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

By Leo Katz

University of North Carolina and Michigan State College

1. Summary. The exact chance distribution of the number of isolates in a social group is found in this paper, using methods due to Fréchet. The binomial distribution fitted to the first two moments of the exact distribution is shown to give reasonably good approximation and a slightly coarser binomial approximation is indicated.

2. Introduction. Consider a group consisting of N individuals. Each designates d of the others with whom he would prefer to be associated in some specified activity, that is, 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 [8], gave the expected (mean) number of isolates as

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

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,"

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

where $a^{(b)} = a(a - 1)(a - 2)\cdots(a - b + 1)$. This form gives completely nonsensical results in application. Edwards [2] conjectured that the Bronfenbrenner formula gives the probability of a given person's including in his list of d at least

[1] Work done under the sponsorship of the Office of Naval Research at Chapel Hill, North Carolina, and presented at the Chicago meeting of the Institute of Mathematical Statistics, December 27, 1950.

one of $i + 1$ specified names. Edwards then gave correctly the probability of the maximum possible number of isolates,

$$(2)\quad P[N-1-d] = \Pr\{N-1-d \text{ isolates}\} = \binom{N}{N-1-d}\frac{(d+1)^{(N-1-d)}}{\binom{N-1}{d}^N},$$

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

In the last paper cited above, Edwards went 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 nonisolates choose d from a list of $d + 1$ names, and indicated 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 algebraic 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 practical 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.

3. Exact distribution of the 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

$$p_{i_1,i_2,\cdots,i_k} = \Pr\{\text{individuals } i_1, i_2, \cdots, i_k \text{ are isolates}\}.$$

As an immediate consequence of the remark made above and the symmetry of the situation,

$$(3)\quad p_{i_1,i_2,\cdots,i_k} = \left[\frac{\binom{N-k}{d}}{\binom{N-1}{d}}\right]^k \cdot \left[\frac{\binom{N-k-1}{d}}{\binom{N-1}{d}}\right]^{N-k},$$

for every $(i_1, i_2, \cdots, i_k)$. Setting

$$(4)\quad S_k = \binom{N}{k} p_{i_1,i_2,\cdots,i_k} = \binom{N}{k}\binom{N-k}{d}^k\binom{N-k-1}{d}^{N-k}\binom{N-1}{d}^{-N}$$

the principle of inclusion and exclusion ([3], ch. 4) gives immediately

$$(5)\quad P_{[k]} = \Pr\{\text{exactly } k \text{ isolates in the group}\} = \sum_{j=k}^{N-1-d}(-1)^{k+j}\binom{j}{k}S_j.$$

Equation (5) gives the exact probability of k isolates, in a group of N where each individual makes d choices, as a linear combination of the S_k.

The values of S_k may be computed directly from (4) or recursively, noting that $S_0 = 1$ and

$$\frac{S_{k+1}}{S_k} = \frac{N-k-d}{k+1}\left[\frac{N-k-d}{N-k}\right]^{k-1}\left[\frac{N-k-1-d}{N-k-1}\right]^{N-k-1}. \tag{6}$$

The form of the last term in (6) 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 may 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, N is usually between 10 and 100. We shall later consider one such typical example in which $N = 26$, $d = 3$.

If we do not require the values of the individual $P_{[i]}$ but are only interested in the moments of the distribution of isolates, it turns out that the S_k quantities are of central importance. Fréchet [4] has shown that

$$\alpha_{(k)} = k!\, S_k, \tag{7}$$

where $\alpha_{(k)}$ is the kth factorial moment of the distribution, given by $\alpha_{(k)} = \sum_{i=1}^{N-1-d} i^{(k)} P_{[i]}$. We shall have occasion to use these factorial moments in the following section.

4. Approximate distribution of number of isolates. Since we know the exact distribution, an approximate distribution is useful only if it is more easily computed. It is easily shown (see Feller [3]) that, for d fixed, the limiting distribution is Poisson with $\Pr(k) = e^{-\lambda}\lambda^k/k!$, where $\lambda = N(1 - d/(N-1))^{N-1}$. However, for moderate values of N, the approximation is not good; an example is given later.

Following the procedure of Kaplansky [7] produces a modified Poisson approximation which is quite good. The drawback to this procedure is that computations are almost as difficult as for the exact distribution. Therefore, we seek another approximation to satisfy the dual requirements of accuracy and simplicity.

From (4) and (7), the mean and the variance of the number of isolates are respectively,

$$\text{mean} = \alpha_{(1)} = N\left(1 - \frac{d}{N-1}\right)^{N-1}, \tag{8}$$

$$\begin{aligned}\text{variance} &= \alpha_{(2)} + \alpha_{(1)} - \alpha_{(1)}^2 \\ &= N\left(1 - \frac{d}{N-1}\right)^{N-1}\left[1 + (N-1-d)\left(1 - \frac{d}{N-2}\right)^{N-2} - N\left(1 - \frac{d}{N-1}\right)^{N-1}\right].\end{aligned} \tag{9}$$

From (9), we see var (k) = mean (k) $[1 - (d + 1)(1 - d/(N - 2))^{N-2} + O(N^{-2})]$ $\approx$ mean (k) $[1 - (d + 1)e^{-d}]$. Since the variance is less than the mean, the binomial distribution, $b(x; n, p)$, is strongly suggested (choice being restricted to simple distributions). We shall not insist that n be an integer; thus, we have essentially a beta distribution. For this distribution, $\alpha_{(r)} = n^{(r)}p^r$ and, fitting the first two moments, we have

$$np = \alpha_{(1)} = N\left(1 - \frac{d}{N-1}\right)^{N-1}, \tag{10}$$

$$\frac{1}{n} = 1 - \frac{\alpha_{(2)}}{\alpha_{(1)}^2} = 1 - \left(1 - \frac{1}{N}\right)\left[1 - \frac{d}{(N-2)(N-1-d)}\right]^{N-2}. \tag{11}$$

Also, since $\alpha_{(r+1)}/\alpha_{(r)} = (n - r)p$, we form the functions,

$$D_r = \frac{\alpha_{(r+1)}}{\alpha_{(r)}} - r\frac{\alpha_{(2)}}{\alpha_{(1)}} + (r - 1)\alpha_{(1)}, \qquad r = 2, 3, 4, \cdots, \tag{12}$$

which vanish identically for the binomial distribution. These functions are equivalent to the "total criteria" proposed by Guldberg [6] and Frisch [5] 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 $r = 2$ and $r = 3$ in (12) 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 (12) is equivalent to Frisch's total criteria for $g = 1$, $h = 1, 2, 3, \cdots$ in his notation. Since his criteria for all other values of g may be expressed in terms of those for $g = 1$, (12) is equivalent to the complete set of conditions given by Frisch.

Substituting from equation (7) into (12), we have

$$D_r = (r + 1)\frac{S_{r+1}}{S_r} - 2r\frac{S_2}{S_1} + (r - 1)S_1,$$

or, using (4) and (6),

$$\begin{aligned} D_r = {} & (N - r - d)\left(\frac{N - r - d}{N - 1}\right)^{r-1}\left(\frac{N - r - 1 - d}{N - r - 1}\right)^{N-r-1} \\ & - r(N - 1 - d)\left(\frac{N - 2 - d}{N - 2}\right)^{N-2} + N(r - 1)\left(\frac{N - 1 - d}{N - 1}\right)^{N-1}. \end{aligned} \tag{13}$$

For large N, each power of a fraction in (13) of the form $((a - d)/a)^a$ is approximately equal to e^{-d} and $D_r = 0$, approximately. In the limit, every $D_r = 0$, the asymptotic form of the distribution in this sense is, therefore, binomial. Further, the approximation should remain good even for moderate values of N (particularly when r 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 p and n given by (10) and (11). (If $1/n$ in (11) is evaluated to terms of $O(N^{-2})$, we find $1/n = (d+1)/(N-1-d)$ or $n = N/(d+1) - 1$, approximately. This seems consistently to understate the value of n from (11); accordingly, it is suggested that n be approximated by

$$n = \frac{N}{d+1} - \frac{1}{2}. \tag{11a}$$

In the next section, we shall compare this approximation with the exact distribution for a typical pair of values of N and d. We also give, for comparison, the Poisson approximation.

TABLE 1

Comparison of the exact and approximate distributions of the number of isolates for $N = 26$, $d = 3$

i	S_i	$P_{[i]}$(exact)	p_i(approx.)	$p_i - P_{[i]}$	$p'_i = \frac{e^{-\lambda}\lambda^i}{i!}$	$p'_i - P_{[i]}$
0	1.000 0000	.309 794	.311 098	+.0013	.344 989	+.0352
1	1.064 2429	.402 574	.399 727	−.0028	.367 152	−.0354
2	474 9281	.214 316	.215 365+	+.0010	.195 370	−.0189
3	.116 8650+	.061 532	.062 473	+.0009	.069 306	+.0078
4	.017 5606	.010 564	.010 354	−.0002	.018 440	+.0079
5	.001 6882	.001 138	.000 943	−.0002	.003 925−	+.0028
6	.000 10596	.000 079	.000 039	−.00004	.000 696	+.00062
7	.000 0043 61	.000 003	.000 0002	−.000003	.000 106	+.000103
8	.000 0001 17				.000 014	
9	.000 0000 02				.000 002	

5. An example. Moreno and Jennings [8] 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, we will test the accuracy of the approximation in this situation. The computation of the exact probability distribution seems to be best performed in two stages. In the first, the logarithms of the ratios S_{j+1}/S_j of equation (6) are obtained using 7-place tables, and the S_i themselves obtained from the partial sums of the logarithms. These values appear in the second column of Table 1. In the second stage of the computation, the exact probabilities are found by setting the S_i into (5). 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 S_1 and S_2 and equation (7) to obtain directly the factorial moments of (8) and (9). From (10) and (11), we have $p = .1717247$ and $n = 6.197378$. We then compute the binomial probabilities, $p_i = b(i; n, p)$,

$i = 0, 1, 2, \cdots, ([n] + 1)$, where $[n]$ is the largest integer in n, in this case, 6, using $p_0 = (1 - p)^n$ and $p_{i+1}/p_i = (n - i)p/(i + 1)(1 - p)$ as suggested by Guldberg [6] and others. 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 fifth column. The Poisson probabilities and errors appear in the sixth and seventh columns.

The discrepancies for the "binomial" approximation are not particularly 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. The Poisson distribution appears to be "flatter" than the exact, understating probabilities for the central values and overstating for both tails.

As a further check on the accuracy of the approximation, the values of $\gamma_1 = \mu_3/\mu_2^{3/2}$ and $\gamma_2 = \mu_4/\mu_2^2$ were computed for the exact distribution and for the "binomial" approximation. These computations give $\gamma_1 = .7193$ for the exact, .6993 for the approximate distribution; $\gamma_2 = 3.2620$ and 3.1663, respectively.

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] W. Feller, *An Introduction to Probability Theory and Its Applications*, John Wiley and Sons, 1950.

[4] M. Fréchet, *Les Probabilités Associées à un système d'Événements Compatibles et Dépendants, Actualités Scientifiques et Industrielles, Nos. 859 and 942*, Hermann et Cie, Paris, 1940 and 1943.

[5] R. Frisch, "On the use of difference equations in the study of frequency distributions," *Metron*, Vol. 10 (1932), pp. 35–59.

[6] A. Guldberg, "On discontinuous frequency functions and statistical series," *Skandinavisk Aktuarietidskrift*, Vol. 14 (1931), pp. 161–187.

[7] I. Kaplansky, "The asymptotic distribution of runs of consecutive elements," *Annals of Math. Stat.*, Vol. 16 (1945), pp. 200–203.

[8] 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.