OCTAV ONICESCU
Member of the
Romanian Academy

Informational Correlation

"After all, what matters in the last analysis is the result"

W. Leontief
(*Input-Output Economics*)

INTRODUCTION

The classical correlation of two random quantities involves at the same time the values of these quantities and their probabilities, such as we can see in the expression:

$$\Gamma_{x,\,y} = \sum_{j,\,k} (x_j - \bar{x})\,(y_k - \bar{y})\,P_{jk}$$

where p_{jk} is the probability of having concomitantly $x = x_j$, $y = y_k$. The correlation coefficient is yielded by the formula:

$$I_{x,\,y} = \frac{\Gamma_{x,\,y}}{\sqrt{(\sum_j (x_j - \bar{x})^2 p_j)\,(\sum_k (y_j - \bar{y})^2 q_k)}}$$

where p_j is the probability of event x_j and q_k of event y_k to occur. In connection with these definitions we have also got the ones of the regression of a variable upon the other, as ascertained by means of the conditioned probabilities of these variables.

The usage and the extent to which these concepts can be employed are known.

It is known, for instance, that the correlation may be null without variables x and y being independent, or it may be equal to the unit without the variables being identical.

The most important limitation of using the classical correlation as defined above arises from its involving the representation of the quantities through numerical or vectorial values.

The correlation as defined above loses any meaning if x and y are not real values or vectors.

DEFINING THE INFORMATIONAL CORRELATION

If α_j and $\beta_j (j = 1, 2, ..., n)$ are two groups of characteristics of the same system of objects corresponding in pairs, and $p_j\, q_j$ are the respective probabilities, the theoretical informational correlation is, by definition:

$$C_{\alpha,\,\beta} = \sum_{j=1}^{n} p_j q_j$$

whereas the coefficient of informational correlation is:

$$y_{\alpha,\,\beta} = \frac{\sum_j p_j q_j}{\sqrt{(\sum_j p_j^2)\,(\sum_j q_j^2)}} = \frac{C_{\alpha,\,\beta}}{\sqrt{E_\alpha \cdot E_\beta}}$$

where E_α and E_β are the respective informational energies.

PROPERTIES

We can see at once that:

$$1 - y_{\alpha,\,\beta}^2 = \frac{E_\alpha E_\beta - C_{\alpha,\,\beta}^2}{E_\alpha E_\beta}$$

In agreement with Lagrange's identity we have:

$$E_\alpha E_\beta - C_{\alpha,\,\beta}^2 = (\sum_j p_j^2)\,(\sum q_j^2) - (\sum_j p_j q_j)^2 = \sum_{j,\,k} (p_j q_k - p_k q_j)^2$$

which shows that:

$$1 - y_{\alpha,\,\beta}^2 \geqslant 0$$

the equality to nought being verified only when

$$p_j q_k - p_k q_k = 0$$

for every j and k, which yields, taking into account that $\Sigma p_j = 1$, $\Sigma q_k = 1$,

$$p_j = q_j;\ j = 1, 2, \ldots n$$

Thus we can state:

PROPOSITION I

The coefficient of informational correlation is less than the unit, unless the two distributions are identical.

We can at once verify also the subsequent proposition when the correlation is null:

PROPOSITION II

The correlation between α and β is null if, and only if, I_1 and I_2 being two systems of complementary indexes in the system 1, 2, ..., n, we have $p_j = 0$ for $j \in I_1$ and $q_h = 0$ for $h \in I_2$. The two quantities are alien.

PROPOSITION III

If $p_j (j = 1, 2, ..., n)$ are steady while $q_j (1, 2, ..., n)$ varies, the correlation will vary between a minimum which is the smallest among p_j and a maximum which is the largest among p_j:

$$min\, C_{\alpha,\beta} = min(p_1, p_2, ..., p_n)$$
$$max\, C_{\alpha,\beta} = max(p_1, p_2, ..., p_n)$$

In general, we have:

$$max(min\, p_j,\ min\, q_j) \leqslant C_{\alpha,\beta} \leqslant min(max\, p_j,\ max\, q_j).$$

Indeed, $C_{\alpha,\beta}$ is the mean value of p_j 's after the weights q_j and we also know that the mean value has got the above said properties. If $\min(p_1, q_2, ..., p_n) = p_h$, then this value is achieved assuming $q_h = 1$, $q_j = 0$ for $k \neq h$. If $\max(p_1, p_2, ..., p_n) = p_k$, this value is achieved assuming $q_h = 1$, $p_j = 0$ for $j \neq k$. If all p_j are equal, then, whatever q_j may be we have:

$$C_{\alpha,\beta} = \frac{1}{m}.$$

PROPOSITION IV

If all p_j and all q_j are variable, then the only maximum of $C_{\alpha,\beta}$ is 1 and the only minimum is 0. Maximum is achieved when $p_h = q_h = 1$ for h ascertained but whatever, and $p_j = q_j = 0$ for $j \neq h$.

The minimum is achieved as pointed out above.

THE DELAY CORRELATION

There are instances, for example when index j stands for the age in an increasing order, when we may have a particular interest to correlate not probabilities p_j and $q_j (j = 1, 2, ..., n)$ but p_j and $q_{j+\tau}$ where τ is a delay constant. Thus, for instance, when we try to consider the correlation among the age distributions of the same population for two different years through an interval equal to τ years. In such an instance, the correlation becomes:

$$C^{(k)}_{\alpha,\beta} = \sum_j p_j q_{j+k}$$

and

$$I^{(k)}_{\alpha,\beta} = \frac{C^{(k)}_{\alpha,\beta}}{\sqrt{(\sum_j p_j^2)\ (\sum_j q_{j+k}^2)}}.$$

In this correlation the population considered after k years does not include the newborn in the mean time.

THE MULTIPLE INFORMATIONAL CORRELATION

The informational correlation among more than two characteristics should be symmetrical, the same as for two characteristics. By definition, we have as an example for three characteristics:

$$C_{\alpha,\beta,\gamma} = \sum_j p_j q_j r_j$$

and the correlation coefficient becomes:

$$J_{\alpha,\beta,\gamma} = \frac{\sum_j p_j q_j r_j}{\sqrt[3]{(\sum_j p_j^3)(\sum_j q_j^3)(\sum r_j^3)}}.$$

$J_{\alpha,\beta,\gamma}$ is positive, smaller than the unit, unless $p_j = q_j = r_j$ for all $j = 1, 2, \ldots, n$, which is quite an exception.

$C_{\alpha,\beta,\gamma}$, the same as $J_{\alpha,\beta,\gamma}$ are null when J_α, J_β and J_γ represent a partition of the set of indexes, $1, 2, \ldots, n$ in three distinct parts; we have:

$$p_j = 0,\ j \in J_\alpha,\ q_h = 0,\ h \in J_\beta,\ r_k = 0,\ k \in J_\gamma$$

We prove that:

$$J_{\alpha,\beta,\gamma} = 1$$

if, and only if, $p_j = q_j = r_j$ for all j.

THE INFORMATIONAL CORRELATION BETWEEN TWO DENSITY DISTRIBUTIONS

Let $\rho_1(x)$ and $\rho_2(x)$ be the distribution densities of the two characteristics. This means that

$$\int_a^b \rho_i(x)\,dx$$

is the probability for characteristic i to be marked by a parameter whose values are comprised between a and b.

The coefficinet of informational correlation shall be the limit of the ratio:

$$J_n = \frac{\sum_j \rho_1(x_j)\,\rho_2(x_j)\,(x_{j+1} - x_j)^2}{\sqrt{(\sum \rho_1^2(x_j)\,(x_{j+1} - x_j)^2)\sum \rho_2^2(x_j)\,(x_{j+1} - x_j)^2}}$$

when the rate of division on the real axis tends towards zero.

We assume $x_{j+1} - x_j = \frac{1}{n}$ for any j. In such a case we can reduce this factor at the numerator and the denominator and get:

$$J_n = \frac{\sum_j \rho_1(x_j)\,\rho_2(x_j)\,(x_{j+1} - x_j)}{\sqrt{(\sum \rho_1^2(x_j)\,(x_{j+1} - x_j))\,(\sum \rho_2^2(x_j)\,(x_{j+1} - x_j))}}$$

which yields, at the limit, when $n \to \infty$

$$J = \frac{\int_{-\infty}^{+\infty} \rho_1(x)\, \rho_2(x)\, dx}{\sqrt{\int_{-\infty}^{+\infty} \rho_1^2(x)\, dx \int_{-\infty}^{+\infty} \rho_2^2(x)\, dx}} \tag{1}$$

on condition to prove that the integrals at the numerator and the denominator do exist.

This expression is assumed as a definition of the correlation index.

Taking into account that $\rho_1(x)$ is a distribution density, it is a positive function for which $\int_{-\infty}^{+\infty} \rho_1(x)\, dx$ does exist, having value 1. To this end, it is required that $\rho_1(x) \to 0$, when x tends towards $+\infty$ or towards $-\infty$. This means that having ε, there is a positive a, so that:

$$\rho_1(x) < \varepsilon$$

if $x > a$ or $x < -a$. We see that:

$$1 = \int_{-\infty}^{+\infty} \rho_1(x)\, dx = \int_{-a}^{+a} \rho_1(x)\, dx + \int_{+\infty}^{-a} \rho_1(x)\, dx + \int_{a}^{+\infty} \rho_1(x)\, dx. \tag{2}$$

Similarly, we shall have:

$$\int_{-\infty}^{+\infty} \rho_1^2(x)\, dx = \int_{-a}^{+a} \rho_1^2(x)\, dx + \int_{-\infty}^{-\infty} + \int_{a}^{+\infty}. \tag{3}$$

The first integral is finite; at the same time

$$\int_{-\infty}^{-a} \rho_1^2(x)\, dx < \varepsilon \int_{-\infty}^{-a} \rho_1(x); \quad \int_{a}^{\infty} \rho_1^2(x)\, dx < \varepsilon \int_{a}^{\infty} \rho_1(x).$$

Consequently, taking into account (2):

$$\int_{-\infty}^{-a} \rho_1^2(x)\, dx < \varepsilon \int_{-\infty}^{-a} \rho_1(x); \quad \int_{a}^{\infty} \rho_1^2(x)\, dx < \varepsilon \int_{a}^{\infty} \rho_1(x)$$

which shows that integral $\int_{-\infty}^{+\infty} \rho_1^2(x)\, dx$ is finite. A similar proof can be achieved for $\int_{-\infty}^{+\infty} \rho_1(x)\, \rho_2(x)\, dx$. By this formula, (1) is fully justified.

Remark 1. A continuos distribution has, as a rule, no density. However, it can be approximated rightly by a distribution in respect of density. This remark leads us without any difficulty to the following definition of the correlation coefficient of two continuous characteristics:

If $\rho_1^n(x)$ and $\rho_2^n(x)$ are the densities of distributions $F_1^{(n)}(x)$ and $F_2^{(n)}(x)$ approximating the continuous distributions $F_1(x)$, $F_2(x)$, the coefficient of approximate correlation of these ones is, by definition:

$$y^{(n)} = \frac{\int_{-\infty}^{+\infty} \rho_1^{(n)}(x)\, \rho_2^{(n)}(x)\, dx}{\sqrt{\int_{-\infty}^{+\infty} \rho_1^{(n)2}(x)\, dx \left(\int_{-\infty}^{+\infty} \rho_2^{(n)2}(x)\, dx \right.}}.$$

Remark 2. A distribution has, as a rule, a discontinuous part and a continuous part. The discontinuous part may be, at most, a numberable set of states having the respective probabilities $p_1, p_2, \ldots, p_n, \ldots$

The correlation between two numberable distribution is defined by the ever convergent sum $\sum p_j q_j$, for the correlation coefficient by making use also of the informational energies which also are always convergent. $\sum_j p_j^2$ and $\sum q_j^2$.

Remark 3. The correlation between two distribution of distinct lengths is achieved by grouping the respective states into cumulative states in equal numbers and by establishing a significant correspondence among them.

Theoretical examples. The coefficient of informational correlation between two Gaussian distributions

$$\rho_1(x) = \frac{h_1}{\sqrt{\pi}} e^{-h_1^2 x^2}; \quad \rho_2 = \frac{h_2}{\sqrt{\pi}} e^{-h_2^2 x^2}$$

is, according to formula (1):

$$y = \frac{2h_1 h_2}{h_1^2 + h_2^2}$$

which, in general, is smaller than the unit, equal to the unit only if $h_1 = h_2$, and equal to 0 when h_1 or h_2 are nought, which means that one of the densities is null.

The delay correlation between two characteristics with densities $\rho_1(x)$ and $\rho_2(x)$ is, according to the definition,

$$y = \frac{\int_{-\infty}^{+\infty} \rho_1(x)\, \rho_2(x-a)\, dx}{\sqrt{\left(\int_{-\infty}^{+\infty} \rho_1^2(x)\, dx\right)\left(\int_{-\infty}^{+\infty} \rho_2^2(x)\, dx\right)}}$$

and, in particular, if

$$y_a = \frac{\int_{-\infty}^{+\infty} \rho(x)\, \rho(x-a)\, dx}{\int_{-\infty}^{+\infty} \rho^2(x)\, dx}.$$

If $\rho(x)$ is Gaussian

$$\rho(x) = \frac{h}{\sqrt{\pi}} e^{-h^2 x^2}$$

we find for this delay correlation the value:

$$y_a = e^{-\frac{a^2}{2}}.$$

Consider Σ to be a system of elements having the characteristics C_1, C_2 ..., C_n weighted with values $p_1, p_2, \ldots, p_n$.

A random selection $\Sigma^{(s)}$ of Σ has the weights $a_1^{(s)}, a_2^{(s)}, \ldots, a_n$ for the n characteristics to be found in $\Sigma^{(s)}$; the correlation coefficient between $\Sigma^{(s)}$ and Σ is:

$$y^{(s)} = \sum_{j=1}^{n} \pi_j \omega_j^{(s)} \tag{1}$$

where we have assumed:

$$\pi_j = \frac{p_j}{\sqrt{\sum_h p_h^2}}; \quad \omega_j^{(s)} = \frac{\alpha_j^{(s)}}{\sqrt{\sum_j \alpha_j^{(s)2}}}. \tag{2}$$

The expression

$$d_{\pi, \omega^{(s)}} = 1 - (\sum_j \pi_j \omega_j^{(s)})^2 \tag{3}$$

measures the quasi-distance between point $\omega^{(s)}$ of components $\omega_1^{(s)}, \omega_2^{(s)}, \ldots, \omega_n^{(s)}$ and point π of components $\pi_1, \pi_2, \ldots, \pi_n$, both of them on the sphere

$$x_1^2 + x_2^2 + \ldots + x_n^2 = 1.$$

Distance $d_{\pi, \omega^{(s)}}$ is minimal and equal to zero if $\pi = \omega^{(s)}$.

Here arises the problem of inferring distribution π of system Σ, knowing a number k of random selections $\Sigma^{(s)}$.

The characterization is achieved in the following manner:

It is distribution π to render minimal the mean value of distances $d_{\pi, \omega^{(s)}}$ that is to say:

$$\frac{1}{k} \sum_{s=1}^{k} d_{\pi, \omega^{(s)}} = 1 - \frac{1}{k} \sum_{s=1}^{k} (\sum_{j=1}^{n} \pi_j \omega_j^{(s)})^2.$$

This means that π must render maximal the expression:

$$e = \sum_{s=1}^{k} (\sum_{j=1}^{n} \pi_j \omega_j^{(s)})^2 \tag{4}$$

taking into account that

$$\sum_{h=1}^{n} \pi_h^2 = 1.$$

That one or more maxima of expression e do exist, it follows from the remarks that the set of values of expression e, when π assumes all the possible systems of values, is bounded and continuous

$$e \leqslant k.$$

In order to find this maximum we have to meet the following two equalities:

$$\sum_{s=1}^{k} \left(\sum_{j=1}^{n} \pi_j \omega_j^{(s)}\right) \left(\sum_{h=1}^{n} \omega_h^{(s)} \delta\pi_h\right) = 0 \tag{5}$$

$$\sum_h \pi_h \delta\pi_h = 0$$

which, by introducing Lagrange's parameter λ yields:

$$\sum_{j=1}^{n} \pi_j \left(\sum_{s=1}^{k} \omega_j^{(s)} \omega_h^{(s)}\right) = \lambda\pi, \tag{6}$$

$$h = 1, 2, ..., n.$$

If we assume

$$\beta_{jh} = \sum_{s=1}^{k} \omega_j^{(s)} \omega_h^{(s)} \tag{7}$$

system (6) becomes

$$\sum_j \beta_{jh} \pi_j = \lambda\pi_h \tag{8}$$

$$h = 1, 2, ..., n.$$

Considering the significations of components π_j yielded by formulas (2), equation (8) can also be written:

$$\sum_j \beta_{jh} p_h = \lambda p_h \tag{9}$$

having a direct bearing on the unknown probabilities $p_1, p_2, ..., p_n$. This system has got real solutions each time that λ is a real root of the characteristic equation:

$$\begin{aligned} &|\beta_{ih} - \lambda\delta_{jh}| = 0 \\ &j, h = 1, 2, ..., n. \end{aligned} \tag{10}$$

We see that, taking into account expressions (7), we have:

$$\sum_{j,k} \beta_{jh} x_j x_h = \sum_{s=1}^{k} \left(\sum_j \omega_j^{(s)} x_j\right)^2$$

consequently β_{ih} are the symmetrical coefficients of a positive square form and it follows that the roots of characteristics equation (10) are real.

However, we are interested solely in the positive roots, as seen in (9). Moreover, we hold only those roots for which the solutions to equations (9) are all positive.

These solutions are determined outwards by an arbitrary multiplier, taking into account condition:

$$p_1 + p_2 + ... + p_n = 1.$$

If there are several solutions, we choose the one for which expression (4) is maximal.

Remark: It has been required that selections $\Sigma^{(s)}$ should be random. A proof of meeting this condition and at the same time an explanation for it is that vectors $\alpha^{(s)}$ should assume a Gaussian distribution for the diverse values of s. We may leave aside the selections which wander from this distribution. They should but disturb the operation of ascertaining the structure of Σ.

Example. We make up 10 series of 100 tossings of a coin and get the following results:

$\Sigma^{(1)}$58	head	42	tail	$\Sigma^{(6)}$51	„ 49 tail
$\Sigma^{(2)}$54	„	46	„	$\Sigma^{(7)}$49	„ 51 „
$\Sigma^{(3)}$45	„	55	„	$\Sigma^{(8)}$47	„ 53 „
$\Sigma^{(4)}$49	„	51	„	$\Sigma^{(9)}$52	" 48 „
$\Sigma^{(5)}$40	„	60	„	$\Sigma^{(10)}$50	„ 50 „

In this case we have:

$$\omega_1^{(1)2} = 0.65; \quad \omega_1^{(2)2} = 0.58; \quad \omega_1^{(3)2} = 0.40;$$
$$\omega_1^{(4)2} = 0.48; \quad \omega_1^{(5)2} = 0.307 \quad \omega_1^{(6)2} = 0.58;$$
$$\omega_1^{(7)2} = 0.48; \quad \omega_1^{(8)2} = 0.44; \quad \omega_1^{(9)2} = 0.54;$$
$$\omega_1^{(10)2} = 0.50.$$

$$\omega_2^{(1)2} = 0.35; \quad \omega_2^{(2)2} = 0.42; \quad \omega_2^{(2)2} = 0.60;$$
$$\omega_2^{(4)2} = 0.52; \quad \omega_2^{(5)2} = 0.693; \quad \omega_2^{(6)2} = 0.48;$$
$$\omega_2^{(7)2} = 0.52; \quad \omega_2^{(8)2} = 0.56; \quad \omega_2^{(9)2} = 0.46;$$
$$\omega_2^{(10)2} = 0.50.$$

$$\omega_1^{(1)}\omega_2^{(1)} = 0.476; \quad \omega_1^{(2)}\omega_2^{(2)} = 0.47; \quad \omega_1^{(3)}\omega_2^{(3)} = 0.49;$$
$$\omega_1^{(4)}\omega_2^{(4)} = 0.49; \quad \omega_1^{(5)}\omega_2^{(5)} = 0.46; \quad \omega_1^{(6)}\omega_2^{(6)} = 0.49;$$
$$\omega_1^{(7)}\omega_2^{(7)} = 0.49) \quad \omega_1^{(8)}\omega_2^{(8)} = 0.49; \quad \omega_1^{(9)}\omega_2^{(5)} = 0.48;$$
$$\omega_1^{(10)}\omega_2^{(10)} = 0.50;$$

which yields:

$$\beta_{11} = 4.897, \ \beta_{12} = 4.836, \ \beta_{22} = 5.163$$

The characteristic equation has the roots $\lambda_1 = 9.43$ and $\lambda_2 = 0.63$. For $\lambda_1 = 9.43$ we get:

$$p_1 = \frac{1}{1.93} = 0.51; \ p_2 = 0.49$$

probabilities corresponding to the coin tossed cassually. For $\lambda_2 = 0.63$ we get $p_1 > 1$, which is not a probability.

We consider the quasi-distance:

$$\delta_{\pi,\,\omega^{(s)}} = 1 - \sum_j \pi_j \omega_j^{(s)}. \tag{1}$$

The mean value:

$$\frac{1}{k}\sum_s \delta_{\pi,\,\omega^{(s)}} = 1 - \frac{1}{k}\sum_{s,\,j} \pi_j \omega_j^{(s)} \tag{2}$$

to be minimized leads to the finding of values π_j, verifying condition

$$\sum_j \pi_j^2 = 1 \tag{3}$$

for which $\sum_{s,\,j} \pi_j \omega_j^{(s)}$ is maximal. Then we have relationships:

$$\sum_j \left(\sum_s \omega_j^{(s)}\right) \delta\pi_j = 0; \quad \sum_j \pi_j \delta\pi_j = 0 \tag{4}$$

which are met if:

$$\pi_j = \lambda \sum_s \omega_j^{(s)} \tag{5}$$

value of λ is inferred by satisfying condition (3) which

$$\lambda = \frac{1}{\sqrt{\sum_j \left(\sum_s \omega_j^{(s)}\right)^2}}.$$

Consequently:

$$\pi_j = \frac{\sum_s \omega_j^{(s)}}{\sqrt{\sum_h \left(\sum \omega_h^{(s)}\right)^2}}. \tag{6}$$

Hence, taking into account expressions

$$\pi_j = \frac{p_j}{\sqrt{\sum_h p_h^2}} \; (j = 1, 2, \ldots, n)$$

we get:

$$\frac{p_j}{\sqrt{\sum_h p_h^2}} = \frac{\sum_s \omega_j^{(s)}}{\sqrt{\sum_h \left(\sum_s \omega_h^{(s)}\right)^2}} \tag{7}$$

Totalizing in both members as against j we reach the equality

$$\frac{1}{\sqrt{\sum_h p_h^2}} = \frac{k}{\sqrt{\sum_h \left(\sum_s \omega_h^{(s)}\right)^2}} \tag{8}$$

By dividing member by member (7) by (8), there follows the expression of the looked for probabilities:

$$p_j = \frac{1}{k} \sum_{s=1}^{k} \omega_j^{(s)} \tag{9}$$

THE INDEX OF THE DELAY INFORMATIONAL CORRELATION FOR THE POPULATION IN 1956

Writing i for the order index of the age group comprised between 0 and over 100 years, q_1 for the frequency of the citizens of the age group i in the year 1956, and q_i for the frequency of the citizens of the age group i in the year 1966, we have:

$$I_{\alpha\beta} = \frac{\sum_{i=1}^{21} q_1 p_{i+2}}{\sqrt{(\sum_{i=1}^{21} q_i^2)\ (\sum_{i=1}^{21} p_{i+2}^2)}} = \frac{0.066072}{0.068789} = 0.960502.$$

THE INDEX OF THE INFORMATIONAL CORRELATION BETWEEN THE POPULATIONS IN THE YEARS 1956 AND 1966:

$$I_{\alpha\beta} = \frac{\sum_{i=1}^{25} p_i q_i}{\sqrt{(\sum p_i^2)(\sum q_i^2}} = \frac{0.71558}{0.072436} = 0.987878.$$

LA CORRELATION INFORMATIONNELLE

L'auteur introduit la „corrélation informationnelle" de deux caractéristiques de la même population par l'expression $\sum_{j=1}^{n} p_j q_j$ où $p_1 p_2, \ldots p_n$ et $q_1 q_2, \ldots q_n$ représentent, par exemple, les répartitions par n groupes d'âge de la population d'un pays en deux années différentes où encore $p_1 p_2, \ldots p_n$ représentent selon un critiérium bien défini la répartition d'un système S d'objets et $q_1 q_2, \ldots q_n$ la répartition correspondante d'une sélection.

Il étudie les propriétés de la corrélation et du coefficient et donne une application au problème de la sélection.

ИНФОРМАЦИОННАЯ КОРРЕЛЯЦИЯ

Автор вводит информационную корреляцию между двумя характеристиками одного и того же населения при помощи выражения $\sum_{j=1}^{n} p_j q_j$, где $p_1, p_2, \ldots, p_n$ и $q_1, q_2, \ldots, q_n$ представляют собой, например, распределение по n возрастным группам населения какой-либо страны в два различных года или $p_1, p_2, \ldots, p_n$ представляют собой, согласно определенному критерию, распределение систем предметов S и $q_1, q_2, \ldots, q_n$ — распределение селекции.

Автор рассматривает свойства корреляции и коэффициента корреляции и приводит пример применения селекции.