ON A CERTAIN STATISTIC IN A SOCIAL GROUP

By HIROJIRO AOYAMA

(Received June 10, 1957)

1. Introduction

When selecting a leader in a small group, for example, a leader for the athletic games in a small group of pupils, we often need to decide which factor is effective for it. Some pupils select the leader from the standpoint of athletic ability and others from that of popularity in their class room. Here, we are interested in deciding which of these factors is effective for the selection. In this paper we shall show the distribution of a certain statistic for the decision when the group is divided into three sub-groups, that is, upper, middle and lower groups with regard to a certain characteristic.

2. Statistic u of the number of selections from the upper group.

Suppose the pupils select d members in their group in order to decide the leader(s) for an athletic game. We divide the group into three subgroups—upper, middle and lower subgroups—which consist of N_1, N_2 and N_3 pupils, respectively. If we denote by $f(N_1-1, N_2, N_3; d)$ the number of possible selections for by any pupil in the upper subgroup, we have obviously

$$(1)\qquad f(N_1-1, N_2, N_3; d)\equiv\binom{N-1}{d}=\binom{N_1-1}{d}+\binom{N_1-1}{d-1}\binom{N_2+N_3}{1}$$
$$+\binom{N_1-1}{d-2}\binom{N_2+N_3}{2}+\cdots+\binom{N_1-1}{0}\binom{N_2+N_3}{d}$$

where $N=N_1+N_2+N_3$. In this last expression the first term equals to the number of cases in which d persons are all selected from the upper subgroup, the second term equals to the number of cases in which $d-1$ persons are selected from the upper subgroup and one person from the other subgroups, and so on.

Similarly we denote by $f(N_1, N_2-1, N_3; d)$ and $f(N_1, N_2, N_3-1; d)$ the numbers of possible selections by two pupils each from the middle and lower subgroups, respectively. Then we have

$$(2)\qquad f(N_1, N_2-1, N_3; d)\equiv\binom{N-1}{d}=\binom{N_1}{d}+\binom{N_1}{d-1}\binom{N_2+N_3-1}{1}+\binom{N_1}{d-2}\binom{N_2+N_3-1}{2}+\cdots+\binom{N_1}{0}\binom{N_2+N_3-1}{d}$$

$$(3)\qquad f(N_1, N_2, N_3-1; d)\equiv\binom{N-1}{d}=f(N_1, N_2-1, N_3; d)$$

Assume the selection of members occurs independently. Then the probabilities, that a pupil in the upper subgroup select d, $d-1$, $d-2$, $\cdots$, 0 members from his group, are $\binom{N_1-1}{d}\Big/\binom{N-1}{d}$, $\binom{N_1-1}{d-1}\binom{N_2+N_3}{1}\Big/\binom{N-1}{d}$, $\cdots$, and $\binom{N_2+N_3}{d}\Big/\binom{N-1}{d}$ which we denote by p_d, p_{d-1}, $\cdots$, p_1, and p_0, respectively. Similarly we denote those probabilities with regard to the middle and lower subgroups q_d, q_{d-1}, $\cdots$, q_1, q_0 and r_d, r_{d-1}, $\cdots$, r_1, r_0, respectively. For example, q_i is the probability that a pupil in the middle subgroup select i members from the upper subgroup and $d-i$ members from the other subgroups.

As we assume the selections of pupils occur independently, we may think the probability $P(x_1, x_2, \cdots, x_d, x_0)$ that x_i pupils in the upper subgroup select i members from the same group and $d-i$ members from the other groups is defined by the multinomial probability, that is,

$$(4)\qquad P(x_1, x_2, \cdots, x_d, x_0)=\frac{N_1!}{x_1!\,x_2!\cdots x_d!\,x_0!}\prod_{i=0}^{d}p_i^{x_i}$$

where $$N_1=x_1+x_2+\cdots+x_d+x_0$$

In this case the total number of persons selected from the upper subgroup is $x=x_1+2x_2+3x_3+\cdots+dx_d$, when we count twice a person selected two times and so on. Similarly, as to the probabilities $P(y_1, y_2, \cdots, y_d, y_0)$ and $P(z_1, z_2, \cdots, z_d, z_0)$ for other groups we have

$$(5)\qquad P(y_1, y_2, \cdots, y_d, y_0)=\frac{N_2!}{y_1!\,y_2!\cdots y_d!\,y_0!}\prod_{i=0}^{d}q_i^{y_i}$$

$$N_2=y_1+y_2+\cdots+y_d+y_0$$

$$y=y_1+2y_2+\cdots+dy_d$$

$$(6)\qquad P(z_1\,z_2, \cdots, z_d, z_0)=\frac{N_3!}{z_1!\,z_2!\cdots z_d!\,z_0!}\prod_{i=0}^{d}r_i^{z_i}$$

$$N_3=z_1+z_2+\cdots+z_d+z_0$$
$$y=z_1+2z_2+\cdots+dz_d$$

Now we shall consider the distribution of $u=x+y+z$.* This statistic u represents the total number of persons selected from the upper subgroup by all members. Hence the u obtained by an experiment is significant if it is greater than a certain u_0 which is calculated by the probability distribution for a given significance level.

3. $E(u)$ and $D^2(u)$.

If the selections of all members are assumed to occur independently, we can derive the expectation $E(u)$ and the variance $D^2(u)$ of the statistic u. From (4), (5) and (6) we have

$$(7)\qquad E(u)=\sum_{i=1}^{d} i\{E(x_i)+E(y_i)+E(z_i)\}$$
$$=N_1\sum_{i=1}^{d}(ip_i)+N_2\sum_{i=1}^{d}(iq_i)+N_3\sum_{i=1}^{d}(ir_i)$$

On the other hand we can prove easily

$$(8)\qquad \sum_{i=1}^{d} i\binom{N_1-1}{i}\binom{N_2+N_3}{d-i}=(N_1-1)\binom{N-2}{d-1}$$

and

$$(9)\qquad \sum_{i=1}^{d} i\binom{N_1}{i}\binom{N_2+N_3-1}{d-i}=N_1\binom{N-2}{d-1}$$

hence we get by some reduction from (7), (8) and (9)

$$(10)\qquad E(u)=N_1d$$

As for the variance of u it follows from the independence among x, y and z that

$$(11)\qquad D^2(u)=D^2(x)+D^2(y)+D^2(z)$$

Here we get

$$(12)\qquad D^2(x)=\sum_{i=1}^{d} i^2D^2(x_i)+2\sum\sum_{i<j} ij\ \mathrm{cov}\,(x_i,\ x_j)$$
$$=N_1\left\{\sum_{i=1}^{d} i^2p_i(1-p_i)-2\sum\sum_{i<j} ijp_ip_j\right\}$$

* Essentially we consider only two groups of sizes N_1 and N_2+N_3.

$$=N_1\left\{\sum_{i=1}^{d} i^2 p_i-\left(\sum_{i=1}^{d} ip_i\right)^2\right\}$$

Further we have

$$\text{(13)}\qquad \sum_{i=1}^{d} i^2\binom{N_1-1}{i}\binom{N_2+N_3}{d-i}=(N_1-1)(N_2-2)\binom{N-3}{d-2}+(N_1-1)\binom{N-2}{d-1}$$

For we have

$$(N_1-1)(N_1-2)(a+b)^{N_1-3}\equiv\frac{\partial^2}{\partial a^2}\left\{(a+b)^{N_1-1}\right\}$$

$$=\sum_{k=2}^{N_1-1} k(k-1)\frac{(N_1-1)!}{k!(N_1-1-k)!}a^{k-2}b^{N_1-1-k}$$

and

$$(N_1-1)(a+b)^{N_1-2}\equiv\frac{\partial}{\partial a}\left\{(a+b)^{N_1-1}\right\}=\sum_{k=1}^{N_1-1} k\frac{(N_1-1)!}{k!(N_1-1-k)!}a^{k-1}b^{N_1-1-k}$$

and by adding these two equations

$$a(N_1-1)(N_1-2)(a+b)^{N_1-3}+(N_1-1)(a+b)^{N_1-2}$$
$$=\sum_{k=1}^{d} k^2\binom{N_1-1}{k}a^{k-1}b^{N_1-1-k}\,.$$

Multiply this equation by $(a+b)^{N_2+N_3}$ and compare the coefficient of $a^{d-1}b^{N-1-d}$ of two sides, and then we have (13).

Hence we have from (12) and (13)

$$\text{(14)}\qquad D^2(x)=N_1\left[\frac{(N_1-1)(N_1-2)\binom{N-3}{d-2}+(N_1-1)\binom{N-2}{d-1}}{\binom{N-1}{d}}-\left(\frac{(N_1-1)d}{N-1}\right)^2\right]$$

Similarly we get

$$\text{(15)}\qquad D^2(y)=N_2\left[\frac{N_1(N_1-1)\binom{N-3}{d-2}+N_1\binom{N-2}{d-1}}{\binom{N-1}{d}}-\left(\frac{N_1 d}{N-1}\right)^2\right]$$

and

$$\text{(16)}\qquad D^2(z)=N_3\left[\frac{N_1(N_1-1)\binom{N-3}{d-2}+N_1\binom{N-2}{d-1}}{\binom{N-1}{d}}-\left(\frac{N_1 d}{N-1}\right)^2\right],$$

therefore, we have

$$D^2(u)=N_1d\left(1-\frac{N_1+d-1}{N-1}+\frac{(N_1-1)d}{(N-1)^2}\right) \tag{17}$$

4. The limiting distribution and the practically approximated distribution of u.

From the equation (17) we have for large N

$$D^2(u)\sim N_1d \tag{18}$$

hence keeping in mind the equation (10), we have as the limiting distribution the Poisson distribution with the mean N_1d, that is

$$P(u)=\frac{(N_1d)^ue^{-N_1d}}{u!} \tag{19}$$

But this approximation is not adequate for small N, because the variance $D^2(u)$ is smaller than $E(u)$, as is seen from

$$D^2(u)=N_1d\left(1-\frac{N_1-1}{N-1}-\frac{N-N_1}{(N-1)^2}d\right)$$

Now let us calculate the factorial moments $\alpha_{(k)}=E(u(u-1)\cdots(u-k+1))$. If these moments satisfy the equation

$$D_k\equiv\frac{\alpha_{(k+1)}}{\alpha_{(k)}}-k\frac{\alpha_{(2)}}{\alpha_{(1)}}+(k-1)\alpha_{(1)}=0\,, \tag{20}$$

we can use the binomial approximation $b(u;\,n,\,p)$ by Katz' criteria [1].

Then we have

$$\alpha_{(1)}\equiv E(u)=N_1d\,. \tag{21}$$

As for $\alpha_{(2)}$ we have

$$\begin{aligned} u(u-1)&=(x+y+z)(x+y+z-1)\\ &=x(x-1)+y(y-1)+z(z-1)+2xy+2xz+2yz \end{aligned} \tag{22}$$

and

$$\begin{aligned} x(x-1)&=\sum_{i=1}^{d}i^2x_i^2+2\sum\sum_{i<j}ijx_ix_j-\sum_{i=1}^{d}ix_i\\ &=\sum_{i=1}^{d}i^2x_i^{(2)}+\sum_{i=1}^{d}i(i-1)x_i+2\sum\sum_{i<j}ijx_ix_j \end{aligned} \tag{23}$$

where

$$x_i^{(2)}=x_i(x_i-1)$$

Therefore we have

$$(24)\qquad E(x(x-1))=N_1^{(2)}\sum_{i=1}^{d}i^2p_i^2+N_1\sum_{i=2}^{d}i^{(2)}p_i+2N_1(N_1-1)\sum_{i<j}\sum ijp_ip_j$$

$$=N_1^{(2)}\Big(\sum_{i=1}^{d}ip_i\Big)^2+N_1\sum_{i=1}^{d}i^{(2)}p_i$$

$$=N_1^{(2)}\Big(\frac{N_1-1}{N-1}\,d\Big)^2+N_1^{(3)}\frac{d^{(2)}}{(N-1)^{(2)}}$$

by the above mentioned formula :

$$\sum_{i=1}^{d}i^{(2)}p_i=\frac{(N_1-1)(N_1-2)d(d-1)}{(N-1)(N-2)}$$

where $N_1^{(2)}=N_1(N_1-1)$, $N_1^{(3)}=N_1(N_1-1)(N_1-2)$ and $i^{(2)}=i(i-1)$. Similarly we can derive the analogous equation for $E\{y(y-1)\}$ and $E\{z(z-1)\}$, so that

$$(25)\quad E(u(u-1))=N_1^{(2)}\Big(\frac{N_1-1}{N-1}\,d\Big)^2+N_2^{(2)}\Big(\frac{N_1}{N-1}\,d\Big)^2+N_3^{(2)}\Big(\frac{N_1}{N-1}\,d\Big)^2$$

$$+2N_1N_2\Big(\frac{N_1-1}{N-1}\,d\Big)\Big(\frac{N_1}{N-1}\,d\Big)+2N_1N_3\Big(\frac{N_1-1}{N-1}\,d\Big)\Big(\frac{N_1}{N-1}\,d\Big)$$

$$+2N_2N_3\Big(\frac{N_1}{N-1}\,d\Big)^2+\frac{N_1^{(2)}d^{(2)}}{N-1}=N_1d\Big[\Big\{N_1+\frac{N-N_1}{(N-1)^2}\Big\}d-\frac{N_1-1}{N-1}\Big]$$

We can also rewrite (25) as follows :

$$(26)\qquad E(u(u-1))=N_1^2[p1]^2+N_2^2[q1]^2+N_3^2[r1]^2+2N_1N_2[p1][q1]$$

$$+2N_1N_3[p1][r1]+2N_2N_3[q1][r1]$$

$$+N_1\{[p(2)]-[p1]^2\}+N_2\{[q(2)]-[q1]^2\}+N_3\{[r(2)]-[r1]^2\}$$

where

$$[p1]=\sum_{i=1}^{d}ip_i=\frac{N_1-1}{N-1}\,d\,,$$

$$[p(2)]=\sum_{i=1}^{d}i^{(2)}p_i=\frac{(N_1-1)^{(2)}d^{(2)}}{(N-1)^{(2)}}$$

and so on.

Thus we have

$$(27)\qquad E(u(u-1))=(N_1d)^2+N_1\{[p(2)]-[p1]^2+N_2\{[q(2)]-[q1]^2\}$$

$$+N_3\{[r(2)]-[r1]^2\}=(N_1d)^2+O(1/N)$$

Similarly we have

$$
\begin{aligned}
(28)\qquad E(u(u-1)(u-2)) &= (N_1 d)^3 + 3N_1^2[p1]\{[p(2)]-[p1]^2\} \\
&\quad + 3N_2^2[q1]\{[q(2)]-[q1]^2\} + 3N_3^2[r_1]\{[r(2)]-[r1]^2\} \\
&\quad + 3N_1N_2\{[p1]([q(2)]-[q1]^2) + [q1]([p(2)]-[p1]^2)\} \\
&\quad + 3N_2N_3\{[q1]([r(2)]-[r1]^2) + [r1]([q(2)]-[q1]^2)\} \\
&\quad + 3N_3N_1\{[r1]([p(2)]-[p1]^2) + [p1]([r(2)]-[r1]^2)\} \\
&\quad + N_1\{[p(3)]-3[p(2)][p1]+2[p1]^3\} \\
&\quad + N_2\{[q(3)]-3[q(2)][q1]+2[q1]^3\} \\
&\quad + N_3\{[r(3)]-3[r(2)][r1]+2[r1]^3\} \\
&\leqq (N_1 d)^3 + 3N^2 \max\,([p1]\{[2p)]-[p1)^2\}, \\
&\quad [q1]\{[q(2)]-[q1]^2\}\ ,\ \cdots\ , \\
&\quad [p1]([r(2)]-[r1]^2)) + 3N \max\,([p(3)]-3[p(2)][p1] \\
&\quad + 2[p1]^3, \cdots) \\
&= (N_1 d)^3 + O(1/N) + O(1/N^2)
\end{aligned}
$$

where evidently $[p(k)]$, $[q(k)]$ and $[r(k)]$ exist only for $d \geqq k$.

Therefore we may think that we have generally

$$(29)\qquad E(u(u-1)\cdots(u-k+1)) \sim (N_1 d)^k$$

although the author can not have a general expression of $\alpha_{(k)}$. Thus the criterion D_k of Katz is approximately zero for large N. Accordingly we may use the approximated binomial distribution, and we have from (10) and (17)

$$(30)\qquad p = \frac{N_1+d-1}{N-1} - \frac{N_1-1}{(N-1)^2}d$$

$$(31)\qquad n = N_1 d/p$$

We shall here present an example for $d=1$, $N_1=N_3=4$, $N_2=12$ in the following table. In this case it holds $p \fallingdotseq 0.20$, $n \fallingdotseq 20$ for the bino-

	Exact value	Binomial approximation	Poisson approximation
0	.0115	.0115	.0183
1	.0575	.0576	.0733
2	.1368	.1369	.1465
3	.2056	.2054	.1954
4	.2186	.2182	.1954
5	.1748	.1746	.1563
6	.1092	.1091	.1042
7	.0545	.0545	.0595
8	.0223	.0222	.0298
9	.0073	.0074	.0132
10	.0020	.0020	.0053
11	.0004	.0005	.0019
12	.0001	.0001	.0006

mial approximation and the result is better than that for the Poisson approximation.

For $d=3$ we have also the following table. In this case we get $p\fallingdotseq 0.292683$ and $n\fallingdotseq 41$ but the binomial approximation for $p=0.3$, $n=40$ is better, that is, $p=\dfrac{N_1+d-1}{N-1}-\dfrac{(N_1-1)d}{2(N-1)^2}$ is better than equation (30).

	Exact value	b (u; 41, 0.292683)	b (u; 40, 0.3)
0	.0000 006	.0000 007	.0000 006
1	.0000 108	.0000 116	.0000 110
2	.0000 911	.0000 958	.0000 912
3	.0004 978	.0005 155	.0004 951
4	.0019 073	.0020 263	.0019 630
5	.0061 354	.0062 048	.0060 572
6	.0152 84	.0154 049	.0151 428
7	.0314 08	.0318 722	.0315 220

THE INSTITUTE OF STATISTICAL MATHEMATICS

REFERENCE

[1] Katz, Leo: The distribution of the number of isolates in a social group, *Annals of Math. Stat.*, vol. 23, 1952.