} \,|\, N(t)=n\}>1-\alpha , \tag{3.2}$$

where $c=c(\alpha)$ is determined by

$$\int_{-c}^{c}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx=1-\alpha ,$$

which is found from the table of standard normal distribution.

As the quantities $E_n^*(L)$ and $V_n^*(L)$ are regarded as functions of the integral-valued variable n, we shall obtain random variables $E_{N(t)}^*(L(t))$

---

[1] If n is not so large, normal appoximation can not be used, and we must content ourselves with less sharp evaluation. As is seen from the proof of section 2.3 [IV], probability distribution $\{p_{n,l}^*\}_{l=1,\ldots,n}$ can be regarded as the distribution of the sum of n independent random variables. Hence Bernstein's inequality (see [5]) can be used, and we have

$$\Pr\{|L-E_n^*(L)|\leqq\lambda\}\geqq 1-2\exp\left(\frac{-\lambda^2}{2V_n(L)+\frac{2}{3}\lambda}\right).$$

We have only to determine λ as a function of $V_n(L)$ so that the right-hand side may be equal to $1-\alpha$.

and $V^*_{N(t)}(L(t))$ by substituting the random variable $N(t)$ for n. It now follows from (3.2) that

$$\begin{aligned}\Pr\{|L(t)-E^*_{N(t)}(L(t))|<c\sqrt{V^*_{N(t)}(L(t))}\} &= \sum_n P_n(t)\cdot\Pr\{|L-E^*_n(L)|<c\sqrt{V^*_n(L)}\,|\,N(t)=n\} \\ &>(1-\alpha)\sum_n P_n(t)=1-\alpha\,.\end{aligned}$$

This means that the probability that two-dimensional random variable $(N(t), L(t))$ will satisfy

$$|L-E^*_N(L)|<c\sqrt{V^*_N(L)} \tag{3.3}$$

is greater than $1-\alpha$. Hence, solving (3.3) in δ, we should have an inequality of the type

$$f(N, L)<\delta<g(N, L)\,,$$

and obtain a confidence interval for δ with confidence coefficient greater than $1-\alpha$.

The foregoing method is easy to deal with owing to its time-free property. We may perform the actual survey at any moment,[1] and estimate δ by means of the observed values of N and L, regardles of how much time has passed since communication began.

As δ is contained in $E^*_n(L)$ and $V^*_n(L)$ in a rather complicated form, it is not easy to solve (3.3) explicitly and to obtain confidence limits for δ for the observed values n and l. When n is not so large, we may draw the graph of $f_n(\delta)=E^*_n(L)\pm c\sqrt{V^*_n(L)}$ by calculating the numerical values of $E^*_n(L)$ and $V^*_n(L)$ as functions of δ, and may find back from the graph the confidence limits for δ for observed values n, l. This method, however, is hardly possible for larger values of n, and we must find another expedient method. Since $E^*_n(L)$ and $V^*_n(L)$ are finite sums of series, we can apply Euler's summation formula to them and rewrite them in more convenient forms. Thus we can obtain their approximate expressions for sufficiently large n. In fact, it follows by some calculations that

$$E^*_n(L)\sim 1+\delta\log\frac{n-1+\delta}{1+\delta}\,,$$

$$V^*_n(L)\sim\log\frac{n-1+\delta}{1+\delta}-\frac{\delta}{1+\delta}+\frac{\delta}{n-1+\delta}\,.$$

[1] See the remark at the end of this section.

With the aid of these expressions the upper and lower confidence limits for δ are given as follows:

i) If the observed value l of $L(t)$ is far smaller than $\log n$, n being the observed value of $N(t)$, the confidence limits are given by

$$x+\frac{c^2}{2\log n}\pm\frac{c\sqrt{4\log n+c^2}}{2\log n},$$

where $x=(l-1)/\log n$.

ii) If the magnitude of l is not so different from that of $\log n$, the confidence limits are given by

$$x\pm c\sqrt{\frac{x}{\log n}}.$$

iii) If l is far larger than $\log n$ and slightly less than n, the confidence limits are given by

$$y+\frac{c^2}{2n}\pm\frac{c\sqrt{4ny+c^2}}{2n},$$

where $y=l/n$.

iv) If $l\sim n$, the confidence limits are given by

$$\frac{2nz+c^2\pm c\sqrt{4nz+c^2}}{4z^2},$$

where $z=n-l$.

REMARK: In the above-mentioned method of estimation the time for observation was arbitrary, but was fixed prior to the experiment. We give here a remark on the extension of it to the case where the time for observation is a random variable τ. Since $N(t)$ is a Markov process, we can consider a Markov time τ with respect to this process, in the wider sense that for any $t>0$ the event $\{\tau\leqq t\}$ belongs to the minimal σ-field containing both $\mathfrak{B}_N(t)$ and $\mathfrak{B}_0$, where $\mathfrak{B}_N(t)$ is the σ-field of events determined by $\{N(s), s\leqq t\}$ and $\mathfrak{B}_0$ is a σ-field independent of our basic process $U(t)$. Then the conditional probability distribution of $L(\tau)$, given $N(\tau)=n$, still depends only on n, and the above-mentioned method of estimation apply. Examples of such a Markov time τ are: (1) minimum of t such that $N(t)=n_0$ for some prescribed number n_0, (2) random time determined quite independently of the process, (3) mixed type of (1) and (2).

3.2. TEST OF HOMOGENUITY OF COMMUNICATION IN THE GROUP π.

The results obtained in section 2.2 serves for testing hypothesis H that the communication of information I in the group π is homogeneous both in time and in persons.

Suppose that the information source S stops its function as soon as it once communicates the information to the first person. Then, according to section 2.2 [IV], if the hypothesis H_0 be true, the conditional expectation $E_n(K)$ and the conditional variance $V_n(K)$ are given by

$$E_n(K)=\frac{n}{2},$$

$$V_n(K)=\frac{n}{12}.$$

They contain neither t nor λ, so, in the same way as in section 3.1, we have only to observe N and K without considering the time when the observation is performed.

For large n[1] we can use the normal approximation (cf. [V] in section 2.2), and

$$\Pr\{|K(t)-E_n(K)|>c\sqrt{V_n(K)}\,|N(t)=n\}\leqq\alpha,$$

which is equivalent to

$$\Pr\left\{K^2-nK+\frac{1}{4}n^2-\frac{c^2}{12}n>0\,\middle|\,N(t)=n\right\}\leqq\frac{1}{c^2}.$$

Where c is determined by α in the same way as in section 3.1. It follows that

$$\begin{aligned}&\Pr\left\{K^2-NK+\frac{N^2}{4}-\frac{c^2}{12}N>0\right\}\\&=\sum_{n=0}^{M}P_n(t)\cdot\Pr\left\{K^2-nK+\frac{1}{4}n^2-\frac{c^2}{12}n>0\,\middle|\,N(t)=n\right\}\\&\leqq\alpha\sum_{n=0}^{M}P_n(t)=\alpha.\end{aligned}$$

Hence, if we determine the critical region w by the outside of the

1) If n is not so large, normal approximation is not valid. In this case, Meidell's inequality for unimodal distribution (see [6]) can be used and we have

$$\Pr\{|K(t)-E_n(K)|>c\sqrt{V_n(K)}\,|N(t)=n\}\leqq\frac{4}{9}\frac{1}{c^2},$$

which is slightly sharper than ordinary Tchebycheff's inequality.

parabolla (see Fig. 3)

$$K^2-NK+\frac{1}{4}N^2-\frac{c^2}{12}N>0\,,$$

we have a test of the hypothesis H_0 with significance level α.

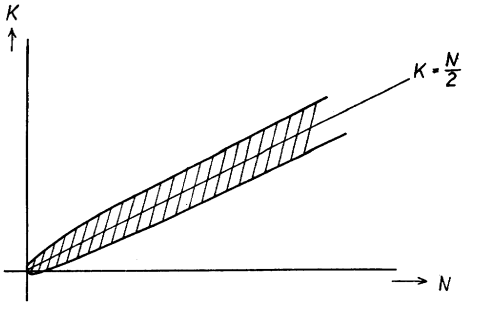

Fig. 3.

REMARK: Similarly to the remark in section 3.1, we can extend the results to the case where the time for observation is a Markov time. The method of testing hypothesis H_0 is quite identical with that for a fixed time.

§ 4. Conclusion

The results obtained in the preceding sections, based on our stochastic model mentioned above, can be used in the following manner.

1) We take a social group π and test the hypothesis that the communication in π is homogenous in time and in persons, by the method stated in section 3.2.

2) Suppose that the same information I is communicated from each of various information sources S_i's to each of corresponding groups π_i's, in which it is assummable that communication is homogeneous and corresponding intensities λ_i's are all equal (to some λ). Then, observing the values of $N(t)$ and $K(t)$ at time t, we can estimate the ratios $\delta_i=\mu_i/\lambda$ of the intensities of communication. These ratios will make it possible to compare the intensities of information sources S_i's, and suggest the optimal of several media of communication of information I.

3) We are able to evaluate constructively the stochastic behavior of pattern $U(t)$, based on the knowledge of the ratio δ of intensities λ, μ and homogeneity of communication in the group π. But it must be noted that the intensities λ, μ depend on the group π, information I and source S, so we need to know them prior to our experiment.

In this paper we have considered the simplest stochastic model for communication in a social group. The assumption of homogeneity has been essential in our argument, but it does not seem to hold in every actual problem. One of simple modifications is to assume that the social group consists of several number of subgroups, in each of which homo-

geneity is assumed. Unfortunately, however, even such a simple modification does not keep the time-free property of the conditional probabilities and makes it impossible to treat the problems in such an easy way as is mentioned above. This is the reason why we have restricted ourselves to the simplest case. It seems that there remains much to be developed in future.

THE INSTITUTE OF STATISTICAL MATHEMATICS

REFERENECES

[1] The Research Group of Mass Communication channel, *Research Report General Series No. 1, Inst. Stat. Math.*

[2] Taga, Y. and Suzuki, T. On the estimation of average length of chains in the communication-pattern, *Ann. Inst. Stat. Math.*, Vol. 9, No. 3 (1957).

[3] Feller, W. *An Introduction to Probability Theory and its Applications*, 1, New York.

[4] Loève, M. *Probability Theory*, New York.

[5] Bernstein, S. Sur une modification de l'inégalité de Tchebichef *Ann. Sc. Instit. Sov. Ukraine*, Sect. Math. 1 (1924). (Russian, French summary).

[6] Meidell, B. Sur la probabilité des erreurs, *Comptes Rendus*, Paris 176 (1923), 280-2.