THE DISTRIBUTION OF THE NUMBER OF ISOLATES IN A GROUP.[1]

by

Leo Katz

Institute of Statistics

University of North Carolina

and

Michigan State College

July, 1950

Institute of Statistics
Mimeo. Series 36
For Limited Distribution

1. Work done under the sponsorship of the Office of Naval Research, Project NR 042 031 at Chapel Hill, North Carolina.

Introduction. Consider a group consisting of N individuals. Each designates d of the others with whom he should prefer to be associated in some specified activity, i.e., each chooses d from N - 1 possible associates. In the context of the group and the specified activity, an individual is said to be an isolate if he is chosen by none of his fellow group members. It is immediately obvious that the number of isolates depends upon the size of the group, the number of choices permitted and the extent to which the group, as a social organism, provides acceptance for joint activities for the individuals who compose the group. Thus, when N and d are fixed, the number of isolates becomes an important characteristic of the group structure. When it is important to state whether the number of isolates is unusually large or small, it is necessary that the chance distribution of this number be known.

The history of attacks on the distribution problem is brief. Lazarsfeld, in a contribution to a paper by Moreno and Jennings [6][1], gave the expected (mean) number of isolates as

$$N\left[(N - d - 1)/(N - 1)\right]^{N-1},$$

but made no attempt to obtain the distribution. Bronfenbrenner [1] gave (without proof) an incorrect version of the distribution function. He gave the expression, which he claimed was "---developed deductively and checked by empirical methods ---,"

$$(1) \qquad F(i) = \Pr\{\text{i or less isolates}\} = 1 - \frac{(N - i - 2)^{(d)}}{(N - 1)^{(d)}},$$

where $a^{(b)} = a(a - 1)(a - 2) \ldots (a - b + 1)$. This form is not only incorrect; it gives completely nonsensical results in application. Edwards [2] conjectured that the Bronfenbrenner formula gives the probability

1. Numbers in square brackets refer to References at end of paper.

of a given person's including in his list of d at least one of $i + 1$ specified names. Edwards then gave correctly the probability of the maximum possible number of isolates,

$$(2) \qquad f(N-1-d) = \Pr\left\{N - 1 - d \text{ isolates}\right\} = \binom{N}{N-1-d} \frac{(d+1)^{(N-1-d)}}{\binom{N-1}{d}^{N}}$$

where

$$\binom{a}{b}, \quad b \leq a,$$

is the binomial coefficient $a!/\,[b!(a-b)!]$. Note that there cannot be $N - d$ isolates, since d persons can be chosen for a maximum total of $(N-1)d$ times, less than the Nd choices actually made.

In the last paper cited above, Edwards goes on to set up the probability of $N - 2 - d$ isolates by eliminating irrelevant cases from those in which the isolates name d from a list of $d + 2$ while the non-isolates choose d from a list of $d + 1$ names, and indicates that the process might be continued to obtain the probabilities of $N - 3 - d$ isolates, etc. The form of these results, it is stated, would indicate a complicated algebraical expression for the required probability distribution and the question is then raised whether the existing technique of experimentation should not be modified to meet the requirement of simple mathematical treatment.

In this paper, we shall first obtain the exact distribution of the number of isolates on the assumption of random choice and second, we shall obtain an approximation which does satisfy the requirement of simple mathematical treatment. An example will be given to indicate the accuracy of the approximation for a typical application.

Exact distribution of number of isolates. It should first be remarked that any division of the group into those who are isolates and those who may not be produces two distinct patterns of choices. Each isolate selects d from among all those in the second group, but each member of the second group must select d from among those members of the second group not including himself. Let

$$P_i = \Pr\left\{\text{exactly } i \text{ isolates in group}\right\}$$

and

$$R_i = \Pr\left\{i \text{ specified individuals are isolates}\right\} .$$

Then,

$$(3)\quad R_i = \sum_{j=1}^{N-1-d} \Pr\left\{j \text{ isolates}\right\} \Pr\left\{\text{specified } i \text{ are among the } j \mid j \text{ isolates}\right\} ,$$

or

$$(3a)\quad R_i = \sum_{j=i}^{N-1-d} P_j \frac{\binom{j}{i}}{\binom{N}{i}} .$$

As a consequence of the remark made at the beginning of this section, we see that

$$(4)\quad R_i = \left[\frac{\binom{N-i}{d}}{\binom{N-1}{d}}\right]^i \left[\frac{\binom{N-1-i}{d}}{\binom{N-1}{d}}\right]^{N-i} ,$$

i.e the probability that all members of the group select from the N - i unspecified members. Therefore, to obtain the P_i, it is necessary to solve the system of equations,

$$(5) \qquad \sum_{j=i}^{n-1-d} \binom{j}{i} P_j = \binom{N}{i} R_i = Q_i, \qquad i = 0, 1, \ldots, N-1-d.$$

If the system (5) is written in matrix form as $AP = Q$, where P and Q are column vectors of the P_i's and Q_i's respectively and A is an $(N - d) \times (N - d)$ matrix with

$$a_{ij} = \binom{j}{i},$$

using the usual convention that

$$\binom{j}{i} = 0$$

if $i > j$, then the determinant $|A| = 1$, A has an inverse with element

$$a^{-1}_{ij} = (-1)^{i+j} \binom{j}{j - i}$$

and the system of equations (5) has the solution

$$(6) \qquad P_i = \sum_{j=i}^{N-1-d} (-1)^{i+j} \binom{j}{j - i} Q_j, \qquad i = 0, ., \ldots, N-1-d.$$

Equation (6) gives the exact probability of i isolates as a linear combination of the Q_j. Computation of the values of the Q_j may be made recursively, noting that $Q_o = 1$ and

$$(7) \qquad \frac{Q_{j+1}}{Q_j} = \frac{N - j - d}{j + 1} \left(\frac{N - j - d}{N - j}\right)^{j-1} \left(\frac{N - j - 1 - d}{N - j - 1}\right)^{N-j-1}$$

The form of the last term in (7) suggests interesting asymptotic behavior. We are, however, less interested in asymptotic characteristics of the distribution than in its properties for moderate values of N. We shall take the asymptotic behavior to give an indication of what may be a reasonable approximation, but the quality of the approximation must be judged

by results for typical cases; here, $\underline{N}$ is usually between 10 and 100. We shall later consider a typical example in which $\underline{N} = 26$, $\underline{d} = 3$.

If we do not require the values of the individual $\underline{P_i}$ but are only interested in the moments of the distribution of isolates, it turns out that the Q_j quantities are of central importance.

Let a_{ki}, $1 \leq i \leq k$, be the Stirling numbers of the second kind defined by

$$m^k = \sum_{i=1}^{k} a_{ki}\, m^{(i)}. \tag{8}$$

These coefficients were first tabulated by Stirling in 1730; a lengthy account of their history appears in Jordan's work on Finite Differences [5], where the table is extended to k = 12. More recently, Stevens [7] has given a table to k = 25.

Form the sum,

$$\sum_{i=1}^{k} a_{ki}\, i!\, Q_i = \sum_{i=1}^{k} \left\{ a_{ki}\, i! \sum_{j=i}^{N-1-d} \binom{j}{i} P_j \right\} \tag{9}$$

by equation (5). But

$$\binom{j}{i} = \frac{j^{(i)}}{i!} \quad \text{and} \quad j^{(i)} \equiv 0$$

for $j < i$. Then, since the order of summation may be interchanged,

$$\sum_{i=1}^{k} a_{ki}\, i!\, Q_i = \sum_{j=0}^{N-1-d} \left\{ P_j \sum_{i=1}^{k} a_{ki}\, j^{(i)} \right\} = \sum_{j=0}^{N-1-d} j^k P_j = \mu'_k, \tag{10}$$

the k-th moment (about zero) of the distribution. Jordan points out that the coefficients a_{ki} are related to Boole's "differences of zero" by the formula $i!\, a_{ki} = \Delta^i 0^k$. Thus the coefficient of Q_i in (10) is

$$b_{ki} = \Delta^i 0^k = \left[\Delta^i x^k\right]_{x=0}.$$

The b_{ki} may be computed recursively by

(11) $$b_{k+1,i} = i(b_{k,i-1} + b_{ki}),$$

where $b_{11} = 1$ and $b_{ko} = 0$ for all k. A short table of the coefficients suitable for moments up to the eighth order follows:

Table of $b_{ki} = \Delta^i 0^k$.

k \ i	1	2	3	4	5	6	7	8
1	1							
2	1	2						
3	1	6	6					
4	1	14	36	24				
5	1	30	150	240	120			
6	1	62	540	1,560	1,800	720		
7	1	126	1806	8,400	16,800	15,120	5,040	
8	1	254	5796	40,824	126,000	191,520	141,120	40,320

The factorial moments are even simpler, being given by

(12) $$\mu_{[k]} = \sum_{j=o}^{N-1-d} j^{(k)} P_j = k!\, Q_k.$$

We shall have occasion to use the factorial moments in the following section.

Approximate distribution of isolates. From (10), we compute the mean and the variance of the distribution of isolates and obtain, upon some simplification,

$$\text{mean}\,(i) = N\left(1 - \frac{d}{N-1}\right)^{N-1}$$

(13) and

$$\text{var}(i) = N\left(1 - \frac{d}{N-1}\right)^{N-1}\left[\,1 + (N-1-d)\left(1 - \frac{d}{N-2}\right)^{N-2} - N\left(1 - \frac{d}{N-1}\right)^{N-1}\right] .$$

From the second of these,

$$\text{var}(i) = \text{mean}(i)\left[\,1 - (d+1)\left(1 - \frac{d}{N-2}\right)^{N-2} + \text{terms in } (N-1)^{-2}\right].$$

The form of the variance strongly suggests the binomial frequency distribution with both parameters, n and p, to be determined. For the binomial distribution $\mu_{[k]} = n^{(k)} p^k$, and, fitting the first two moments, we should have

$$\hat{p} = \mu_{[1]} - \frac{\mu_{[2]}}{\mu_{[1]}} \quad \text{and} \quad \hat{n} = \frac{\mu_{[1]}}{\hat{p}} . \tag{14}$$

Also, since $\mu_{[k+1]}/\mu_{[k]} = (n - k)p$, we form the functions,

$$D_k = \frac{\mu_{[k+1]}}{\mu_{[k]}} - k\,\frac{\mu_{[2]}}{\mu_{[1]}} + (k - 1)\,\mu_{[1]}, \quad k = 2, 3, 4, \ldots, \tag{15}$$

which vanish identically for the binomial distribution. These functions are equivalent to the "total criteria" proposed by Guldberg [4] and Frisch [3] for judging whether an observed series may be approximated by a binomial frequency function. In their work, the approximation is considered to be good when the criterion functions of the moments of the observed series are close to zero. We shall extend the notion to cover the case of approximation of a more complicated probability law by the binomial law.

Setting k = 2 and k = 3 in (15) gives two functions which are exactly equivalent to the two criteria given by Guldberg (allowing for an omitted term in his second result). Also, the complete set (15) is equivalent to Frisch's total criteria for g = 1, h = 1, 2, 3,... in his notation. Since his criteria for all other values of g may be expressed in terms of those for g = 1, (15) is equivalent to the complete set of conditions given by Frisch.

Substituting from equation (12) into (15), we have

$$D_k = (k+1)\frac{Q_{k+1}}{Q_k} - 2k\frac{Q_2}{Q_1} + (k-1)Q_1$$

or, using (4), (5) and (7),

$$(16)\quad D_k = (N-k-d)\left(\frac{N-k-d}{N-1}\right)^{k-1}\left(\frac{N-k-1-d}{N-k-1}\right)^{N-k-1} - k(N-1-d)\left(\frac{N-2-d}{N-2}\right)^{N-2}$$

$$+ N(k-1)\left(\frac{N-1-d}{N-1}\right)^{N-1}.$$

For large N, each power of a fraction in (16) of the form

$$\left(\frac{a-d}{a}\right)^{a}$$

is approximately equal to e^{-d} and $D_k = 0$ (approx.). In the limit, every $D_k = 0$ and the asymptotic form of the distribution is, therefore, binomial. Further, the approximation should remain good even for moderate values of N (particularly when k is small) since the errors made by the exponential approximation are not only small but tend to compensate for each other.

We may, then, use a binomial probability law approximation with $\hat{p}$ and $\hat{n}$ given by (14) with

$$\mu_{[1]} = N\left(\frac{N - 1 - d}{N - 1}\right)^{N - 1}$$

(17) and

$$\mu_{[2]} = N(N - 1)\left(\frac{N - 1 - d}{N - 1}\right)^{N}\left(\frac{N - 2 - d}{N - 2}\right)^{N - 2}.$$

In the next section, we shall compare the approximation with the exact distribution for a typical pair of values of N and d.

An example. Moreno and Jennings [6] considered in some detail the case (N = 26, d = 3). Since, also, a number of later writers have treated the same case as a reasonably typical one, let us test the accuracy of the approximation in this situation. The computation of the exact probability distribution seemed to be best performed in two stages. In the first, the logarithms of the ratios Q_{j+1}/Q_j of equation (7) were obtained using 7-place tables, and the Q_i themselves obtained from the partial sums of the logarithms. These values appear in the second column of the table below. In the second stage of the computation, the exact probabilities were found by setting the Q_i into (6). The exact probabilities are given to six decimals in the third column of the table.

In the computation of the approximate probabilities, we take advantage of the already computed values of Q_1 and Q_2 and equation (12) to avoid the direct computation of the factorial moments of (17). From (14), we have $\hat{p} = .1717247$ and $\hat{n} = 6.197378$. We then compute the binomial probabilities, p_i, i = 0, ., 2, ..., [n] + 1, where [n] is the largest integer in n, in this case, 6, using $p_o = (1 - \hat{p})^{\hat{n}}$ and $p_{i+1}/p_i = (\hat{n}-i)\hat{p}/(i+1)(1-\hat{p})$ as suggested by Guldberg [4]. The approximate probabilities, p_i,

appear in the fourth column of the table to six decimals. It will be seen that the fit to three decimals is almost exact and certainly good enough for tests of significance. The discrepancies, $p_i - P_i$, are given in the last column.

Comparison of the Exact and Approximate Distribution of the Number of Isolates for N = 26, d = 3.

i	Q_i	P_i(exact)	p_i(approx.)	$p_i - P_i$
0	1.000 0000	·309 794	.311 098	+ .0013
1	1.064 2429	.402 574	.399 727	- .0028
2	.474 9281	.214 316	.215 365$^+$	+ .0010
3	.116 8650$^+$	.061 532	.062 473	+ .0009
4	.017 5606	.010 564	.010 354	- .0002
5	.001 6882	.001 138	.000 943	- .0002
6	.000 10596	.000 079	.000 039	- .00004
7	.000 0043 61	.000 003		
8	.000 0001 17			
9	.000 0000 02			

The discrepancies are not systematic except in the upper tail of the distribution, where the binomial gives zero probability for all numbers of isolates above seven. Although numbers through 22 are possible, they are so unlikely to occur by chance that this possibility may be practically disregarded. For example, the exact probability of eight isolates by chance is about one in ten million.

REFERENCES

1 U. Bronfenbrenner, "The measurement of sociometric status, structure and development," Sociometry, Vol. 6 (1943), pp. 363-397. Reprinted as Sociometry Monograph, No. 6, Beacon House, New York, 1945.

2 D. S. Edwards, "The constant frame of reference problem in sociometry," Sociometry, Vol. 11 (1948), pp. 372 - 379.

3 R. Frisch, "On the use of difference equations in the study of frequency distributions," Metron, Vol. 10 (1932), pp. 35-59.

4 A. Guldberg, "On discontinuous frequency functions and statistical series," Skand. Aktuar. tids., Vol. 14 (1931), pp. 161-187.

5 C. Jordan, Calculus of Finite Differences, Rottig and Romwalter, Budapest, 1939.

6 J. L. Moreno and H. H. Jennings, "Statistics of social configurations," Sociometry, Vol. 1 (1938), pp. 342-374. Reprinted as Sociometry Monograph, No. 3 , Beacon House, New York, 1945.

7 W. L. Stevens, "Significance of grouping," Annals of Eugenics, Vol. 8 (1937), pp. 57-73.