and therefore

$$2(e_2, T, E_2) = 0.$$

Owing to this, E. Čech, who pointed out this circumstance to us, suggested that in the present and in the similar instance for any p, the singular cell be also considered as degenerate. The more extended meaning to be thus attached to degenerate cells, while justifiable, is not however essential.

PRINCETON UNIVERSITY

VARIABLES CORRELATED IN SEQUENCE*

BY A. T. CRAIG

1. *Introduction.* If each of n variables, $x_1, x_2, \cdots, x_n$, represents a quantitative character of an individual, and if the variables are correlated in sequence, that is, x_1 is correlated with x_2, x_2 is correlated with x_3, $\cdots$, and in general x_i is correlated with x_{i+1}, it seems natural to inquire about the correlation between a character, say x_1, of one individual and a character, say x_3, of a second individual, with the condition imposed that the two individuals have identical measurements with regard to the character x_2. It is this problem with which we shall be primarily concerned in the present paper. As we proceed, we shall place appropriate restrictions upon the nature of the correlation which exists between the variables. We shall, however, make no assumptions regarding the correlation between the variables other than that between them in adjacent pairs.

In order to provide a convenient point of departure and to exhibit a set of variables correlated in sequence, we shall first consider a rather elementary problem which arises when measurements are made under a constant law of probability.

2. *The Correlation between Measurements under a Constant Law of Probability.* Let the variable t obey a constant law of probability $f(t) = 1/a$, $0 \leqq t \leqq a$. Let successive sets of n independent measurements each, say $t_1, t_2, \cdots, t_n$, be made upon t. We may, without loss of generality, suppose the meas-

* Presented to the Society, April 8, 1932, under the title *Some properties of correlated variables.*

urements of the sets to be arranged in ascending order of magnitude so that $0 \leqq t_1 \leqq t_2 \leqq \cdots \leqq t_n \leqq a$. If we denote t_p, t_q, t_r, $p \leqq q \leqq r$, by x, y, z respectively, then the simultaneous law of probability of the pth, qth, and rth measurements is given by

$$\begin{aligned}\phi(x, y, z) &= \binom{n}{p-1}\binom{n-p+1}{q-p-1}\binom{n-q+2}{r-q-1}\binom{n-r+3}{n-r}\binom{3}{1}\binom{2}{1} \\ &\cdot\left[\int_0^x f(t)dt\right]^{p-1}\left[\int_x^y f(t)dt\right]^{q-p-1}\left[\int_y^z f(t)dt\right]^{r-q-1} \\ &\cdot\left[\int_z^a f(t)dt\right]^{n-r} f(x)f(y)f(z) \\ &= Kx^{p-1}(y-x)^{q-p-1}(z-y)^{r-q-1}(a-z)^{n-r},\end{aligned}$$

where

$$K = \frac{n!}{(p-1)!(q-p-1)!(r-q-1)!(n-r)!a^n}.$$

The simultaneous laws of probability of the variables taken in pairs are then

$$\begin{aligned}\phi_1(x, y) &= \int_y^a \phi(x, y, z)dz = K_1x^{p-1}(y-x)^{q-p-1}(a-y)^{n-q}, \\ \phi_2(x, z) &= K_2x^{p-1}(z-x)^{r-p-1}(a-z)^{n-r}, \\ \phi_3(y, z) &= K_3y^{q-1}(z-y)^{r-q-1}(a-z)^{n-r},\end{aligned}$$

where

$$K_1 = \frac{n!}{(p-1)!(n-q)!(q-p-1)!a^n},$$

$$K_2 = \frac{n!}{(p-1)!(n-r)!(r-p-1)!a^n},$$

$$K_3 = \frac{n!}{(q-1)!(n-r)!(r-q-1)!a^n}.$$

Further, the laws of probability of x, y, and z are respectively

$$\begin{aligned}\phi_4(x) &= K_4x^{p-1}(a-x)^{n-p}, \\ \phi_5(y) &= K_5y^{q-1}(a-y)^{n-q}, \\ \phi_6(z) &= K_6z^{r-1}(a-z)^{n-r},\end{aligned}$$

where

$$K_4 = \frac{n!}{(p-1)!(n-p)!a^n},$$

$$K_5 = \frac{n!}{(q-1)!(n-q)!a^n},$$

$$K_6 = \frac{n!}{(r-1)!(n-r)!a^n}.$$

For an assigned value of y, the mean of the array of x is

$$\bar{x}_y = \frac{\int_0^y x\phi_1(x, y)dx}{\phi_5(y)} = \frac{py}{q}.$$

Similarly,

$$\bar{y}_x = \frac{(n-q+1)x + a(q-p)}{n-p+1}.$$

Thus of the two variables x and y, the regression of each on the other is linear and the coefficient of correlation is*

$$r_{xy} = \left[\frac{p(n-q+1)}{q(n-p+1)}\right]^{1/2}.$$

By the same procedure, we find

$$r_{yz} = \left[\frac{q(n-r+1)}{r(n-q+1)}\right]^{1/2}, \quad r_{xz} = \left[\frac{p(n-r+1)}{r(n-p+1)}\right]^{1/2}.$$

It may be observed that

$$r_{xz} = r_{xy}r_{yz}.$$

We shall next determine the regression surfaces of each variable on the other two. For assigned values of y and z, the mean of the array of x is

$$\bar{x}_{yz} = \frac{\int_0^y x\phi(x, y, z)dx}{\phi_3(y, z)} = \frac{py}{q}.$$

* See Karl Pearson, *On the mean character and variance of a ranked individual*, etc., Biometrika, vol. 23 (1931), p. 391.

Similarly

$$\bar{y}_{xz} = \frac{z(q-p) + x(r-q)}{r-p},$$

and

$$\bar{z}_{xy} = \frac{y(n-r+1) + a(r-q)}{n-q+1}.$$

Thus the regression surfaces are planes and consequently the multiple and partial correlation coefficients may be computed. Indeed $r_{x\cdot yz} = r_{xy\cdot z} = r_{xy}$.

If we set $u = y - x^*$ and $v = z - y$, then $\psi(u, v)$, the simultaneous probability function of the differences between the qth and pth measurements and the rth and qth measurements, is readily found. We have

$$\psi_1(u, v, z) = \phi(z - u - v, z - v, z), \quad u + v \leqq z \leqq a.$$

Hence,

$$\psi(u, v) = \int_{u+v}^{a} \phi(z - u - v, z - v, z)dz$$

$$= \frac{n!\,u^{q-p-1}(a - u - v)^{n+p-r}}{(q-p-1)!(r-q-1)!(n+p-r)!a^n}.$$

We observe that the surface $w = \psi(u, v)$ is limited by the lines $v = 0$, $u = 0$, $u + v = a$. It follows immediately that the regression of each variable on the other is linear and that

$$r_{uv} = -\left[\frac{(q-p)(r-q)}{(n+p+1-q)(n+q+1-r)}\right]^{1/2}.$$

As a special application of the function $\phi(x, y, z)$ to the statistical theory of sampling, we may take

* The determination of the mean value of the difference between the pth and the $(p+1)$st items is known as the Galton difference problem. This problem has been extensively studied and we cite the following references:

Karl Pearson, *Note on Francis Galton's problem*, Biometrika, vol. 1 (1901–1902), pp. 390–399; *On the probable errors of frequency constants*, Biometrika, vol. 13 (1920), pp. 113–132; *On the mean character and variance of a ranked individual*, Biometrika, vol. 23 (1931), pp. 364–397; vol. 24 (1932), pp. 203–297.

H. L. Rietz, *On a mean difference problem that occurs in statistics*, American Mathematical Monthly, vol. 17 (1910), pp. 235–240.

T. Hojo, *Distributions of medians, quartiles and interquartile distance*, Biometrika, vol. 23 (1931), pp. 315–360.

$$n = 2m + 1, \qquad p = 1, \qquad q = m + 1, \qquad r = 2m + 1$$

and determine the correlation function $F(\xi, W)$ of the range $W = z - x$ and the median $\xi = y$ in samples of $2m+1$ items drawn at random from a universe characterized by a constant law of probability. We then have

$$F(\xi, W) = \int \phi(x, \xi, x + W)dx.$$

The limits of integration are given below for ξ and W in that part of the ξW-plane bounded by the lines indicated at the right:

$\xi - W, \xi$	$W = 0, \quad W = \xi, \quad W = a - \xi;$
$0, \xi$	$\xi = 0, \quad W = \xi, \quad W = a - \xi;$
$0, a - W$	$W = a, \quad W = \xi, \quad W = a - \xi;$
$\xi - W, a - W$	$\xi = a, \quad W = \xi, \quad W = a - \xi.$

Let us now consider the problem of determining the correlation between the pth measurement of one set and the rth measurement of a second set with the condition that the two sets have identical qth measurements. Since x and z are to be matched as to y, we first choose at random a value of y. The probability that y lies in the interval $(y, y+dy)$ is, to within infinitesimals of higher order, $\phi_5(y)dy$. From the array of x corresponding to this value of y, we choose at random a value of x. The probability that x lies in the interval $(x, x+dx)$ is, to a first approximation, $[\phi_1(x, y)/\phi_5(y)]dx$. Finally, from the array of z corresponding to the same value of y, we choose at random a value of z. The probability that z lies in the interval $(z, z+dz)$ is, to a first approximation, $[\phi_3(y, z)/\phi_5(y)]dz$. The probability of the joint occurrence of these events is then the product of the separate probabilities and we have for the simultaneous probability function of x and z when chosen in this manner

$$\bar{\phi}_2(x, z) = \int_x^z \frac{1}{\phi_5(y)} \phi_1(x, y)\phi_3(y, z)dy.$$

Upon performing the integration, we observe that

$$\bar{\phi}_2(x, z) = \phi_2(x, z).$$

Accordingly, if $\bar{r}_{xz}$ be the coefficient of correlation between x and z when they are matched identically with respect to y, then

$$\bar{r}_{xz} = r_{xz} = r_{xy}r_{yz}.$$

If then, we regard x, y, z as three variables correlated in sequence with no information concerning the correlation between x and z, we may say that the coefficient of correlation between the x of one set and the z of another set, the two sets having identical y's, is equal to the product of the coefficients of correlation between x and y and between y and z.

3. *Variables Normally Correlated in Sequence.* Let

$$x_1, x_2, \cdots, x_n$$

be n variables normally correlated in sequence. That is, there are given $n-1$ normal correlation functions

$$F_i(x_i, x_{i+1}) = \frac{1}{2\pi\sigma_i\sigma_{i+1}(1 - r_i^2)^{1/2}} e^{\lambda}, \quad (i = 1, 2, \cdots, n - 1),$$

where

$$\lambda = -\frac{1}{2(1 - r_i^2)}\left\{\frac{x_i^2}{\sigma_i^2} + \frac{x_{i+1}^2}{\sigma_{i+1}^2} - \frac{2r_i x_i x_{i+1}}{\sigma_i\sigma_{i+1}}\right\},$$

and where r_i is the coefficient of correlation between x_i and x_{i+1}. If it is desired, $x_1, x_2, \cdots, x_n$ may be regarded as the pth, qth, rth, $\cdots$, sth measurements in successive sets of N. We do not, however, assume a knowledge of probability functions of more than two variables. For example, we have no function comparable to the function $\phi(x, y, z)$ of §2. Let it be required to determine the correlation between x_{j-1} of one set of measurements and x_{j+1} of another set, the two sets having identical values for x_j. As will appear obvious presently, it is sufficient to consider the variables x_1, x_2, x_3. We first choose at random a value of x_2. For this assigned x_2, we choose at random a value of x_1 and a value of x_3 from the arrays corresponding to the assigned x_2. Then, to a first approximation, the probability that x_2 lies in the interval (x_2, x_2+dx_2), that x_1 lies in the interval (x_1, x_1+dx_1), and that x_3 lies in the interval (x_3, x_3+dx_3) is

$$f_2(x_2)dx_2 \frac{1}{f_2(x_2)} F_1(x_1, x_2)dx_1 \frac{1}{f_2(x_2)} F_2(x_2, x_3)dx_3,$$

where

$$f_2(x_2) = \int_{-\infty}^{\infty} F_1(x_1, x_2)dx_1 = \int_{-\infty}^{\infty} F_2(x_2, x_3)dx_3.$$

Accordingly, the correlation function $\psi_{13}(x_1, x_3)$ of x_1 and x_3 when matched in this manner is given by

$$\psi_{13}(x_1, x_3) = \int_{-\infty}^{\infty} \frac{1}{f_2(x_2)} F_1(x_1, x_2)F_2(x_2, x_3)dx_2$$

$$= \frac{1}{2\pi\sigma_1\sigma_3(1 - r_1^2r_2^2)^{1/2}} e^{\mu},$$

where

$$\mu = -\frac{1}{2(1 - r_1^2r_2^2)}\left\{\frac{x_1^2}{\sigma_1^2} + \frac{x_3^2}{\sigma_3^2} - \frac{2r_1r_2x_1x_3}{\sigma_1\sigma_3}\right\}.$$

Thus the variables are normally correlated with the coefficient of correlation $r = r_1r_2$. If then, the variables x_{j-1} and x_{j+1} are matched with respect to x_j, the coefficient of correlation is $r = r_{j-1}r_j$.

We may now, by making use of $\psi_{13}(x_1, x_3)$ and $F_3(x_3, x_4)$, match in a somewhat analogous manner x_1 and x_4 identically as to x_3 and obtain $\psi_{14}(x_1, x_4)$ which is a normal correlation function with $r = r_1r_2r_3$. Clearly the procedure may be continued so as to involve any two of the n variables.

4. *Variables Non-Normally Correlated in Sequence.* In §2 we considered a special case of variables x, y, z non-normally correlated in sequence, but possessing the property of linearity of regression, and found the coefficient of correlation between the measurement x of one set and the measurement z of another set, the two sets having the same measurement y, to be the product of the coefficients of correlation between x and y and between y and z. We now propose to show that linearity of regression is a sufficient condition that this be true in general. We take the three variables x, y, z correlated in sequence in accord with $F(x, y)$ and $G(y, z)$. We assume the functions to be continuous and the regression of y on x and of z on y to be

linear; that is,

$$\int yF(x, y)dy = (a_1x + b_1)f_1(x)$$

and

$$\int zG(y, z)dz = (a_2y + b_2)f_2(y),$$

where

$$f_1(x) = \int F(x, y)dy,$$

$$f_2(y) = \int F(x, y)dx = \int G(y, z)dz,$$

and a_1, a_2, b_1, b_2 are constants. Then

$$\psi(x, z) = \int \frac{1}{f_2(y)} F(x, y)G(y, z)dy.$$

We have

$$\frac{\int z\psi(x, z)dz}{\int \psi(x, z)dz} = \frac{\int\int z \frac{1}{f_2(y)} F(x, y)G(y, z)dzdy}{f_1(x)}$$

$$= a_1a_2x + a_2b_1 + b_2.$$

Thus the regression of z on x is linear and $\bar{r}_{xz}(\sigma_z/\sigma_x) = a_1a_2$. But $r_{xy}(\sigma_y/\sigma_x) = a_1$ and $r_{yz}(\sigma_z/\sigma_y) = a_2$. Therefore $\bar{r}_{xz} = r_{xy}r_{yz}$.

THE UNIVERSITY OF IOWA