This result implies that the distribution $\left\{\frac{\bar{n}_w - m_1(\bar{n})}{\sigma_{\bar{n}}}; P(\bar{n})\right\}$ has the limiting normal distribution $\left\{x, \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}\right\}$, as $w \to \infty$.

A SEQUENCE OF DISCRETE VARIABLES EXHIBITING CORRELATION DUE TO COMMON ELEMENTS

By Carl H. Fischer

University of Michigan

1. **Introduction.** Studies of correlation due to common elements have been made more or less sporadically over the past thirty years in attempts to throw more light on the meaning of correlation. Numerous examples may be cited. One of the earliest was a study by Kapteyn [1] in which he showed that two sums, each of n elements drawn from a normal population with k elements in common, had a correlation coefficient of k/n. This was considerably generalized by the writer [3] who considered sums of different numbers of elements drawn from quite arbitrary continuous distributions. The work was extended to include sequences of three or more such sums. Antedating this latter paper, Rietz [2] has devised various urn schemata in one of which pairs of drawings of s balls each were produced with t balls held in common. The coefficient of correlation between the numbers of white balls in each of the pairs of drawings was found to be t/s.

Fairly recently some interest has been shown in this subject in connection with the study of heredity; hence it appeared that it might be of value to present the following study by elementary methods of a sequence of discrete variables in which each member is linked to the adjacent members by various specified numbers of common elements.

2. **Two variables.** A pair of discrete variables is defined as follows: The first, x, is equal to the number of white balls in a set of s_1 balls drawn one at a time from an urn which is so maintained that the probability of drawing a white ball is always a constant, p. The second, y, is equal to the number of white balls in a second set of s_2 balls formed by drawing t_{12} balls at random from the s_1 balls of the first set plus $s_2 - t_{12}$ balls drawn directly from the urn. The numbers s_1 and s_2 may or may not be equal.

Evidently the marginal distribution of x follows the Bernoulli law and is given by $\binom{s_1}{x} q^{s_1-x} p^x$.[1] The first step in finding $P(x, y:t_{12})$, the bivariate distribution

[1] By $\binom{a}{b}$ is meant the number of combinations of a items taken b at a time. It shall be understood that $\binom{a}{b} = 0$ if $b < 0$ or $b > a$.

function of x and y with t_{12} balls in common between the two drawings, is to write the product of the three probabilities: of obtaining x white balls in the first set; of drawing d of these whites in the t_{12} balls chosen at random from this set; of drawing exactly $y - d$ white balls among the $s_2 - t_{12}$ balls drawn directly from the urn to complete the second set. This product may readily be reduced to the form shown below in (1), symmetric in x and y and in s_1 and s_2, which is then summed on d from 0 to t_{12}. Thus

$$(1) \qquad P(x, y:t_{12}) = \sum_{d=0}^{t_{12}} \binom{s_1 - t_{12}}{x - d}\binom{t_{12}}{d}\binom{s_2 - t_{12}}{y - d} q^{s_1+s_2-t_{12}-x-y+d} p^{x+y-d}.$$

The marginal distribution of x has already been given. From the symmetry of (1) it is obvious that the corresponding marginal distribution of y must be characterized by the Bernoulli distribution function $\binom{s_2}{y} q^{s_2-y}p^y$. The variances of the marginal distributions are s_1pq and s_2pq, respectively.

We next proceed to demonstrate that both of the regression curves are linear and to find the equations of the lines. Consider an array of x on y for some fixed value of y. The mean of the array is

$$(2) \qquad \bar{x}_y = \binom{s_2}{y}^{-1} q^{-(s_2-y)} p^{-y} \sum_{x=0}^{s_1} xP(x, y:t_{12}).$$

The summation in the right member of (2) may be expanded and then rewritten as

$$(3) \qquad \sum_{d=0}^{t_{12}} \binom{t_{12}}{d}\binom{s_2 - t_{12}}{y - d} q^{s_2-y} p^y \sum_{x=0}^{s_1} x \binom{s_1 - t_{12}}{x - d} q^{s_1-t_{12}-x+d} p^{x-d}.$$

The inner summation in (3) is seen to equal $d + p(s_1 - t_{12})$ and hence (2) becomes

$$\begin{aligned} \bar{x}_y &= \binom{s_2}{y}^{-1}\left\{\sum_{d=0}^{t_{12}} \binom{t_{12}}{d}\binom{s_2 - t_{12}}{y - d}[d + p(s_1 - t_{12})]\right\} \\ &= \binom{s_2}{y}^{-1}\left\{t_{12} \sum_{d=0}^{t_{12}} \binom{t_{12} - 1}{d - 1}\binom{s_2 - t_{12}}{y - d} + p(s_1 - t_{12})\binom{s_2}{y}\right\}. \end{aligned}$$

Then the equation of the line of regression of x on y becomes

$$(4) \qquad \bar{x}_y = t_{12}y/s_2 + p(s_1 - t_{12}).$$

By symmetry, the line of regression of y on x may be seen to be

$$\bar{y}_x = t_{12}x/s_1 + p(s_2 - t_{12}).$$

The square of the correlation coefficient is equal to the product of the slopes of the two regression lines, hence

$$(5) \qquad r_{xy} = t_{12}/(s_1s_2)^{\frac{1}{2}}.$$

If $s_1 = s_2 = s$ we have the familiar result t/s.

3. **Three variables.** A third variable, z, may now be defined as the number of white balls in a set of s_3 balls formed by drawing t_{23} balls at random from the s_2 of the second set plus $s_3 - t_{23}$ drawn directly from the urn. It is evident from the results on two variables that the marginal distribution of z follows the Bernoulli law and that the equations of the regression lines of z on y and y on z are

$$\bar{z}_y = t_{23}y/s_2 + p(s_3 - t_{23});$$

$$\bar{y}_z = t_{23}z/s_3 + p(s_2 - t_{23}).$$

The correlation coefficient, r_{yz}, is equal to $t_{23}/(s_2 s_3)^{\frac{1}{2}}$.

The relationship between x and z remains to be investigated. The probability of the joint occurrence of x whites on the first drawing and z whites on the third when it is specified that the s_1 and s_3 balls of the two sets shall include the same g balls in common is given by the right member of (1) with g, z, and s_3 replacing t_{12}, y, and s_2, respectively. When this expression is multiplied by the probability that the first and third sets do contain exactly g balls in common and the product is summed on g over the range 0 to t_{12}, we have $P(x, z{:}t_{12}, t_{23})$, the bivariate distribution function of x and z. Thus

$$P(x, z{:}t_{12}, t_{23}) = \sum_{g=0}^{t_{12}} \binom{t_{12}}{g}\binom{s_2 - t_{12}}{t_{23} - g}\binom{s_2}{t_{23}}^{-1} P(x, z{:}g). \tag{6}$$

The mean of the array of x and z for any fixed z may be written, after inverting the order of summation:

$$\bar{x}_z = \sum_{g=0}^{t_{12}} \left\{\left[\binom{s_3}{z}^{-1} q^{-(s_3-z)} p^{-z} \sum_{x=0}^{s_1} xP(x, z{:}g)\right]\binom{t_{12}}{g}\binom{s_2 - t_{12}}{t_{23} - g}\binom{s_2}{t_{23}}^{-1}\right\}. \tag{7}$$

The expression within the square brackets of (7) is identical in form with the right member of (2), and hence we now have

$$\bar{x}_z = \sum_{g=0}^{t_{12}} \left\{[gz/s_3 + p(s_1 - g)]\binom{t_{12}}{g}\binom{s_2 - t_{12}}{t_{23} - g}\binom{s_2}{t_{23}}^{-1}\right\}.$$

This reduces readily to

$$\bar{x}_z = \frac{t_{12}t_{23}}{s_2 s_3} z + \frac{s_1 s_2 - t_{12}t_{23}}{s_2} p. \tag{8}$$

By symmetry,

$$\bar{z}_x = \frac{t_{12}t_{23}}{s_1 s_2} x + \frac{s_2 s_3 - t_{12}t_{23}}{s_2} p.$$

The coefficient of correlation between x and z is found to be

$$r_{xz} = \frac{t_{12}t_{23}}{s_2(s_1 s_3)^{\frac{1}{2}}}. \tag{9}$$

It will be observed that

$$r_{xz} = r_{xy}r_{yz}. \tag{10}$$

Interesting relationships also exist among the partial and multiple correlation coefficients and the multiple regression surfaces. It will be convenient here to measure each variate from its mean and to replace the subscripts x, y, and z, on r by 1, 2, and 3, respectively. Then the multiple regression surface of each variable on the other two may be conveniently expressed in terms of the cofactors of the correlation determinant. From the results found by the writer [4] for the case where each element r_{ij} of the correlation determinant may be expressed as the product $r_{i,i+1}\cdot r_{i+1,i+2}\cdots r_{j-1,j}$, we now have

$$R_{11} = 1 - r_{23}^2, \qquad R_{12} = -r_{12}(1 - r_{23}^2),$$
$$R_{22} = 1 - r_{13}^2, \qquad R_{23} = -r_{23}(1 - r_{13}^2),$$
$$R_{33} = 1 - r_{12}^2, \qquad R_{13} = 0.$$

Then the regression planes of x on y and z and of z on x and y are given, respectively, by

$$x = \frac{r_{12}\sigma_1}{\sigma_2}y = \frac{t_{12}}{s_2}y,$$

$$z = \frac{r_{23}\sigma_3}{\sigma_2}y = \frac{t_{23}}{s_2}y.$$

The regression plane of y on x and z is

$$y = \frac{\sigma_2}{1 - r_{12}^2 r_{23}^2}\left\{\frac{r_{12}(1 - r_{23}^2)}{\sigma_1}x + \frac{r_{23}(1 - r_{12}^2)}{\sigma_3}z\right\}$$
$$= \frac{(s_2^2 s_3 - s_2 t_{23}^2)t_{12}}{s_1 s_2^2 s_3 - t_{12}t_{23}}x + \frac{(s_1 s_2^2 - s_2 t_{12}^2)t_{23}}{s_1 s_2^2 s_3 - t_{12}t_{23}}z.$$

The three multiple correlation coefficients are

$$r_{1\cdot 23} = r_{12}, \qquad r_{3\cdot 12} = r_{23}, \qquad r_{2\cdot 13} = \left[\frac{1 - (1 - r_{12}^2)(1 - r_{23}^2)}{1 - r_{12}^2 r_{23}^2}\right]^{\frac{1}{2}}. \tag{11}$$

The partial correlation coefficients are

$$r_{12\cdot 3} = r_{12}\left[\frac{1 - r_{23}^2}{1 - r_{12}^2 r_{23}^2}\right]^{\frac{1}{2}}, \qquad r_{23\cdot 1} = r_{23}\left[\frac{1 - r_{12}^2}{1 - r_{12}^2 r_{23}^2}\right]^{\frac{1}{2}}, \qquad r_{13\cdot 2} = 0. \tag{12}$$

4. **k variables.** A sequence of k variables may be formed successively as were the three considered above. It will be convenient here to designate the variables by x_i $(i = 1, 2, \cdots k)$. We also define h_i as the total number of balls held in common between the first and the i-th drawings. Then, as special cases, $h_1 = s_1$ and $h_2 = t_{12}$.

The bivariate distribution functions, regression lines, and correlation coefficients associated with any two consecutive variables in the sequence and with any two variables separated by only one other variable can, from the preceding results, be written at once.

It is not difficult to derive the bivariate distribution function for x_1 and x_k by an extension of the method used in deriving (6). We then have

$$P(x_1, x_k : t_{12}, t_{13} \cdots t_{k-1,k})$$

$$= \sum_{h_k} \sum_{h_{k-1}} \cdots \sum_{h_3} \left\{ \prod_{i=3}^{k} \left[\binom{h_{i-1}}{h_i} \binom{s_{i-1} - h_{i-1}}{t_{i-1,i} - h_i} \binom{s_{i-1}}{t_{i-1,i}}^{-1} \right] P(x_1, x_k : h_k) \right\}. \tag{13}$$

The equation of the line of regression of x_1 on x_k is

$$x_1 = \sum_{x_1=0}^{s_1} x_1 P(x_1, x_k : t_{12}, t_{23} \cdots t_{k-1,k}).$$

This may be reduced, by repeated applications of the steps illustrated in the corresponding case for three variable, to the form

$$x_1 = \frac{t_{12} t_{13} \cdots t_{k-1,k}}{s_2 s_3 \cdots s_k} x_k + \frac{s_1 s_2 \cdots s_{k-1} - t_{12} t_{23} \cdots t_{k-1,k}}{s_2 s_3 \cdots s_{k-1}} p. \tag{14}$$

By symmetry, we have

$$x_k = \frac{t_{12} t_{23} \cdots t_{k-1,k}}{s_1 s_2 \cdots s_{k-1}} x_1 + \frac{s_2 s_3 \cdots s_k - t_{12} t_{23} \cdots t_{k-1,k}}{s_2 s_3 \cdots s_{k-1}} p.$$

Then the simple correlation coefficient between x_1 and x_k is

$$r_{1k} = \frac{t_{12} t_{23} \cdots t_{k-1,k}}{s_2 s_3 \cdots s_{k-1} (s_1 s_k)^{1/2}} = r_{12} \cdot r_{23} \cdots r_{k-1,k}. \tag{15}$$

It was shown by the writer [4] that for a sequence such as we are considering the multiple correlation coefficient is a function only of the variables immediately adjacent to the one considered, and that the partial correlation coefficient is zero for any pairs except those of consecutive variables in the sequence. Thus, the formulas given in terms of simple correlation coefficients for the case of a sequence of three variables may be interpreted so as to cover the case for k variables.

REFERENCES

[1] J. C. KAPTEYN, "Definition of the correlation-coefficient," *Monthly Notices Roy. Astron. Soc.*, Vol. 72(1912), pp. 518–525.

[2] H. L. RIETZ, "Urn schemata as a basis for the development of correlation theory," *Annals of Math.* Vol. 21(1920), pp. 306–322.

[3] C. H. FISCHER, "On correlation surfaces of sums with a certain number of random elements in common," *Annals of Math. Stat.* Vol. 4(1933), pp. 103–126.

[4] C. H. FISCHER, "On multiple and partial correlation coefficients of a certain sequence of sums," *Annals of Math. Stat.* Vol. 4(1933), pp. 278–284.