PROBABILITY DISTRIBUTIONS OF RANDOM VARIABLES ASSOCIATED WITH A STRUCTURE OF THE SAMPLE SPACE OF SOCIOMETRIC INVESTIGATIONS[1]

By Leo Katz and James H. Powell

Michigan State University and Western Michigan College

1. Summary. In this paper, we consider a disjoint decomposition, at three levels, of the total sample space for n-person, one-dimensional sociometric investigations. This results in a structure particularly suited to determination of the probability distributions of a large class of sociometric variables. Systematic methods for obtaining these distributions are presented and illustrated by two examples; while the first is trivial, the second produces a previously unknown result.

It should be remarked that the methods developed here have application in the theory of communication networks and, indeed, in the study of any network situations which may be represented by either of the two models employed in the paper.

2. Introduction. The simplest model for the organization of a group of individuals is one-dimensional, in the sense that organization for only one activity of the group is considered. Connections between *ordered* pairs of individuals are represented by non-reflexive binary relations. Although a binary model appears superficially to be too barren to show adequately the richness of variability of the response of one individual to another, it is by no means trivial and is precisely the model used in most sociometric investigations, where the relations are lines of communication, authority, liking, etc.

In this model, a particular organization of n individuals has two isomorphic representations, both of which have been used extensively in the literature for descriptive purposes. The older of the two is the linear directed graph on n points, $P_1, P_2, \cdots, P_n$. A connection from man i to man j is represented by a directed line from P_i to P_j, $P_i \rightarrow P_j$; the absence of such a connection, by no line from P_i to P_j. The equivalent matrix representation is an $n \times n$ matrix, $C = (c_{ij})$, where $c_{ij} = 1$ if a connection exists from man i to man j, and $c_{ij} = 0$, otherwise. By convention, $c_{ii} = 0$. Obviously, $c_{ij} = 1$ (or 0) if and only if a directed line exists (or doesn't) from P_i to P_j. Hence, the two representations are isomorphic.

To fix the notation, let $r_i = \sum_j c_{ij}$ be the ith row total of C and $s_j = \sum_i c_{ij}$ be the jth column total. In the graph, r_i is the number of lines issuing from the point P_i, and s_j is the number of lines terminating on the point P_j. Moreover, $\sum_i r_i = \sum_j s_j = t$, the total number of directed lines. Finally, let the vectors

Received May 24, 1956; revised October 18, 1956.

[1] This work was supported by the Office of Naval Research under contract NR 170-115

r and s, with elements r_i and s_j, respectively, be the two n-part, non-negative, ordered partitions of t which represent respectively, the marginal row and column totals of C.

Unless otherwise noted in the sequel, all graphs will be on n points and linearly directed (n-graphs), and all matrices will be $n \times n$ hollow matrices of 1's and 0's. (A matrix is *hollow* if all principal diagonal elements vanish.)

3. Decomposition of the sample space. The sample space of the possible organizations of an n-member group is the space of all possible n-graphs or $n \times n$ hollow matrices of 1's and 0's. In this section, we consider a decomposition of the total sample space, Ω, following lines which hold promise of utility for certain investigations. We define first-order disjoint subspaces, Ω_t, $t = 0, 1, \cdots, [n(n-1)]$, as the collections of n-graphs containing exactly t lines. Obviously,

$$\Omega = \bigcup_{t=0}^{n(n-1)} \Omega_t, \tag{1}$$

since the Ω_t are mutually exclusive and exhaustive.

Continuing in the same vein, we define second-order subspaces, $\omega(\rho)$, $\rho = (r_1, r_2, \cdots, r_n)$, as the collections of graphs with r_i lines emanating from P_i, $i = 1, 2, \cdots, n$. Since $\sum_i r_i = t$, we have

$$\Omega_t = \bigcup_{(\rho)_t} \omega(\rho), \tag{2a}$$

where $(\rho)_t$ is a generic symbol for non-negative, integral, ordered, n-part partitions of t with all $r_i < n$. In a completely dual manner, we might define alternative second-order subspaces, $\omega(\sigma)$ in terms of n-graphs with s_j lines converging on P_j. In this case, we would have

$$\Omega_t = \bigcup_{(\sigma)_t} \omega(\sigma). \tag{2b}$$

Third-order subspaces are defined by $\omega(\rho, \sigma) \equiv \omega(\rho) \cap \omega(\sigma)$, and are identified with spaces of n-graphs with r_i lines emanating from, *and* s_i lines converging on, P_i. Once again, these sets are exclusive and exhaustive in the sense that

$$\omega(\rho) = \bigcup_{(\sigma)_t} \omega(\rho, \sigma), \tag{3a}$$

and

$$\omega(\sigma) = \bigcup_{(\rho)_t} \omega(\rho, \sigma). \tag{3b}$$

We remark that double and triple disjoint decompositions of the larger spaces may also be indicated.

It will be obvious to the reader that there exist isomorphisms among certain of these second and third-order subspaces. It will be less obvious, but important for computations, that these isomorphisms involve *simultaneous* permutations on the elements of the two vectors ρ and σ. We shall not elaborate on this point since it contributes little to the notions with which we are here concerned.

4. Random variables associated with the structure of the sample space. The decomposition described in the previous section imposes a structure on the sample

space. In most sociometric investigations, involving randomness in the existence of connections between ordered pairs of individuals, it has been deemed appropriate to assign uniform probability to each of the points in a third-order subspace, at least. In more extreme cases (the vast majority) it is customary to assume that every possible sample point is equally likely. Sometimes this has been done without even specifying which sample points are possible under the conditions of the experiment.

In the context of the particular experiment, it is usually possible for the experimenter to determine that his universe of discourse consists of Ω or one of the smaller subsets we have described. If, *also*, it happens that the random variable under discussion assumes the same value over all the points of each of certain smaller subspaces, the assumption of uniformity of probability within these subspaces will produce the complete probability distribution of the variable. In this section, we investigate these circumstances.

We say that a random variable defined over n-graphs or $n \times n$ hollow matrices is *associated* with the sample space structure of the previous section if the value of the variable is constant over all points in every $\omega(\rho, \sigma)$ contained in the domain of definition of the variable. Every such variable has a probability distribution which is completely specified as soon as we are able to count the numbers of points in the appropriate subspaces, assuming uniformity of probability on each point. In the next section, we shall present methods for carrying out this enumeration. A variable associated with the structure in the sense of the present definition is necessarily one whose value is somehow determined by, i.e., is a function of, the r_i and s_j, $i, j = 1, 2, \cdots, n$, alone. Indeed, this may be taken as an alternative definition.

To establish that the class of variables associated with the structure has some real substance, we examine a few variables which have been the subjects of sociometric investigations. Gross expansiveness, or average level of expansiveness, has been defined in terms of t alone in the context of the space Ω. Variability in expansiveness is defined as a function (usually a sum of squares) of the r_i, $i = 1, 2, \cdots, n$, sometimes in the context of Ω and sometimes in Ω_t. A number of variables have been defined as functions of the s_j, $j = 1, 2, \cdots, n$, in various contexts ranging down to $\omega(\rho)$. Examples are (1) the number of isolates, i.e., the number of $s_j = 0$ and (2) the choice status of the most highly chosen, i.e., $\max_j s_j$. Both of these are usually studied in the context of some $\omega(\rho)$.

5. Enumeration of the points in various subspaces. In considering the problems of enumeration, it will be more convenient to use the matrix representation because of its more flexible notation. Thus, the total number of matrices (graphs) in Ω is the number of ways in which the $n(n-1)$ elements of C may be specified as either zero or one. By elementary considerations, the number of distinct ways this can be done is

$$\eta = 2^{n(n-1)} \tag{4}$$

The matrices in Ω_t have t ones distributed over $n(n-1)$ positions; the number of ways this can be accomplished is the number of ways of specifying a particular t of the $n(n-1)$ positions. Therefore, the number of matrices (graphs) in Ω_t is given by

$$\eta_t = \binom{n(n-1)}{t}. \tag{5}$$

where $\binom{a}{b}$, $b \leqq a$, is the binomial coefficient $a!/[b!(a-b)!]$. As is well-known,

$$\sum_t \eta_t = \sum_t \binom{n(n-1)}{t} = 2^{n(n-1)} = \eta.$$

The enumeration of matrices in $\omega(\rho)$ is accomplished by considering, for each i, r_i ones distributed over $(n-1)$ positions. This can be done, independently, for each i, in $\binom{n-1}{r_i}$ ways and thus the total number of matrices (graphs) in $\omega(\rho)$ is given by

$$\eta(\rho) = \prod_1^n \binom{n-1}{r_i}. \tag{6a}$$

By a similar argument, the number of matrices in $\omega(\sigma)$ is given by

$$\eta(\sigma) = \prod_1^n \binom{n-1}{s_j}. \tag{6b}$$

It is easily seen that

$$\sum_{(\rho)_t} \eta(\rho) = \sum_{(\sigma)_t} \eta(\sigma) = \eta_t .$$

The only difficult counting problem arises when we attempt to compute the number of points in $\omega(\rho, \sigma)$. This problem was solved by the authors [3] who showed that this number is given by

THEOREM.

$$\eta(\rho, \sigma) = A\left\{\left[\prod_{i=1}^n (1 + \delta_i)^{-1}\right](\rho, \sigma)\right\}.$$

where the δ_i are operators on the pair of vectors defined by $\delta_i(r_1, \cdots, r_i, \cdots, r_n ; s_1, \cdots, s_i, \cdots, s_n) = (r_1, \cdots, r_i - 1, \cdots, r_n ; s_1, \cdots, s_i - 1, \cdots, s_n)$, *the symbol* $A\{\sum a_\alpha(\rho_\alpha, \sigma_\alpha)\}$ *stands for* $\sum a_\alpha A(\rho_\alpha, \sigma_\alpha)$ *and* $A(\rho_\alpha, \sigma_\alpha)$ *is the coefficient of the monomial symmetric function of order corresponding to* σ_α *in the expansion of the unitary (elementary) symmetric function of order corresponding to* ρ_α.

We note that the coefficients $A(\rho_\alpha, \sigma_\alpha)$ are given in tables of David and Kendall [1] for ρ_α and σ_α partitions of t up to $t = 12$. P. V. Sukhatme [5] gave an algorithm for computing $A(\rho_\alpha, \sigma_\alpha)$ for any weight and showed that $A(\rho, \sigma)$ is the number of matrices of elements $c_{ij} = 0$ or 1 with fixed row totals r_i and

column totals s_j but *without* restrictions on the diagonal elements. We present a very much abbreviated alternative to the proof previously given by the authors in the paper cited above.[2]

PROOF. $G_n \equiv \prod_{i,j=1}^{n} (1 + x_i y_j)$ generates the $A(\rho_\alpha, \sigma_\alpha)$ as coefficients of terms $\prod_{i=1}^{n} x_i^{r_{i\alpha}} \prod_{j=1}^{n} y_j^{s_{j\alpha}}$, and we may write

$$G_n = \sum_{(\alpha)} A(\rho_\alpha, \sigma_\alpha) \prod_i x_j^{r_{i\alpha}} \prod_j y_j^{s_{j\alpha}},$$

where the sum extends over all α such that

$$0 \leqq r_{i\alpha} \leqq n - 1, \qquad 0 \leqq s_{j\alpha} \leqq n - 1, \qquad \sum_i r_{i\alpha} = \sum_j s_{j\alpha}.$$

This is most easily seen if each c_{ij} in a matrix C of 0's and 1's is represented as $(x_i y_j)^{c_{ij}}$. Then, each term in the formal expansion of G_n represents one complete configuration of all the c_{ij}, simultaneously. Finally, in each individual term, the total exponent of $x_i(y_j)$ is the sum $\sum_j c_{ij} = r_i (\sum_i c_{ij} = s_j)$, and the coefficient $A(\rho_\alpha, \sigma_\alpha)$ is the number of distinct configurations of the c_{ij} with the indicated row and column totals.

Minor modification of the same reasoning serves to establish that

$$H_n \equiv \prod_{\substack{i,j=1 \\ i \neq j}}^{n} (1 + x_i y_j)$$

is a generating function for the $\eta(\rho_\alpha, \sigma_\alpha)$.

Next, we observe that

$$H_n = \left[\prod_{i=1}^{n} (1 + x_i y_i)^{-1} \right] G_n.$$

In this equation, the coefficient of $(\prod_i x_i^{r_i} \prod_j y_j^{s_j})$ in the left-hand member is $\eta(\rho, \sigma)$ and, in the right-hand member, is $\sum_{\alpha_1} \sum_{\alpha_2} \cdots \sum_{\alpha_n} (-)^{\sum_i \alpha_i} A(\rho - \alpha, \sigma - \alpha)$, where the α_i range over all non-negative integers. Equating these coefficients gives the expanded form of the statement of the theorem.

We note that the last sum in the proof above may be written in finite terms, since, as soon as any $\alpha_i > \min(r_i, s_i)$, the corresponding $A(\rho - \alpha, \sigma - \alpha) \equiv 0$, by the definition of Sukhatme as a number of certain matrices of 0's and 1's.

6. Probability distributions of associated random variables. It is now clear that we have laid down a program for computing, exactly, the probability distributions for any and all random variables associated with this structure of the sample space. In particular instances, it may be possible to effect certain economies in the computations by exploiting the isomorphisms among subsets so as to avoid duplication.

When the variable in question has constant values on sub-spaces no larger than an $\omega(\rho, \sigma)$, the computations are always formidable, though never impossible. In such circumstances, it would seem desirable to develop approximate

[2] This alternative proof follows lines of a suggestion by J. S. Frame

distributions for these variables, treating the exact methods as procedures for testing the validity of the approximations over the ranges of group size, etc., to be covered. For *very* small groups, it will usually be feasible to carry out the exact computations.

7. Examples. We shall give two examples of random variables associated with the sample space structure. In each, we consider the null case in which each graph in the appropriate sample space is equally likely, i.e., a uniform probability distribution over the sample space.

EXAMPLE 1. One measure of gross expansiveness, equal to the total number of choices made by group divided by size of group, is given by Loomis and Proctor [4] in a contribution to *Research Methods in Social Relations*. In our notation, this index is $E = t/n$.

The distribution problem, in the null case, is easily solved. Clearly, the appropriate sample space is Ω and our random variable, the number of distinct n-graphs with t lines, is constant over the first-order subspaces, Ω_t, in the disjoint and exhaustive decomposition of Ω. Thus, our random variable is associated with the sample space structure and according to Section 4 and the enumeration formulas of Section 5, the required probabilities are given by

$$P(t = k) = \frac{\eta_k}{\eta} = \frac{\binom{n(n-1)}{k}}{2^{n(n-1)}}. \tag{7}$$

EXAMPLE 2. An *isolate* is an individual represented in the graph by a point, P_1, with no terminating lines and in the matrix by a column of zeros, i.e., $s_i \parallel 0$ in the vector σ. The exact probability distribution of the number of isolates for the case $r_i = d(i = 1, 2, \cdots, n)$ was obtained from first principles by Katz [2], in 1950.

Using the methods already developed, we can now easily extend this result to the general case where the ith individual has r_i outgoing connections, the r_i being not necessarily equal.

The most common setting for this problem is in the sample space $\omega(\rho)$. In the null case, we desire the number of n-graphs having a specified number of points with no terminating lines, i.e., a specified number of zeros in the vector σ. Our random variable, X, the number of zero s_j's, is constant over the third order subspaces $\omega(\rho, \sigma)$ in the decomposition of $\omega(\rho)$; thus, it is associated with the sample space structure. Hence, according to Section 4 and the enumeration formulas of Section 5, the probability of exactly k isolates is given by

$$\begin{aligned} P(X = k \mid \rho) &= \frac{\sum_{(\sigma)_t} I_{A_k}(\sigma)\eta(\rho, \sigma)}{\eta(r)} \\ &= \frac{\sum_{(\sigma)_t} I_{A_k}(\sigma) A\left\{\prod_1^n (1 + \delta_i)^{-1}(\rho, \sigma)\right\}}{\prod_1^n \binom{n-1}{r_i}}, \end{aligned} \tag{8}$$

where A_k is the union of $\omega(\rho, \sigma)$ such that the vectors σ have exactly k vanishing components, and I_A is the indicator function for the set A.

We remark that in some contexts the appropriate sample space might be the larger space Ω_t. However, our enumeration methods will still give us the required probabilities necessary to construct the distribution. In this case, the probability of exactly k isolates is given by

$$P(X = k \mid t) = \frac{\sum_{(\rho)_t} \sum_{(\sigma)_t} I_{A_k}(\sigma)\eta(\rho, \sigma)}{\eta_t}$$

$$= \frac{\sum_{(\rho)_t} \sum_{(\sigma)_t} I_{A_k}(\sigma) A\left\{\prod_1^n (1 + \delta_i)^{-1}(\rho, \sigma)\right\}}{\binom{n(n-1)}{t}}, \tag{9}$$

where the notations are the same as before.

Thus, the probability distribution can be constructed for any index (proposed for the study of group structure) which depends only on the number of isolates in the group. Another such index, equal to the reciprocal of the number of isolates, is given by Loomis and Proctor [4] as a measure of "group integration."

Finally, we note that neither of the distributions (8) and (9) have been given correctly in the literature.

REFERENCES

[1] F. David and M. G. Kendall, "Tables of symmetric functions. Parts II and III," *Biometrika*, Vol. 38 (1951), pp. 435–462.

[2] L. Katz, "The distribution of the number of isolates in a social group," *Ann. Math. Stat.*, Vol. 23 (1951), pp. 271–276.

[3] L. Katz and J. Powell, "The number of locally restricted directed graphs," *Proc. Amer. Math. Soc.*, Vol. 5 (1954), pp. 621–626.

[4] C. Loomis and C. Proctor, "Analysis of Sociometric Data," *Research Methods in Social Relations*, Part 2, Chap. 17, The Dryden Press, New York, 1951.

[5] P. V. Sukhatme, "On bipartitional functions," *Philos. Trans. Roy. Soc. London, Ser. A*, Vol. 237 (1938), pp. 375–409.