Technoeconomics. 2023, Vol. 2, No. 2 (5). Pp. 15–35.
Техноэкономика. 2023, Том 2, № 2 (5). С. 15–35.

Scientific article
UDC 330.47
DOI: https://doi.org/10.57809/2023.2.2.5.2

ON THE PROBLEM OF NOMINAL DATA CORRELATION

Sergey Svetunkov ✉

Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia

✉ sergey@svetunkov.ru

Abstract. This article suggests a new methodological approach to calculating the correlation coefficient between nominal data. In the course of the study, it was revealed that most of the work with nominal data uses two Yule coefficients and Pearson coefficient. Moreover, these coefficients are proposed based on the same assumption that there is no correlation between the variables. As a result of the study, this assumption is questioned by the authors, and a new coefficient of correlation between nominal data is proposed. A comparative analysis of the application of the three classical coefficients with the new coefficient is conducted. The advantages of the new coefficient are shown.

Keywords: correlation coefficient, nominal data, Yule coefficient, Pearson coefficient, correlations

Citation: Svetunkov S.G. On the problem of nominal data correlation. Technoeconomics. 2023. 2. 2 (5). 15–35. DOI: https://doi.org/10.57809/2023.2.2.5.2

Научная статья
УДК 330.4
DOI: https://doi.org/10.57809/2023.2.2.5.2

К ВОПРОСУ О КОРРЕЛЯЦИИ НОМИНАЛЬНЫХ ДАННЫХ

Сергей Светуньков ✉

Санкт-Петербургский политехнический университет Петра Великого,
Санкт-Петербург, Россия

✉ sergey@svetunkov.ru

Аннотация. В данной статье представлен новый методический подход к расчету коэффициента корреляции между номинальными данными. В процессе исследования выявлено, что в большинстве случаев работы с номинальными данными используют два коэффициента Юла и один коэффициент Пирсона. Причем эти коэффициенты предлагаются исходя из одного и того же предположения об отсутствии корреляционных связей между переменными. В результате исследования данное предположение ставится авторами под сомнение и предлагается новый коэффициент корреляции между номинальными данными. Проводится сравнительный анализ применения трёх классических коэффициентов с новым коэффициентом и показываются преимущества нового коэффициента.

Ключевые слова: коэффициент корреляции, номинальные данные, коэффициент Юла, коэффициент Пирсона, корреляционные связи

Для цитирования: Светуньков С.Г. К вопросу о корреляции номинальных данных // Техноэкономика. 2023. Т. 2, № 2 (5). С. 15–35. DOI: https://doi.org/10.57809/2023.2.2.5.2

Introduction

In this scientific research, correlation analysis tools are used to confirm the presence of the cause-and-effect relationships. Due to the originality of the nominal data, the study of their correlation differs significantly from the studies in other sections of the correlation analysis. At the same time, despite the numerous efforts of scientists to develop and improve this section of correlation analysis, it still does not provide a satisfactory solution to the problem of identifying and evaluating correlation. The difficulty in determining the correlation between data measured in nominal data is explained by the fact that no mathematical operations can be performed on these data. Occurrences of some numbers are already the data of the metric scale, and these data can be processed statistically. The number of occurrences of nominal numbers is used to judge whether or not there is a relationship between the nominal numbers.

Materials and Methods

The most convenient way to work with the data on the number of occurrences of nominal numbers is to put them into a table, which is commonly called a "conjugacy table" (Table 1).

Table 1. General vision of the conjugacy table

	The first characteristic of attribute A, x_1	The second characteristic of attribute A, x_2	Total
The first characteristic of attribute B, y_1	a	b	a+b
The second characteristic of attribute B, y_2	c	d	c+d
Total	a+c	b+d	N=a+b+c+d

Attribute A in the conjugacy table is, for example, the sex of a person. And attribute B in the same table is a preference between lipstick (the first characteristic) and strong alcohol (the second characteristic) a, b, c, d are the numbers of observations on the conjugate features, e.g. a is the number of women who stop to look at the lipstick counter with interest and b is the number of men who do the same. The researcher's task is to infer from the observations of the numbers a, b, c, d whether or not there is a correlation between the attributes A and B. This is not an easy task (Cramer, 1946).

If there is some correlation between the attributes A and B, it manifests itself in certain proportions between the numbers a, b, c and d. But no one knows this proportion a priori. Furthermore, these proportions are different for different properties. Therefore, the task is to try to describe this proportion with the help of some tool and to evaluate the strength of its manifestation, or in other words, to measure the strength of the correlation between the attributes. How can one generally determine the presence or absence of a correlation between the nominal data? Let us give some description of such situations, using the most general idea of the presence or absence of a correlation (Yates, 1934).

There is no correlation if a change in one attribute no effect has whatsoever on another attribute - it remains unchanged. For example, attribute A is the sex of the TV viewer interviewed in Russia, and attribute B is his attitude towards the two parties in Honduras: y1 is the Liberal Party of Honduras and y2 is the National Party of Honduras. Since ordinary Russian viewers have no idea about the political system of Honduras and will express their attitude towards them not by their characteristics but by their names, there is no correlation between the attributes in this case (Reagle, Vinod). Some of the respondents will like the word "Liberal", others – "National" (Boon, 2020). Therefore, this is the type of data most likely to be obtained in this case.

Table 2. An example of a situation where there is no relationship in the conjugacy table

	Male TV viewers, x_1	Female TV viewers, x_2	Total
Preference for the Liberal Party, y_1	149	101	250
Preference for the National Party, y_2	151	99	250
Total	300	200	500

Here, there is no correlation, as a relationship between two random factors, and the coefficient reflecting this situation should be equal to zero. The lack of correlation in the nominal data manifests itself in the fact that when the characteristics of one attribute A change, the characteristics of another attribute B do not change.

The most common coefficient in practice today, with the help of which a researcher tries to assess the strength of the correlation between two groups of nominal data, is Yule's association coefficient:

$$Q = \frac{ad - bc}{ad + bc} \quad (1)$$

The main idea for substantiating the form of this coefficient, which Yul outlined in an article in 1912, is as follows: "If the two attributes are combined entirely independently, the proportion that possesses, say, the first character will be the same, or more or less approximately the same, amongst those which possess and those which do not possess the second. If these two proportions differ, the two attributes are not independent but associated: positively associated if the proportion possessing the first character is greater amongst the objects or individuals possessing the second character than amongst those not possessing it, negative in the contrary case". Mathematically, this idea, taking into account the notation of Table 1, will be written as follows:

$$\frac{a}{a+c} = \frac{b}{b+d} \quad (2)$$

Whence:

$$ab + ad = ab + cb \rightarrow ad - cb = 0 \quad (3)$$

That is, if there is no relationship between the attributes, as Yule understood it, then the right-hand side of (3) will be equal to zero. And if the relationship ("association" - according to Yule) exists, then the equality to zero is violated. In this case, the difference (ad-cb) will be greater or less than one (Thompson, 2019).

In order to transform this condition into a computationally friendly coefficient that varies modulo from zero to one, Yule divided the right-hand side of (3) by its conjugate value and obtained formula (1).

Exactly the same result can be obtained if the proportions are calculated not by columns, but by rows, because:

$$\frac{a}{a+b} = \frac{c}{c+d} \rightarrow ac + ad = ac + cb \rightarrow ad - cb = 0 \quad (4)$$

One can make sure that the right part (3) is the numerator (1).

While explaining the reason for the fact that he designated the new coefficient with the letter Q, Yule wrote that: "I took the symbol from the first letter of Quetelet". At the time of Yule, the name Lambert-Adolph-Jacques Quetelet, as one of the founders of statistics, was widely known all over the world.

For the convenience of studying the properties of the association coefficient (1), Yul transformed it into this form:

$$Q = \frac{ad - bc}{ad + bc} = \frac{1 - \frac{bc}{ad}}{1 + \frac{bc}{ad}} = \frac{1-k}{1+k} \quad (5)$$

He went on to introduce such designations:

$$p_0 = \frac{a}{a+c}; \; p_1 = \frac{b}{b+d} \quad (6)$$

Taking into account the coefficient k, these two components can be written as:

$$p_0 = \frac{1}{1+\sqrt{k}}; \; p_1 = \frac{\sqrt{k}}{1+\sqrt{k}} \quad (7)$$

Their difference:

$$\omega = p_0 - p_1 \tag{8}$$

is also some measure of the relationship: "...why not use ω itself as the coefficient of association instead of the function Q?".

He called this coefficient ω the coefficient of colligation. It is more convenient to use this coefficient in a form approximating the form of the association coefficient (1) with those variables used in the coefficient of association. This can be easily done by substituting for k its values taken from (6):

$$\omega = \frac{1-\sqrt{k}}{1+\sqrt{k}} = \frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}} \tag{9}$$

Obviously, the condition $|\omega| \le 1$ is met for this coefficient as well.

As can be seen, both Yule`s coefficients are based on the assumption that there is no relationship between the two attributes X and Y only if there is a condition of proportion (2) or (3) between them. In all other cases, these coefficients will be modulo greater than zero (Deisenroth, Ong, Faisal).

In the previous paragraph, we determined that the lack of correlation between the attributes means the absence of any influence at all from X on Y and vice versa.

This means that either X or Y must be evenly distributed in the conjugacy table, as shown in Table 2. If there is a proportional distribution, then this indicates that there is some established relationship between the attributes. Another thing is that this relationship does not change with the change of the attributes, but it is there (Zudin, 2023).

Consequently, there is still a relationship between the two attributes X and Y, although not significant.

Using the example of Table 2 in the previous paragraph, we have examined the case where there is no correlation between the factors. Let us now apply both of Yule's coefficients to this table, knowing that it simulates a non-correlation situation. We obtain:

$$Q = \frac{70\cdot 100 - 100\cdot 230}{70\cdot 100 + 100\cdot 230} = -0,53; \quad \omega = \frac{\sqrt{70\cdot 100}-\sqrt{100\cdot 230}}{\sqrt{70\cdot 100}+\sqrt{100\cdot 230}} = -0,29$$

The association coefficient shows the presence of a relationship, and the coefficient of colligation indicates that there is a relationship, but rather a weak one. The discrepancy in the readings of these two coefficients should not be surprising. It is a well-known fact that Yule`s coefficient of association modulo is always greater than the coefficient of colligation of Yule.

And the fact that these two coefficients reveal a correlation where, in our opinion, it cannot be, is also not surprising, because both coefficients proceed from the fact that there is no relationship if and only if there is some invariable proportion between the numbers of columns (and rows).

In practice Pearson's mutual conjugacy coefficient is less often used. However, it has found the most widespread use in the scientific environment since, unlike Yule's coefficients, it is obtained by using a statistical distribution. Taking into account the designations we use; Pearson's conjugacy coefficient will take this form:

$$\phi = \frac{ad-bc}{\sqrt{(a+c)(b+d)(a+b)(c+d)}} \tag{10}$$

K. Pearson is one of the founders of mathematical statistics and his contribution to this science is enormous. In particular, he proposed and carefully studied the x^2 distribution which is known today in mathematical statistics:

$$\chi^2 = \sum_i \frac{(y_i - y_i^{\wedge})^2}{y_i^{\wedge}} \quad (11)$$

Pearson suggested using the x^2 distribution to identify the correlation between the nominal data. The essence of his proposal is as follows. In the conjugacy table there are some real data. They need to be compared with such calculated values where there is no relationship between the attributes (Edwards, David).

These calculated values are most often called "theoretical". Having these two groups of data, it is possible to calculate (11).If the real values coincide with the "theoretical" ones or are close to them, then x^2 will be equal to or close to zero.

The obtained value of x_2 can be compared with the critical value (from the table, which is available in all textbooks on mathematical statistics) and if the calculated value of x_2 exceeds the critical value, it indicates that the assumption of no association is not true.

The real values of the numbers are available in the original conjugacy table. These are values a, b, c, and d. But how can we find unknown "theoretical values" at which there is no any relationship?

To do this, one can use Yule's suggestion (2).

Let b/a = k (k>0). Since there is no relationship according to Yule when (2) is met, then by substituting b = ak into it, we obtain that the following equality should be met: d = ck. Or: b/a = d/c = k. Let us substitute these values in the conjugacy table.

Table 3. General view of the conjugacy table

	x_1	x_2	Total
y_1	a	ak	a(1+k)
y_2	c	ck	c(1+k)
Total	a+c	(a+c)k	N=(a+c)(1+k)

Suppose now we know only the final rows and columns of the conjugacy table 5, and the values inside it are not known to us.

Table 4. Conjugacy table 5 in the absence of internal numbers

	x_1	x_2	Total
y_1			a(1+k)
y_2			c(1+k)
Total	a+c	(a+c)k	N=(a+c)(1+k)

Under these conditions, how can one calculate "theoretical values" when there is no relationship? To do this, the simple proportion shown in Table 5 should be used.

Table 5. Simple proportion in the conjugacy

	x_1	x_2	Total
y_1	?		a+b
y_2			c+d
Total	a+c	b+d	N

Unknown or "theoretical" elements of the table, with the help of this simple proportion, will be found as follows:

$$a' = \frac{(a+b)(a+c)}{N}, \quad b' = \frac{(a+b)(b+d)}{N}, \quad c' = \frac{(c+d)(a+c)}{N}, \quad d' = \frac{(c+d)(b+d)}{N} \tag{12}$$

Substituting the total columns and rows of Table 6 into (12), we obtain that in the absence of a relationship, as Yule understood it, we obtain that a'=a, b'=b, c'=c, d'=d. That is, now for any values of the numbers in the conjugacy table, those very "theoretical" values for which Yule`s coefficients will be zero, can always be calculated using the total values of rows and columns, which, according to Yule, indicates a lack of correlation.

The distribution of x_2 in relation to the case under consideration will be written as follows:

$$\chi^2 = \frac{(a-a')^2}{a'} + \frac{(b-b')^2}{b'} + \frac{(c-c')^2}{c'} + \frac{(d-d')^2}{d'} \tag{13}$$

Let us show how this method works.

Table 6. Conditional example

	x_1	x_2	Total
y_1	34	66	100
y_2	88	62	150
Total	122	128	250

According to the total values of this table, "theoretical" values can be calculated.

Table 7. Calculation of "theoretical" values according to Table 6

	x_1	x_2	Total
y_1	48,8	51,2	100
y_2	73,2	76,8	150
Total	122	128	250

The value of the criterion x^2 (13) in this case is calculated by the formula:

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(y_{ij} - y\!\!\!E_{ij})^2}{y\!\!\!E_{ij}} \tag{14}$$

where y_{ij}- is the real value of the indicator, located in the i row and in the j column, $y\!\!\!E_{ij}$ is its "theoretical" value.

Substitute in formula (14) the real values from Table 8 and the "theoretical" values from Table. 9. We obtain:

$$\chi^2 = \frac{(34-48,8)^2}{48,8} + \frac{(66-51,2)^2}{51,2} + \frac{(88-73,2)^2}{73,2} + \frac{(62-76,8)^2}{76,8} = 14,649 \tag{15}$$

Let us compare the value of the criterion x^2 with the critical values. It is equal to 3.841 with a significance level of 0.05. Our value (15) exceeds significantly the critical value, so it can be argued that there is correlation between the nominal numbers Table. 7. But how close is this correlation? The answer to this question cannot be derived from (15), but the researcher is interested not only in the fact that there is a relationship between the attributes, but also in the degree of this relationship. And it is impossible to determine it by analyzing the data obtained.

Therefore, based on (13), Pearson proposed a coefficient that modulo will vary from minus one to plus one:

$$\phi = \sqrt{\frac{\chi^2}{N}} \tag{16}$$

Substituting in (17) the values of x^2 and the number of observations N, we obtain:

$$\phi = \sqrt{\frac{14,649}{250}} = 0,242 \quad (17)$$

It follows from (16) that in the case where the real values coincide with the "theoretical" ones and x^2 is equal to zero, Pearson's conjugacy coefficient will be also equal to zero. And for $x^2 \rightarrow \infty$ it tends modulo to one. That is, the requirements for the limits of the correlation coefficient change are satisfied here: in the absence of correlation, the coefficient is equal to zero, and in the presence of strong correlation, it tends to one.

Since the Pearson correlation coefficient for the case in question turned out to be insignificant φ =0,242, it should be stated that this coefficient diagnoses a weak relationship between the two attributes. But this multi-iterative approach is not very convenient for practical application. To calculate Yule's coefficient, simply substitute the values from the conjugacy table into his formula and the result is immediately available. It is necessary to simplify the calculation of Pearson's conjugacy coefficient. And this can be done.

If we now replace x^2 in (16) with its value from (13) and, in its turn, replace (13) with the "theoretical" values as determined from (12), we obtain a formula suitable for calculations. Once this has been done, by reducing and grouping, it is possible to obtain a formula suitable for calculating Pearson's conjugacy coefficient (10). Without claiming that his coefficient is better than Yule's one, Pearson gives an example of an evaluation of vaccination effectiveness, the one Yule had previously used. "Taking the small-pox returns for the epidemic of 1890, we have".

Table 8. Example of K. Pearson

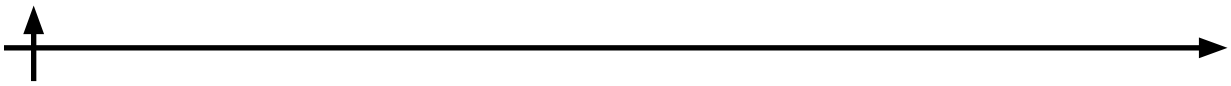
	Recoveries	Deaths	Total
Present	1562	42	1604
Absent	383	94	477
Total	1945	136	2081

For the data in this table, Yule's association coefficient will be 0.803, and Pearson's conjugacy coefficient will be 0.29. The first coefficient indicates that there is a strong correlation between the attributes, and the second coefficient indicates that if there is a correlation, it is very weak (Wu, Gan, Ma, 2007).

Having received this conclusion, Pearson pointed out that the comparison of these two coefficients should be carried out on a large number of examples, and only then would it be possible to conclude which of the two coefficients should be preferred. In cases where the numbers in the conjugacy tables are added so that the proportions (2) or (12) are satisfied, then both coefficients will be equal to zero (Agarwal, 2006). But, as follows from (2), if at least one of the numbers in the conjugacy table is zero, then both Yule`s coefficients (1) and (9) will be equal to one, regardless of whether there is a relationship between the attributes or not. They will be close to one, even if one of the numbers in the conjugacy table is extremely small, compared to the rest of the numbers. The example in Table 11 illustrates this peculiarity.

Table 9. Conditional example

	x_1	x_2	Total
y_1	1	5000	20
y_2	200	100000	5001
Total	201	105000	100200



For this table we have: Q = -0.8182, ω = -0.5195 and φ = -0.0088. This means that the Yule`s coefficients indicate a strong relationship between the factors, and the Pearson's conjugacy coefficient indicates its absence (Pearson, 1904).

The sensitivity of Yule`s coefficients to the presence of small numbers in the conjugacy table has led to a theoretical preference for Pearson's conjugacy coefficient over Yule`s coefficients (Rauber, Nesbitt).

Another significant advantage of Pearson's conjugacy coefficient over Yule`s coefficients is that it can be used for conjugacy tables of dimension greater than 2x2 - with any number of columns l and rows m. None of Yule's coefficients is suitable for this.

As studies have shown, the x^2 changes in leaps and bounds in conjugacy tables as new data arrive, because it is based on calculating integers that change in leaps and bounds. And x^2 statistics were originally proposed for continuous distributions. In some cases, the analysis of conjugacy tables leads to misunderstandings in the interpretation of the values obtained. Therefore, in 1934, Yates's correction was proposed, which attenuates these jumps. To do this, when calculating x^2, the following correction is introduced into the numerator when calculating x^2.

$$\chi^2 = \sum_i \frac{(|y_i - y_i| - 0,5)^2}{y_i} \tag{18}$$

Different studies have also been carried out on Pearson's conjugacy coefficient itself. The famous Russian statistician Alexander Tschuprow proposed a correction to this coefficient for the case of a conjugacy table with more than four elements:

$$C = \sqrt{\frac{\chi^2}{N\sqrt{(l-1)(m-1)}}} \tag{19}$$

where l and m are the number of rows and columns in the conjugacy table.

Another popular coefficient that develops Pearson's idea and is based on x^2 is Cramer's V coefficient.

$$V = \sqrt{\frac{\chi^2}{N \min(l-1)(m-1)}} \tag{20}$$

There are several other auxiliary coefficients, but they have no independent meaning, so we will not consider them here.

Let us look at a number of other examples to see if the claims of practitioners about the accuracy of diagnosing the correlation between the nominal data of Yule`s and Pearson`s coefficients are justified. Using the rule in Table 5, we will generate a situation in where there is no correlation according to Yule (and Pearson), and where their coefficients are equal to zero. Let a = 10, c = 120, k = 2. Then:

Table 10. Lack of relationship according to Yule and Pearson

	x_1	x_2	Total
y_1	3	9	12
y_2	10	30	40
Total	13	39	

As expected, all three coefficients considered earlier are equal to zero for the data in Table 12. Now let us do this. Reduce the number d in the conjugacy table from 30 to 11.

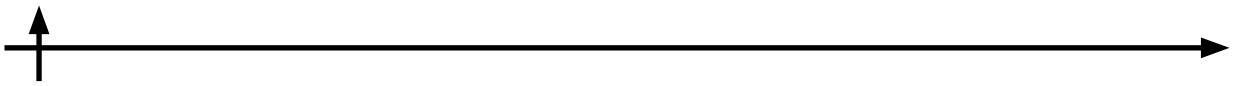


Table 11. Lack of relationship according to Yule and Pearson

	x_1	x_2	Total
y_1	3	9	12
y_2	10	11	21
Total	13	20	

As can be seen from this table, as the number of observations for one attribute increases, so does the number of observations for another attribute: in the totals column - from 12 to 21, in the totals row - from 13 to 20. That is, we see that there is a positive correlation - an increase in one corresponds to an increase in the other. And what do the coefficients discussed above diagnose us according to the data in this table?

Here are their values: Q = -0,463, ω = 0,246 and φ = -0,223. That is, they diagnose a negative relationship between the factors - the opposite of what actually exists.

And this is understandable because the coefficients are equal to zero under the conditions of Table 12, and reducing the value of d in this table by at least one (Table 14) causes all the coefficients to become negative (Q = -0,017, ω = -0,008 and φ = -0,006).

Table 12. Inverse relationship according to Yule and Pearson (negative coefficients)

	x_1	x_2	Total
y_1	3	9	12
y_2	10	29	39
Total	13	38	

And increasing the value of d by the same unit (Table 15) results in all three coefficients becoming positive (Q = 0,016, ω = 0,008 and φ = 0,006).

Table 13. Direct relationship according to Yule and Pearson (positive coefficients)

	x_1	x_2	Total
y_1	3	9	12
y_2	10	31	41
Total	13	40	

In both the first and the second cases, an increase in the values of the quantities of one attribute is accompanied by an increase in the quantities of the values of the other attribute. This means a positive relationship. Here is an example that confirms this paradox even more vividly.

Table 14. Conditional example

	x_1	x_2	Total
y_1	10	200	210
y_2	120	250	370
Total	30	50	580

An analysis of the numbers in Table 16 shows that when moving from x1 to x2, the numbers in each row increase: from 10 to 200, from 120 to 250. And as you move from y1 to y2, the numbers in the columns increase: from 10 to 120, from 200 to 250. If the growth in the

indicators of one attribute is accompanied by the growth in the indicators of another attribute, then we have an obvious positive relationship. What do the coefficients Q, ω and φ give us for this case?

And here's what we get: Q = - 0,811, ω = - 0,512, φ = -0,319.

That is, they unanimously signal the presence of a negative relationship between the factors. But the increase in the value of one indicator is accompanied by a similar increase in the value of another indicator, not by its decrease!

Results and Discussion

The assumption of Yule and Pearson that a zero correlation is diagnosed when the proportions between the values of the numbers in the conjugacy table b/a = d/c = k are kept, is not true. The coefficients they propose will therefore give a distorted picture of the situation. An alternative coefficient, based on different assumptions about the presence or absence of a relationship, is required (Pearson, 1900).

The lack of correlation between the attributes means that if one of the attributes changes, the characteristics of the other attribute will not react in any way. This gives reason to propose an appropriate correlation coefficient for its use according to the data of the conjugacy tables.

First of all, let us note that the conjugacy tables under consideration between two attributes can be represented graphically in a three-dimensional space. The axes of this space are the attributes x and y, and the number of observed occurrences of each of these attributes n (Fig. 1).

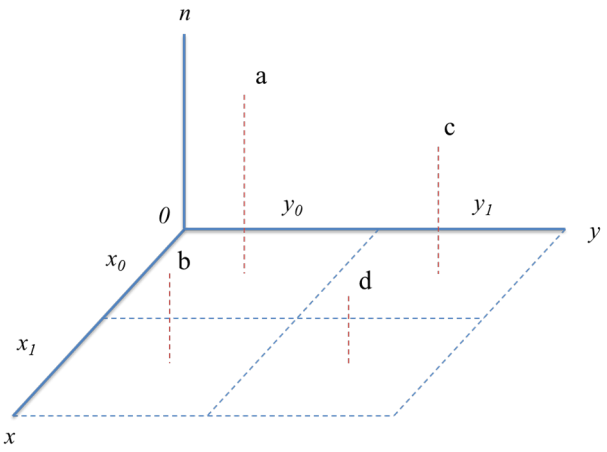


Fig. 1. A Table of feature conjugacy in graphical interpretation

Four points in three-dimensional space are projected onto each of the planes that make up the space. The n0y and n0x planes are of interest since the quantities n are not projected onto the x0y plane, and for any distribution of numbers of any nominal scale, the same points will be depicted on this plane.

The first of the considered planes n0y looks like this.

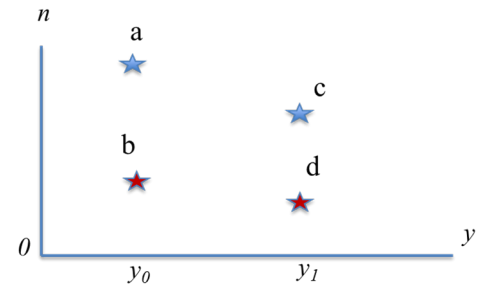


Fig. 2. Projections of the conjugacy table points on the plane n0y

It is well known that one and only one straight line can be drawn through two points. Let us do this by drawing a straight line on the plane in question through points a and c, since these two points reflect a change in one attribute, and through points b and d - another one, since they reflect a change in the number of observations of another attribute. We obtain:

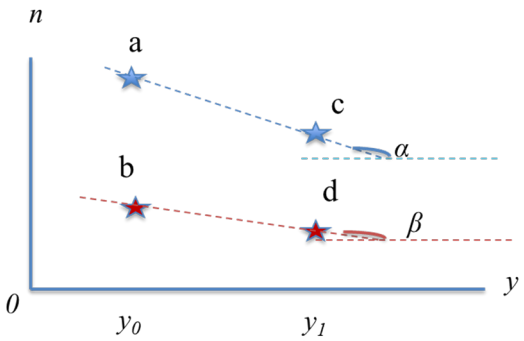


Fig. 3. Straight lines on the plane n0y

Therefore, in the calculated coefficient, we should calculate the arithmetic mean of the moduli of the tangents on each plane. It will characterize how far the points in the conjugacy table are from equal values. For the plane n0y we have:

$$\frac{|a-c|+|b-d|}{2} \tag{21}$$

And for another plane n0x:

$$\frac{|a-b|+|c-d|}{2} \tag{22}$$

If we multiply these two arithmetic means of the tangent moduli and extract the square root of their product, then the geometric mean of these means modulo tangents will be obtained:

$$\frac{1}{2}\sqrt{(|a-c|+|b-d|)(|a-b|+|c-d|)} \tag{23}$$

The geometric mean (23) can be as large as you like - this is the tangent. To bring the desired coefficient within the required limits of minus one to plus one, the value (23) must be scaled so that it does not exceed modulo one. How to do this?

To do this, the numbers in the conjugacy table a, b, c, d should be divided by the maximum value of these numbers. In this case, none of the tangents of the angles will exceed one, and

their geometric mean (23), scaled in this way will be between zero and one. Then we get:

$$\frac{1}{2}\frac{\sqrt{(|a-c|+|b-d|)(|a-b|+|c-d|)}}{\max(a,b,c,d)} \tag{24}$$

At the same time (24) will be always positive, regardless of whether there is a negative or positive correlation between the attributes. When calculating the required coefficient, it is necessary to ensure that it has a "+" sign for a positive connection and a sign "-" for a negative connection (Mills, 2017).

If one plane shows a generally upward tendency and the same tendency is generally reflected on the second plane, this indicates that the relationship has a positive sign. But if these two tendencies, both on the first and on the second plane, are of a downward nature and their tangents are negative, then this is also a direct dependence, since a decrease in the values of one attribute results in a decrease in the values of another attribute, i.e. they change in one direction (Janning).

But if multidirectional trends are observed on the planes, that is, an increase in one attribute is accompanied by a decrease in the other one, and then this means feedback between the attributes.

Consequently, the sign of the direction of the relationship between the nominal data can be determined by multiplying each other the sums of the tangents on each of the planes, i.e.

$$(ac+bd-(bc+ad))\cdot(ab+cd-(cb+ad)) \tag{25}$$

The sign of this product should be applied to the required coefficient.

MS Excel has such a built-in function that determines the sign of any mathematical operations. If the researcher is using another software product that does not have this function, the sign to be put before the calculated coefficient, can be found as follows:

$$\xi = \frac{(ac+bd-(bc+ad))\cdot(ab+cd-(cb+ad))}{|(ac+bd-(bc+ad))\cdot(ab+cd-(cb+ad))|} \tag{26}$$

The coefficient ξ, as one can see, takes only two values - plus one or minus one.

Taking into account this sign, which determines the direction of the correlation, the required coefficient will have the following form:

$$S = \xi\frac{1}{2}\frac{\sqrt{(|a-c|+|b-d|)(|a-b|+|c-d|)}}{\max(a,b,c,d)} \tag{27}$$

It will be equal to zero if the tangent of both lines is zero on at least one of the planes considered, and in all cases, it will be modulo greater than zero but less than one.

Now it is necessary to understand how to interpret the values of the coefficient (27), modulo in the range from zero to one.

For example, what does S=0.5 show? Is the correlation between the attributes strong or weak? The fact that it is straight is indicated by the "+" sign, and what is the strength of the relationship?

At first glance, it seems that a linear scale for interpreting correlation coefficients values could be used to diagnose the strength of the relationship, the one which the researchers usually use when calculating Yule`s and Pearson`s coefficients. Normally they are given this interpretation.

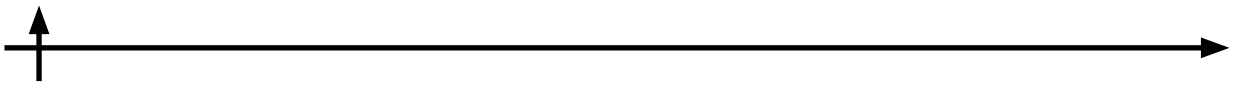


Table 15. Standard interpretation of correlation coefficients values

The value of the coefficient module, k (Q ∨ ω ∨ φ)	Strength of relationship
0 <k <0,1	Absence of relationship
0,1 ≤ k <0,2	Weak
0,2 ≤ k <0,4	Average
0,4 ≤ k <0,6	Relatively strong
0,6 ≤ k <0,8	Strong
k ≥ 0,8	Very strong

But in fact, this table cannot be applied directly to the interpretation of the values of the coefficient (27). The fact is that the coefficient S is a tangent, which has been formed in a complicated way by averaging and calculating the geometric mean. And the tangent is a non-linear function, and the linear interpretation presented in Table 17 cannot be applied to it.

For the calculated coefficient (27), the angles of inclination act as an argument, which, depending on the degree of connection of attributes, vary linearly from zero to the maximum angle whose tangent modulus is equal to one. This linear change is transformed by (27) into a non-linear change of the new correlation coefficient (Yule, 1912). Therefore, it is necessary to "tie" the change in the angle of inclination to the linear scale of Table 17, and then to find the correspondence of the coefficient (27) to one or another degree of the relationship between the nominal data. Let us do this.

Our starting point is that the maximum value of (27) is modulo one. The tangent is equal to one if the argument (the angle of inclination φ) is π/4. The minimum modulo value of the coefficient (27) is zero and it corresponds to the zero angle of inclination. Consequently, a change in the argument from 0 to π/4 corresponds to a change in the strength of the correlation tie from its absence (at zero angle of inclination) to the highest degree (when the angle of inclination is equal to π/4).

Then we can propose the following interpretation of the values of the coefficient modulus S (with rounding off to convenient numbers).

Table 16. Interpretation of coefficient values (27)

The value of the coefficient module, k (Q ∨ ω ∨ φ)	The value of the argument (angle φ) corresponding to the segment of the relationship degree scale	The modulus of the coefficient S as the tangent of the argument	Strength of relationship
0<k <0,1	0< φ < π/40	0 ≤ S_{gs} <0,08	Absence of relationship
0,1 ≤ k <0,2	π/40 ≤ φ < π/20	0,08 ≤ S_{gs} <0,16	Weak
0,2 ≤ k <0,4	π/20 ≤ φ < π/10	0,16 ≤ S_{gs} <0,33	Average
0,4 ≤ k <0,6	π/10 ≤ φ < 3π/20	0,33 ≤ S_{gs} <0,5	Relatively strong
0,6 ≤ k <0,8	3π/20 ≤ φ < π/5	0,5 ≤ S_{gs} <0,73	Strong
k ≥ 0,8	φ ≥ π/5	S_{gs} ≥ 0,73	Very strong

Now it is possible to get an answer to the question of what the value of the coefficient S=0.5 indicates. If it were Yule`s association coefficient, it would diagnose a relatively strong degree of association (Table 17). And such a value of the new coefficient diagnoses a strong relationship between the attributes (Leibniz, Clarke, 2000).

Let us test how the new coefficient works on those examples that questioned the acceptability of existing coefficients for diagnosing the degree of connection between the attributes. Thus, in

tables 14 and 15 the conditional values of the numbers were given at which these coefficients had negative (Table 14) and positive (Table 15) values, although the essence of the relationship did not change much - it was direct and positive. This is easily verified by both the numbers in these tables and their graphical representation, which is shown in Fig. 4.

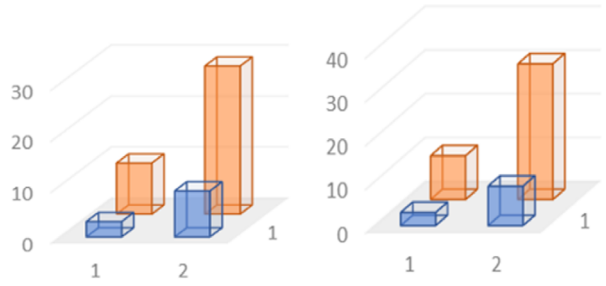


Fig. 4. Graphical representation of the data in Table 14 (left) and Table 15 (right)

It can be seen from the figure that in both one case and another, the growth of one attribute is accompanied by the growth of another attribute, i.e. the condition of a positive relationship between them is fulfilled. They differ from each other in that the last number in the tables, which corresponds to the upper right column in the figures, changes its values from d=29 (Table 14) up to 31 (Table 15), while the other numbers remain unchanged.

On the graph, this change by two units is not even noticeable. However, for the graph of the values presented on the left side of the figure, the following coefficients are calculated: Q = -0,017, ω = -0,008 and φ = - 0.006, and for the values presented on the right side of the figure, they become positive: Q = 0,016, ω = 0,008 and φ = 0,006.

The proposed coefficient (27) for the first case of Table 14 is equal to S = 0.448, and for the case of Table. 15, it is equal to S = 0.448. Both in one case and the other, a positive relationship between the attributes is diagnosed, which, in accordance with the recommendations of the Table can be interpreted as relatively strong.

Let us check how the proposed coefficient (27) works on other conditional examples that were given in the tables of the previous paragraphs, except for the examples just discussed in Tables 14 and 15, and compare it with what the coefficients Q, ω, φ show.

For the convenience of subsequent interpretation of the values of all four coefficients, let us summarize in a single table of the correspondence between the values of these correlation coefficients to the strength of the relationship from Tables 17 and 18.

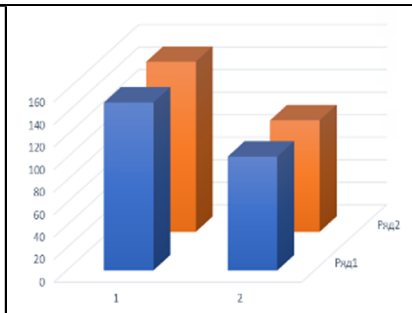
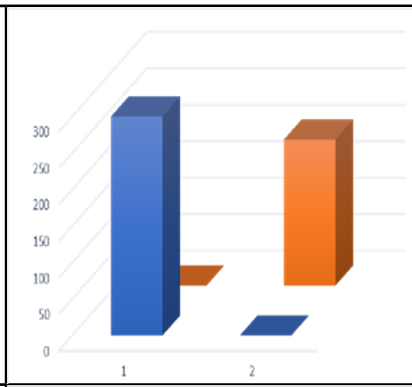
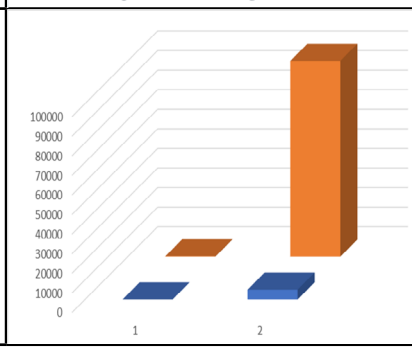
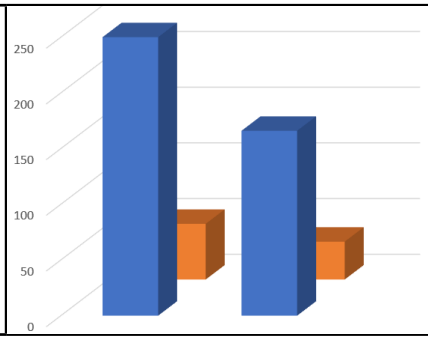
Table 17. Standard interpretation of correlation coefficients values

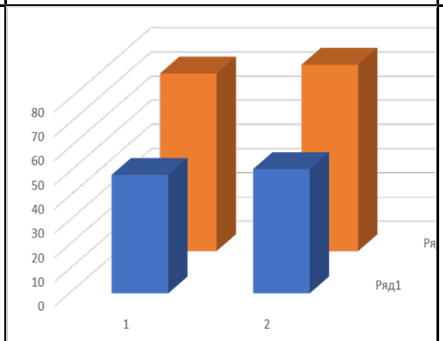
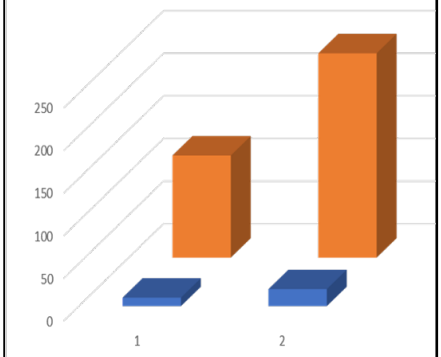
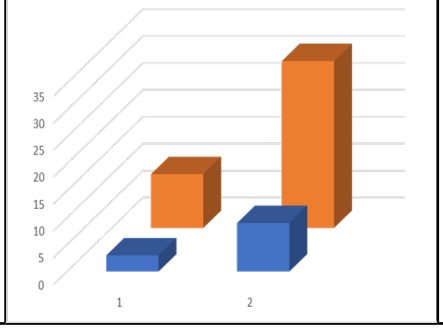
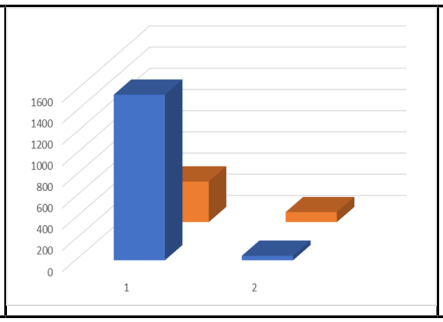
k (Q ∨ ω ∨ φ)	Strength of relationship	S
0<k <0,1	Absence of relationship	0 ≤ S <0,08
0,1 ≤ k <0,2	Weak	0,08 ≤ S <0,16
0,2 ≤ k <0,4	Average	0,16 ≤ S <0,33
0,4 ≤ k <0,6	Relatively strong	0,33 ≤ S <0,5
0,6 ≤ k <0,8	Strong	0,5 ≤ S <0,73
k ≥ 0,8	Very strong	S ≥ 0,73

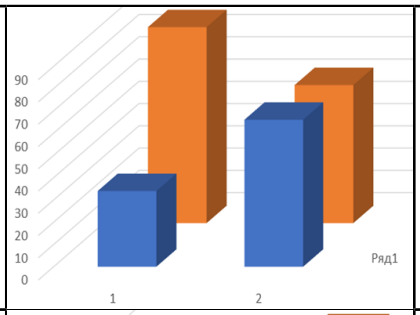
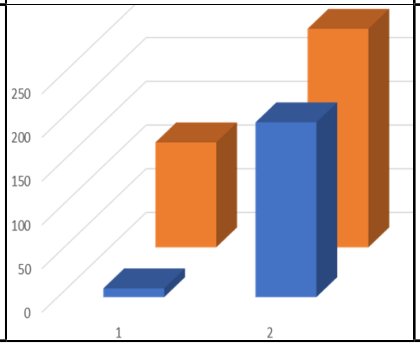
But before interpreting certain values of the calculated coefficients, it should be noted that in correlation practice it has long been believed that Yule`s coefficient slightly overstates its values with respect to the true value of the strength of the relationship, and that his colligation coefficient and Pearson's conjugacy coefficient slightly understate their values with respect to

the real degree of correlation. But since we are offering an alternative to these coefficients, we will not go into these details any further. Let us divide all these tables into groups with a typical situation and summarize the results of the calculation of all the coefficients.

Table 18. Comparative analysis of correlation coefficient calculations from previous tables

Number of the table	Representation of data in three-dimensional form	Q	ω	φ	S
Lack of correlation					
2		-0,017 No correlation	-0,008 No correlation	-0,008 No correlation	-0,066 No correlation
Strong correlation					
3		0,999 Very strong correlation	0,984 Very strong correlation	0,983 Functional correlation	0,826 Very strong correlation
11		-0,818 Very strong negative	-0,519 Relatively strong negative	-0,009 no correlation	0,499 strong
There is a correlation, but Yule`s and Pearson`s coefficients show that there is none					
4		0 Lack of correlation	0 Lack of correlation	0 Lack of correlation	0,364 Relatively strong

Number of the table	Representation of data in three-dimensional form	Q	ω	φ	S
9		0 Lack of correlation	0 Lack of correlation	0 Lack of correlation	0,112 Weak positive
12		0 Lack of correlation	0 Lack of correlation	0 Lack of correlation	0,449 Relatively strong
15		0,016 Lack of correlation	0,008 Lack of correlation	0,005 Lack of correlation	0,451 Relatively strong positive
An example of Yule and Pearson					
10		0,803 Very strong positive	0,503 Relatively strong positive	0,291 Average positive	-0,477 Relatively strong negative

Number of the table	Representation of data in three-dimensional form	Q	ω	φ	S
The correlationis positive, and the Yule`s and Pearson`s coefficients show a negative correlation					
8		-0,467 Relatively strong negative	-0,248 Average negative	-0,242 Average negative	0,329 Average positive
16		-0,811 Very strong negative	-0,512 Relatively strong negative	-0,319 Average negative	0,452 Relatively strong positive

These are very interesting results. It is an undeniable fact that as a person ages, he is more careful about his health, at least due to the fact that he gets sick more often and chronic diseases appear. From this unconditional fact, it is logical to conclude that, among the various aspects of increased attention to their health by the elderly, there should also be a growing interest in healthy lifestyles as one of the ways of independent health care for the residents of the region (Geddes, 2022).

However, the coefficients have shown a negative direction of the correlation. However, of the three classical coefficients, only one indicates a weak negative correlation - this is Yule`s association coefficient and the S coefficient shows the presence of an average negative relationship. The other two show that, in fact, there is essentially no correlation here.

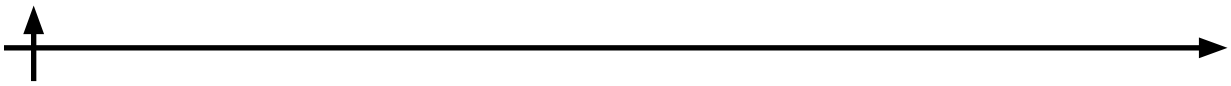
2. Young people from another region of Russia were asked, among other things, on the problem whether they trusted political parties or youth associations more?

The results of this survey are presented below.

Table 19. Conjugacy Table of Healthy Lifestyle

The meaning of the signs	x_1 - from 18 to 45 years	x_2 - from 45 years and older	Total
y_1 - interested	215	234	449
y_2 - not interested	175	136	311
Total	390	370	760

Q	ω	φ	S
-0,167 (Weak)	-0,084 (Lack of correlation)	-0,082 (Lack of correlation)	-0,191 (Average)



These are very interesting results. It is an undeniable fact that as a person ages, he is more careful about his health, at least due to the fact that he gets sick more often and chronic diseases appear. From this unconditional fact, it is logical to conclude that, among the various aspects of increased attention to their health by the elderly, there should also be a growing interest in healthy lifestyles as one of the ways of independent health care for the residents of the region (Geddes, 2022).

However, the coefficients have shown a negative direction of the correlation. However, of the three classical coefficients, only one indicates a weak negative correlation - this is Yule`s association coefficient and the S coefficient shows the presence of an average negative relationship. The other two show that, in fact, there is essentially no correlation here.

2. Young people from another region of Russia were asked, among other things, on the problem whether they trusted political parties or youth associations more?

The results of this survey are presented below.

Table 19. Conjugacy Table of Healthy Lifestyle

The meaning of the signs	x_1 - political parties	x_2 - youth associations	Total
y_1 - trust	14	35	49
y_2 - do not trust	28	6	34
Total	42	41	83

Q	ω	φ	S
-0,842 very strong inverse	-0,547 relatively strong inverse	-0,529 relatively strong inverse	0,614 strong direct

Yule`s association coefficient shows that young people trust youth associations more than political parties. Two other classical coefficients also confirm the direction of this dependence but note that this dependence is relatively strong. But the new coefficient S diagnoses the opposite situation. It points out that young people show a direct dependence between the attributes, i.e. they trust parties but not youth organizations (Tschuprow). Since all youth organizations registered in Russia are pro-government and work in line with the policies of the main pro-government parties, and since there is at least some disagreement with the authorities and pro-government parties among the registered parties, the conclusion given by the coefficient S should be preferred - among the youth there is a large percentage of nihilists or rebels who do not agree with any government (represented by adults), so they will trust heterogeneous parties more than homogeneous youth organizations.

Conclusions

Yule`s and Pearson`s classical coefficients are based on such assumption about the situation of the absence of correlation between data, which introduces inaccuracy into the procedure for estimating the degree and direction of correlation between the data. The new coefficient was proved based on other prerequisites and the situation of lack of relationship between attributes, where a change in one attribute does not affect in any way a change in another attribute.

A comparative analysis of all coefficients - both old and new – has shown that the coefficient S successfully copes with its task assigned to it. It both assesses the degree of correlation and identifies its direction. It both assesses the extent of the relationship and identifies its direction. This coefficient is slightly more difficult to calculate than Yule`s and Pearson`s coefficients. But who calculates such coefficients by hand nowadays? And for computer calculations, the new coefficient presents no difficulties. This new coefficient has one significant drawback

compared to Pearson's conjugacy coefficient - it can only work with two-dimensional conjugacy tables. And Pearson's conjugacy coefficient, which is calculated using the x^2 distribution, can be applied to conjugacy tables of any dimension.

Therefore, the coefficient S for multidimensional cases and conjugacy tables can only be used by reducing them to a two-dimensional case by a method well known to practitioners - property A, and all other properties are not A. The solution to the problem in this way becomes quite labor-intensive, but today, with the digitalization of the scientific process, it should not embarrass anyone.

REFERENCES

Agarwal B.L. 2006. Basic Statistics. New Age International, 788.

Boon J. 2020. Relationships and the Course of Social Events During Mineral Exploration: An Applied Sociology Approach. Switzerland: Springer, 125.

Cramer H. 1946. Mathematical Methods of Statistics. Princeton University Press, 280-282.

Deisenroth M.P., Ong C.S., Faisal A.A. Mathematics for Machine Learning. United Kingdom: Cambridge University Press, 398.

Edwards A., David H. Annotated Readings in the History of Statistics. United States: Springer New York, 252.

Geddes P. 2022. Civics: as Applied Sociology. DigiCat, 109.

Janning M. A Guide to Socially-Informed Research for Architects and Designers. United Kingdom: Taylor & Francis, 178.

Leibniz G.W., Clarke S. 2000. Correspondence. Roger Ariew Edition: Hackett Publishing Company, 110.

Mills T.C. 2017. A Statistical Biography of George Udny Yule. A Loafer of the World. Cambridge Scholars Publishing, 530. doi: 10.4135/9781526421036935452

Pearson K. 1904. On the Theory of Contingency and its Relation to Association and Normal Correlation. Dulau & Co., 35.

Pearson K.X. 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 5 (50), 157-175. doi: 10.1080/14786440009463897

Rauber R.M., Nesbitt S.W. Radar Meteorology: A First Course. Germany: Wiley, 496.

Reagle D., Vinod H.D. Preparing for the Worst: Incorporating Downside Risk in Stock Market Investments. Germany: Wiley, 320.

Thompson N. 2019. Applied Sociology. United Kingdom: Routledge, 220 p.

Tschuprow A.A. Principles of the mathematical theory of correlation. Wm. Hodge and Co., 194. doi:10.2307/3606786

Wu J., Gan G., Ma C. 2007. Data Clustering: Theory, Algorithms, and Applications. Society for Industrial and Applied Mathematics, 466. doi: 10.1137/1.9780898718348

Yates F. 1934. Contingency tables involving small numbers and the chi-square test. Supplement to the Journal of the Royal Statistical Society 1 (2), 217-235. doi: 10.2307/2983604

Yule G.U. 1912. On the Methods of Measuring Association Between Two Attributes. Journal of the Royal Statistical Society 75 (6), 579–652. doi:10.2307/2340126

Zudin Y.B. 2023. Theory of Periodic Conjugate Heat Transfer Mathematical Engineering. Springer Nature, 440.

Handbook of Meta-analysis in Ecology and Evolution. United States: Princeton University Press, 520.

СПИСОК ИСТОЧНИКОВ

Agarwal B.L. 2006. Basic Statistics. New Age International, 788.

Boon J. 2020. Relationships and the Course of Social Events During Mineral Exploration: An Applied Sociology Approach. Switzerland: Springer, 125.

Cramer H. 1946. Mathematical Methods of Statistics. Princeton University Press, 280-282.
Deisenroth M.P., Ong C.S., Faisal A.A. Mathematics for Machine Learning. United Kingdom: Cambridge University Press, 398.
Edwards A., David H. Annotated Readings in the History of Statistics. United States: Springer New York, 252.
Geddes P. 2022. Civics: as Applied Sociology. DigiCat, 109.
Janning M. A Guide to Socially-Informed Research for Architects and Designers. United Kingdom: Taylor & Francis, 178.
Leibniz G.W., Clarke S. 2000. Correspondence. Roger Ariew Edition: Hackett Publishing Company, 110.
Mills T.C. 2017. A Statistical Biography of George Udny Yule. A Loafer of the World. Cambridge Scholars Publishing, 530. doi: 10.4135/9781526421036935452
Pearson K. 1904. On the Theory of Contingency and its Relation to Association and Normal Correlation. Dulau & Co., 35.
Pearson K.X. 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 5 (50), 157-175. doi: 10.1080/14786440009463897
Rauber R.M., Nesbitt S.W. Radar Meteorology: A First Course. Germany: Wiley, 496.
Reagle D., Vinod H.D. Preparing for the Worst: Incorporating Downside Risk in Stock Market Investments. Germany: Wiley, 320.
Thompson N. 2019. Applied Sociology. United Kingdom: Routledge, 220 p.
Tschuprow A.A. Principles of the mathematical theory of correlation. Wm. Hodge and Co., 194. doi:10.2307/3606786
Wu J., Gan G., Ma C. 2007. Data Clustering: Theory, Algorithms, and Applications. Society for Industrial and Applied Mathematics, 466. doi: 10.1137/1.9780898718348
Yates F. 1934. Contingency tables involving small numbers and the chi-square test. Supplement to the Journal of the Royal Statistical Society 1 (2), 217-235. doi: 10.2307/2983604
Yule G.U. 1912. On the Methods of Measuring Association Between Two Attributes. Journal of the Royal Statistical Society 75 (6), 579–652. doi:10.2307/2340126
Zudin Y.B. 2023. Theory of Periodic Conjugate Heat Transfer Mathematical Engineering. Springer Nature, 440.
Handbook of Meta-analysis in Ecology and Evolution. United States: Princeton University Press, 520.

INFORMATION ABOUT AUTHOR / ИНФОРМАЦИЯ ОБ АВТОРЕ

SVETUNKOV Sergey G. – Professor, Doctor of Economic Sciences
E-mail: sergey@svetunkov.ru
СВЕТУНЬКОВ Сергей Геннадьевич – профессор, д.э.н.
E-mail: sergey@svetunkov.ru

Статья поступила в редакцию 13.06.2023; одобрена после рецензирования 20.06.2023; принята к публикации 21.06.2023.
The article was submitted 13.06.2023; approved after reviewing 20.06.2023; accepted for publication 21.06.2023.