УДК 330.43 : 519.2

5.2.2. Математические, статистические и инструментальные методы в экономике (физико-математические науки, экономические науки)

О КОРРЕЛЯЦИИ

Орлов Александр Иванович
д.э.н., д.т.н., к.ф.-м.н.
профессор
РИНЦ SPIN-код: 4342-4994
prof-orlov@mail.ru

Орлов Антон Александрович
ассистент
РИНЦ SPIN-код: 6151-2018
antorlov@mail.ru
Московский государственный технический университет им. Н.Э. Баумана, Россия, 105005, Москва, 2-я Бауманская ул., 5

Термин «корреляция» означает «связь между переменными». Применительно к анализу данных этот термин обычно используется в сочетании «коэффициент корреляции». Методы изучения корреляции широко применяются при анализе данных в различных областях знаний (десятки тысяч публикаций в РИНЦ). Однако многие вопросы требуют тщательного рассмотрения. Им и посвящена настоящая статья. Согласно вероятностно-статистической модели корреляционного анализа исходные данные представляют собой выборку из двумерного распределения (как правило, отличного от нормального). Введены коэффициенты корреляции Пирсона, Спирмена и Кендалла. Показана некорректность термина "корреляционно-регрессионный анализ", широко используемого в публикациях, рассмотренных в статье. Дело в том, что модель корреляционного анализа – лишь одна из моделей регрессионного анализа. Корреляционный анализ позволяет оценивать степень связи, прогнозировать значение одной переменной по значению другой, но не позволяет управлять, а именно, изменяя значение одной переменной, целенаправленно менять значение другой. Так, изменение (например, уменьшение) веса взрослого человека не приводит к изменению его роста, хотя коэффициент корреляции между этими переменными значителен. Для того, чтобы с помощью регрессионной зависимости разрабатывать управленческие решения, необходима серия предварительных экспериментов. В ней исследователь задает (по определенным правилам) значения независимой переменной и измеряет соответствующие значения зависимой. Эта область прикладной статистики называется «планирование эксперимента». Рассмотрен ряд вариантов шкалы Чеддока,

UDC 330.43 : 519.2

5.2.2. Mathematical, statistical and instrumental methods of economics (physical and mathematical sciences, economic sciences)

ABOUT CORRELATION

Orlov Alexander Ivanovich
Dr.Sci.Econ., Dr.Sci.Tech., Cand.Phys-Math.Sci.,
professor
RSCI SPIN-code: 4342-4994
prof-orlov@mail.ru

Orlov Anton Alexandrovich
assistant
RSCI SPIN-code: 6151-2018
antorlov@mail.ru
Bauman Moscow State Technical University, Moscow, Russia

The term "correlation" means "relationship between variables". In relation to data analysis, this term is usually used in combination with "correlation coefficient". Methods of studying correlation are widely used in data analysis in various fields of knowledge (tens of thousands of publications in the Russian Science Citation Index). However, many issues require careful consideration. This article is devoted to them. According to the probabilistic-statistical model of correlation analysis, the initial data are a sample from a two-dimensional distribution (usually different from normal). Pearson, Spearman and Kendall correlation coefficients are introduced. The incorrectness of the term "correlation-regression analysis", widely used in the publications considered in the article, is shown. The fact is that the correlation analysis model is only one of the regression analysis models. Correlation analysis allows you to assess the degree of connection, predict the value of one variable based on the value of another, but does not allow you to control, namely, by changing the value of one variable, purposefully change the value of another. Thus, a change (for example, a decrease) in the weight of an adult does not lead to a change in his height, although the correlation coefficient between these variables is significant. In order to develop management decisions using regression dependence, a series of preliminary experiments is necessary. In it, the researcher sets (according to certain rules) the values of the independent variable and measures the corresponding values of the dependent variable. This area of applied statistics is called "experimental planning". A number of variants of the Chaddock scale, designed to interpret the values of the correlation coefficients in verbal form, are considered. In a non-parametric formulation, the problem of assessing the significance of correlation coefficients by testing the statistical hypothesis that the theoretical correlation

предназначенной для интерпретации значений коэффициентов корреляции в словесной форме. В непараметрической постановке рассмотрена проблема оценки значимости коэффициентов корреляции путем проверки статистической гипотезы о том, что теоретический коэффициент корреляции равен 0. Получена теорема об асимптотической нормальности коэффициентов корреляции Пирсона, Спирмена и Кендалла. Результаты проверки значимости сопоставлены с оценками по шкале Чеддока. Рассмотрены некоторые направления дальнейших исследований

Ключевые слова: СТАТИСТИЧЕСКИЕ МЕТОДЫ, ПРИКЛАДНАЯ МАТЕМАТИЧЕСКАЯ СТАТИСТИКА, КОРРЕЛЯЦИЯ, ВЕРОЯТНОСТНО-СТАТИСТИЧЕСКАЯ МОДЕЛЬ, КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫЙ АНАЛИЗ, ПЛАНИРОВАНИЕ ЭКСПЕРИМЕНТА, КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ ПИРСОНА, КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ СПИРМЕНА, КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ КЕНДАЛЛА, ШКАЛА ЧЕДДОКА, ПРОВЕРКА ГИПОТЕЗ

coefficient is 0 is considered. A theorem on the asymptotic normality of the Pearson, Spearman and Kendall correlation coefficients is obtained. The results of the significance test are compared with the estimates on the Chaddock scale. Some directions for further research are considered

Keywords: STATISTICAL METHODS, APPLIED MATHEMATICAL STATISTICS, CORRELATION, PROBABILISTIC-STATISTICAL MODEL, CORRELATION- REGRESSION ANALYSIS, DESIGN OF EXPERIMENTS, PEARSON CORRELATION COEFFICIENT, SPEARMAN CORRELATION COEFFICIENT, KENDALL CORRELATION COEFFICIENT, CHADDOCK SCALE, HYPOTHESIS TESTING

http://dx.doi.org/10.21515/1990-4665-202-020

1. Introduction

The term "correlation" means "relationship between variables." When applied to data analysis, the term is usually used in combination with "correlation coefficient". Such coefficients are used to measure the magnitude and direction of the relationship between random variables.

In [1] the results of the search for publications are presented in the scientific electronic library eLIBRARY.RU using the keywords: “Correlation”, “Spearman’s Correlation”, “Kendall’s Correlation”, “Pearson’s Correlation”, “Concordance”. In table 1 provides a short excerpt.

Table 1

Number of search results

Search subject	Total	Economy	Mathematics
Correlation	38614	11266	5505
Pearson correlation coefficient	11922	974	368
Spearman correlation coefficient	12545	1221	333
Kendall correlation coefficient	4301	135	87
Concordance coefficient	846	699	322

Table data 1 show that methods for studying correlation are widely used in data analysis in various fields of knowledge. However, as shown below, many issues require careful consideration. This article is dedicated to them.

It is important to note that a large number of authors do not report which correlation coefficient they use. In such cases, we are most often talking about the Pearson correlation coefficient.

2. Correlation coefficients

As shown in [2 - 4], the description of data analysis methods should begin with the formulation of the corresponding probabilistic-statistical model.

Let be a two-dimensional random vector, i.e. a function defined on the space of elementary events Ω = {ω} with values in R2. In an outdated paradigm of mathematical methods, studies often assume that the distribution is bivariate normal. However, it is well known that the distributions of real data, as a rule, are not normal (Gaussian) [5]. Therefore, we will assume that the distribution of the random vector is arbitrary, i.e. We will consider a nonparametric model. In this case, we assume that the usual assumptions [5] of the Central Limit Theorem of probability theory are satisfied, allowing us to conclude that the asymptotic statements given below are valid.$(X,Y) = (X(\omega),Y(\omega))(X,Y)$ (X,Y)

To measure the relationship between the coordinates of a random vector, one or another correlation coefficient is used. Among them, the best known is the linear paired Pearson correlation coefficient:(X,Y)

$$r = \frac{M\{(X-M(X))(Y-M(Y))\}}{\sigma(X)\sigma(Y)},$$

where M(X) is the mathematical expectation of the random variable X, and is its standard deviation (i.e. the square root of the variance), etc.$\sigma(X)$

In applied statistics, the initial data is a sample, i.e. n pairs of numbers (i.e. n two-dimensional vectors)(x_i, y_i), $i = 1,2,...,n$, where n is the sample size. In the probabilistic-statistical model under consideration, the sample elements are

independent two-dimensional random vectors, identically distributed with the Sample linear paired Pearson correlation coefficient rn, as is known, the number$(X,Y) = \big(X(\omega), Y(\omega)\big)$.

$$r_n = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Here $\bar{x}$ - arithmetic mean of numbers x_i, $i = 1,2,...,n$.

If rn = 1, then $y_i = ax_i + b$, $i = 1,2,...,n$, for some numbers a and b, and a > 0. If rn = –1, then there is also a linear relationship between the variables xi and yi, i.e. $y_i = ax_i + b$, $i = 1,2,...,n$, but here a < 0. From these statements we can conclude that the closeness of the correlation coefficient to 1 (in absolute value) indicates a fairly close linear relationship between the variables under consideration.

In the probabilistic-statistical model under consideration, the sample correlation coefficient is a consistent estimate of the theoretical, i.e. converges (in probability) to the theoretical coefficient with an unlimited increase in the sample size:

$$\lim_{n\to+\infty} r_n = r.$$

In theoretical considerations it is often assumed that random vectors $(x_i, y_i) = (x_i(\omega), y_i(\omega))$, $i = 1,2,...,n$,have a bivariate normal distribution. As already mentioned, the distributions of real data, as a rule, differ from normal ones [5]. Why is the idea of a bivariate normal distribution widespread? The fact is that the theory in this case is simpler. In particular, the equality of 0 of the theoretical correlation coefficient is equivalent to the independence of random variables. If the assumption of bivariate normality is not met, then the equality of the theoretical correlation coefficient to 0 does not imply the independence of the random variables. It is not difficult to construct an example of a random vector for which the correlation coefficient is 0, but the coordinates are dependent. But

there is another way - to move to nonparametric correlation coefficients, which are equally suitable for any continuous distribution of a random vector.

To analyze data measured in interval and ratio scales, Pearson's linear pairwise correlation coefficient can be used, since its value does not change with acceptable transformations in these scales. But for data measured on an ordinal scale, it cannot be used, since its value, as a rule, changes with permissible transformations in the ordinal scale. In such cases, it is necessary to use the nonparametric rank correlation coefficients of Spearman and Kendall, as well as other coefficients developed in the theory of rank correlations [6]. We can say that the algorithms for calculating the Spearman coefficient and the Kendall coefficient convert any input data into ordinal data (ranks, i.e. places in an ordered series), and then make it possible to study them.

In accordance with the theory of stability of economic-mathematical methods and models [7, 8], it is advisable to carry out calculations for all three correlation coefficients (Pearson, Spearman and Kendall) and compare the results of the final calculations. If they are close, then there is no need to choose one or the other of the correlation coefficients. If they are different, then you should rely on the Spearman and Kendall rank correlation coefficients, and if you need to choose between them, it is advisable to choose the Kendall correlation coefficient, since it is linearly related to the Kemeny distance [9].

To calculate the Spearman and Kendall rank correlation coefficients, it is necessary to first rank the coordinate values of the vectors, i.e. construct variation series for populations $\{x_1, x_2, ..., x_n\}$ And $\{y_1, y_2, ..., y_n\}$ (separately). Let pj be the rank of xj in the aggregate $\{x_1, x_2, ..., x_n\}$ and qj is the rank of yj in the aggregate $\{y_1, y_2, ..., y_n\}$, where j = 1,2, …, n.

Sample Spearman correlation coefficient ρ_n for sampling $(x_i, y_i) = (x_i(\omega), y_i(\omega)),\ i = 1, 2, ..., n,$ calculated by the formula

$$\rho_n = 1 - \frac{6\sum_{j=1}^{n}(r_j - q_j)^2}{n^3 - n}. \quad (1)$$

Note that ρ_n is a sample linear paired Pearson coefficient constructed from a sample of ranks $(p_i, q_i),\ i = 1, 2, ..., n$.

Example 1. Let's calculate the sample Spearman correlation coefficient. The initial data is a sample of size n = 10, each element of which is a two-dimensional vector:

(2.35; 1), (4; 1.5), (3; 1.2), (1; 0), (2.25; 2),

(11; 5), (15; 8), (17; 9), (19; 10), (22, 10).

To calculate the sample Spearman correlation coefficient:

1. Let us write down the values of the coordinates of the sample elements from top to bottom in Table 2 (in this table i is the number of the sample element, i = 1, 2, ..., 10; xi are the values of its first coordinate, yi are the values of its second coordinate).

2. Let's rank the coordinate values. Let pi be the rank of xi among all elements of the column of values of the first coordinate (i.e., among all x1, x2, ..., x1n), and qi be the rank of yi among all elements of the column of values of the first coordinate (i.e., among all y1, y2, ... , y1n), i = 1, 2, …, 10.

3. For identical elements, as their final ranks (“linked ranks”) we will indicate the arithmetic mean of the ranks assigned to them during numbering - for example, if their numbers are 7 and 8, then we will assign each a “linked rank” of 7.5.

4. Note that the sum of the ranks for each data set of n elements must be equal to the sum of all natural numbers from 1 to n, i.e. $\frac{n(n+1)}{2}$, in this case - 55.

5. Further calculations will be carried out only on the values of ranks; the values of the source data will no longer be used.

6. For each pair of ranks (p_i, q_i), $i = 1, 2, ..., n$, Let's find their difference pi - qi. (The sum of all such differences in the ranks of elements with the same numbers will always be equal to 0.)

Table 2.

Initial data and coordinate ranks of sample elements

Sample element number i	First coordinate*xi*	Rank of the first coordinate	Second coordinate*yi*	Rank of the second coordinate
1	2.35	3	1	2
2	4	5	1.5	4
3	3	4	1.2	3
4	1	1	0	1
5	2.25	2	2	5
6	11	7	5	6
7	15	8	8	7
8	10	6	9	8
9	19	10	10	9.5
10	18	9	10	9.5

7. Let us square each difference in the ranks of elements with the same number.

8. Let’s find the sum of squares of the rank differences (pi – qi)2. For the data under consideration, this amount is 18.5 (Table 3).

Table 3.

Analysis of coordinate ranks of sample elements

Sample element number i	*pi*– rank xi among all x1, x2, …, x1n	*qi*– rank of yi among all y1, y2, …, y1n	Rank difference pi – qi	Squared difference of ranks (*pi*– qi)2
1	3	2	1	1
2	5	4	1	1
3	4	3	1	1
4	1	1	0	0
5	2	5	-3	9
6	7	6	1	1
7	8	7	1	1
8	6	8	-2	4
9	10	9.5	0.5	0.25
10	9	9.5	-0.5	0.25

9. We will find the Spearman rank correlation coefficient using formula (1). For the sample under consideration

$$\rho_n = 1 - \frac{6 \times 18.5}{10^3 - 10} = 0,88.$$

In other words, when calculating the Spearman rank correlation coefficient, we take the original data, for each coordinate of the sample elements we rank them from smaller to larger values, thereby assigning ranks to these values from 1 to the sample size (in the case of identical values, we use related ranks, i.e. . We take the arithmetic mean for the ranks assigned to these values), replace all values with their ranks, and then calculate the sample Pearson correlation coefficient for the resulting data.

When calculating the Kendall rank correlation coefficient, it is necessary to perform pairwise comparisons - to consider the change in the ranks of indicators (synonyms - parameters, coordinate values) when moving from one sample element to another. If there is a simultaneous increase or decrease in ranks for both comparison parameters, then this situation is called a coincidence (changes in ranks for two indicators are in the same direction). If there is an increase in one indicator and a decrease in another indicator, then this situation is called inversion (changes in ranks in two indicators in different directions). Kendall's rank correlation coefficient τ_n for sampling $(x_i, y_i) = (x_i(\omega), y_i(\omega)),\ i = 1, 2, ..., n,$ calculated by the formula

$$\tau_n = \frac{2(P - Q)}{n(n-1)}, \qquad (2)$$

where P is the number of matches; Q is the number of inversions [12, 13].

Like the Spearman coefficient, the Kendall coefficient is based on assigning ascending or descending serial numbers (ranks) to the coordinates of sample elements and further calculations based on them.

The key element of calculating the Kendall coefficient is to look for pairs of ranks that have a different order in each of the population data sets $\{x_1, x_2, ..., x_n\}$

And$\{y_1, y_2, ..., y_n\}$. For a correlation close to 1 according to the Kendall coefficient, large and small values in one set of data should be in the same order as in the other (or in the exact opposite - for reverse correlation).

Example 2. Let's consider the initial data of example 1. The algorithm for calculating the Kendall coefficient can be as follows:

1. Let's take numerical data sets of the same size, interconnected - in which each element of one set corresponds to one and only one element of the second set (that is, essentially representing one sample from pairs of values). As for calculating the Spearman rank correlation coefficient, we number in ascending order all the elements of both data sets from 1 to the sample size - “we will rank the elements.” For identical elements, as their final ranks, we will indicate the arithmetic mean of the ranks assigned to them during numbering (for example, if their numbers are 7 and 8, then we will assign each element a rank of 7.5), i.e., we will move on to related ranks. The sum of the ranks for each data set must be equal to the sum of the natural numbers from 1 to the sample size.

2. Further calculations will be carried out only on the values of ranks; the values of the source data will no longer be used.

3. Let's order the ranks of the elements of the first set (i.e., the ranks of the first coordinate of the sample elements) from smallest to largest and rearrange the series of ranks of the set of values of the second coordinate in accordance with this ordering (see Table 4). The original order of ranks will completely disappear (cf. Table 2). We can say that the rows of the table. 2 are rearranged in accordance with the increasing ranks of the elements of the first coordinate.

4. Consider the resulting sequence of ranks of the second coordinate (Table 4). For each rank in this sequence, we will find how many values greater than this rank occur in the sequence of ranks below in the same column, and indicate this value in Table 4 as the “number of matches.” We will also find out how many values smaller than this rank occur in the sequence of ranks below in the same column, and indicate this value in Table 3 as the “number of inversions”.

5. We add together the calculated numbers of coincidences, as well as the numbers of inversions. For the data under consideration, the total number of matches is P = 40, and the total number of inversions is Q = 5.

Table 4.

Calculation of matches and inversions

Rank of the first coordinate	Rank of the second coordinate	Coincidences	Inversions
1	1	9	0
2	5	5	3
3	2	7	0
4	3	6	0
5	4	5	0
6	8	2	2
7	6	3	0
8	7	2	0
9	9.5	1	0
10	9.5	0	0

6. Calculate the Kendall rank correlation coefficient using formula (2):

$$\tau_n = \frac{2(P-Q)}{n(n-1)} = \frac{2(40-5)}{10\times 9} = \frac{70}{90} = 0,78.$$

Other rank correlation coefficients have also been developed [6].

Since the values of the Spearman and Kendall rank correlation coefficients do not change with any strictly increasing transformations of measurement scales of the original sample data $(x_i, y_i) = (x_i(\omega), y_i(\omega)),\ i = 1, 2, ..., n,$ those. for any permissible transformations in ordinal scales, then these coefficients are used when analyzing data measured in ordinal scales [5].

The Pearson correlation coefficient evaluates the deviation from linearity. Its modulus reaches a maximum (equal to 1) if and only if the variables are related by a linear dependence. At the same time, the Spearman and Kendall rank correlation coefficients assess the deviation from monotonicity. They reach a maximum if and only if the values of the variables (synonyms - parameters, coordinate values, indicators) are equally ordered. And the minimum is when

the ordering is opposite, i.e. When the sign of one of the variables changes, the same ordering is observed.

To analyze data whose distributions do not obey the normal distribution (i.e. for almost all types of real data [10]), we recommend using Spearman and Kendall rank correlation coefficients.

In the basic probabilistic-statistical model, the joint distribution function of a two-dimensional random vector is assumed to be continuous, and therefore the probability of coincidence of the observed observation results is equal to 0. In practice, there are cases when some measurement results coincide, i.e. there are associated ranks. A model for analyzing coincidences when calculating nonparametric rank statistics was developed in article [11]. In relation to the Spearman and Kendall coefficients, the problem of taking into account related ranks is considered in [12, section. 6.10, pp. 152 - 155], [13, p. 207 - 208] and other works.$(X, Y) = (X(\omega), Y(\omega))$

3. About the term "correlation-regression analysis"

This term is widely used in publications. Sections in textbooks [14, 15], teaching aids [16, 17] are devoted to the topics behind it. methodological instructions for practical exercises [18] and for performing calculation tasks [19], scientific publications [20], reference materials [21]. Search in the Russian Science Citation Index (RSCI) for the query "Correlation and regression analysis" in the titles of articles and books for 2019 – 2024. gave 82 titles. Let us dwell on the topics of economics and management. This includes, for example, work on the use of a statistical analysis package to conduct correlation and regression analysis in the course of economic research [22] and on studying the relationship between socio-economic indicators of an organization's activities [23]. Correlation and regression analysis when assessing the investment attractiveness of the oil refining industry [24] and the efficiency of using investment resources of an oil company [25]. Correlation and regression analysis is used as a tool for assessing innovation activity in the regions [26]. It

is used to analyze the dependence of an enterprise's revenue on factors of foreign economic activity [27]. It is argued that with its help it is possible to study the influence of indicators on the revenue of an enterprise [28], the development of small businesses on the standard of living of the population [29], and inflation on the level of wages in the Russian Federation [30]. Correlation and regression analysis is used as a tool for predicting the impact of the functioning of a socio-economic cluster in the housing and communal services sector on the regional economy [31]. It turned out to be useful for analyzing the dependence of the revenue of the largest TNCs of the BRICS countries in the oil and gas industry on factors of economic activity [32]. Based on RSCI data, it can be established that correlation and regression analysis is used in many other fields of science and sectors of the national economy. Let us point out work at the intersection of economics and medicine [33], in which it was used to study the degree of influence of various socio-economic factors on the dynamics of the incidence of opisthorchiasis.

However, it must be stated that the term "correlation-regression analysis" from the point of view of modern applied statistics [5] is incorrect. It unreasonably mechanically combines completely different sections of applied statistics - correlation analysis and regression analysis. Let's discuss the relationship between these sections.

4. Correlation and regression: forecast and control

The correlation coefficients discussed above are intended to quantify the strength of the relationship between two random variables, i.e. between the coordinates of a two-dimensional random vector. The purpose of regression analysis is to restore the relationship between variables, at least one of which is random. The second one can be deterministic. For example, when studying the dynamics of indicators of the financial and economic activities of an enterprise. The variety of regression analysis models is discussed in article [34]. One of these models is designed to analyze a sample from the distribution of a two-

dimensional random vector. This statistical data generation model is also the original model for correlation analysis. It is discussed at the beginning of this article. Other regression analysis models are fundamentally different (see [34]).

The point is that correlation analysis allows you to assess the degree of connection, predict the value of one variable based on the value of another, but does not allow you to control - by changing the value of one variable, purposefully change the value of another.

This fact has long been known. It is called the correlation paradox [35]. As an example, let's discuss the analysis of data on the height and weight of a certain number of people. One can adopt a probabilistic-statistical model of correlation analysis, according to which the specified data are considered as independent identically distributed two-dimensional random vectors. As practical experience shows, the linear pairwise Pearson correlation coefficient is positive and noticeably different from 0. Using the least squares method, you can obtain a linear relationship between height and weight, which allows you to predict height based on a person’s weight (with a certain accuracy, which is expressed using confidence limits). However, it is obvious that this dependence cannot be used for control - a change (for example, a decrease) in the weight of an adult does not lead to a change in his height.

Other examples can be discussed. So, for a city, consider such indicators as the number of televisions for a certain year, the number of murders in the city, the number of diseases and mortality in the same year. You can verify that for each two of these four indicators the correlation coefficient between them is very close to 1. As a consequence, for any of these indicators you can fairly accurately predict the value of any other. At the same time, it is clear that, for example, the complete elimination of televisions will not significantly reduce the values of the other three indicators.

The most common reason for a situation where there is a correlation between two quantities in the absence of a direct cause-and-effect relationship

between them is the presence of some other, third factor on which both of these quantities depend. For example, in the case considered above, such a factor will be the population of the city - with an increase in this value, the number of televisions in the city, and the number of murders, and the number of diseases will certainly increase, simply because people buy televisions and murders, they also suffer from diseases, and The more people in the city, the more cases of murders and diseases. We can say that in the situation there is a hidden (latent) variable - the number of residents in the city, and on this variable the listed indicators with a sharp increase in the city's population depend almost linearly, which leads to the fact that the correlation coefficients between them are close to 1.

The real connection between the quantities between which a high correlation coefficient is observed is not always obvious. Thus, based on the results of the analysis of the chronicles of some cities of the Roman Empire, it turned out that with an almost constant population of the city, over time, a direct correlation began to be observed between the number of fountains (at that time they served as the end point of the aqueducts through which clean water was supplied to the city) in the city districts and mortality in the same areas from diseases, the descriptions of which corresponded to the descriptions of myocardial infarction and stroke. At first glance, it may seem that the values of the number of fountains and the number of heart attacks are connected by the fact that, they say, water from fountains causes heart attacks, but for medical science this is nonsense - no mechanisms for such connections have ever been identified. In reality, supplying city areas with clean water from aqueducts led to a sharp decrease in the mortality rate of the population of these areas from intestinal infections and poisoning (an alternative to fountains were rivers within the city into which waste was poured, as well as wells filled with contaminated groundwater), and the number of people who ultimately lived to ages at which mortality from myocardial infarction and cerebral stroke became significant.

Thus, paradoxically, the increase in mortality from such diseases precisely indicated the improvement in the health of the population of the cities of the Roman Empire from the construction of aqueducts and fountains (although, undoubtedly, this will be shown even better by the decrease in mortality from intestinal infections).

Understanding the entire chain of deep connections between correlating variables, however, sometimes makes it possible to use information about correlation to manage the situation. Thus, Charles Darwin, in his work “The Origin of Species by Means of Natural Selection,” noted a direct correlation between the number of cats in rural areas of Great Britain and the yield of red clover there. One might think that these two indicators simply simultaneously depend on some third variable (for example, the climatic conditions of the area), but the researcher found that the chain of connections is actually more complex: red clover was pollinated only by bumblebees (but not by bees), bumblebees (unlike bees) lived in nests accessible to mice, which destroyed these nests, and cats, in turn, hunted mice. Accordingly, the more cats lived in any territory, the fewer mice remained there, the less the threat to bumblebee nests, the more bumblebees participated in pollinating clover, which increased its productivity. Therefore, Charles Darwin's advice to British farmers to "Take care of cats and breed them" was a justifiable response to the farmers' request for a scientific idea to increase the yield of clover (the main fodder crop of that time in Great Britain).

The quality of a regression model is measured by the coefficient of determination. If the probabilistic-statistical model of correlation analysis is valid, then for the paired linear regression model the coefficient of determination is equal to the square of the usual correlation coefficient between the independent and dependent variables. However, the coefficient of determination in OLS can be used more widely (for example, when one of the variables is determined) than the correlation coefficient. Therefore, we cannot say that the

coefficient of determination is equal to the square of the correlation coefficient, although the formulas are the same. Many errors when using correlation and determination coefficients in the analysis of specific practical data are associated with the unlawful transfer of the properties of the correlation analysis model to other regression analysis models [36]. A comparison of probabilistic-statistical models of correlation and regression was carried out in article [37].

In order to develop management decisions using regression dependence, a series of preliminary experiments is necessary. In it, the researcher sets (according to certain rules) the values of the independent variable and measures the corresponding values of the dependent variable. This area of applied statistics is called experimental design. The fundamental works are the book by V.V. Nalimov [38], the founder of experimental planning in our country, and reference book [39]. In our country, educational and methodological publications [40 - 43] and scientific articles [44 - 46] on the theory and practice of experiment planning continue to be published.

5. Chaddock scale

It is used to interpret the results obtained from calculating correlation coefficients in verbal form. In other words, for the convenience of presenting the values of correlation coefficients, a transition to a linguistic variable is used.

When assessing correlation, its different degrees are distinguished, for example, according to Table. 5.

Table 5

Degrees of correlation

Degree of correlation	Direct correlation	Inverse correlation
Absent	0	0
Weak	(0; 0.3)	(0; -0.3)
Moderate	[0.3; 0.5)	[-0.3; -0.5)
Significant	[0.5; 0.7)	[-0.5; -0.7)
Strongly expressed	[0.7; 0.9)	[-0.7; -0.9)
Very strong	[0.9; 1)	[-0.9; -1)
Functional	1	-1

Source: compiled by the authors [47] based on the results of the study [48].

For the first time such a linguistic scale fordegrees of correlation suggestedAmerican sociologist and statistician Robert Emmett Chaddock (1879–1940) in 1925[49].

In literary and Internet sources there are variants of the scale in Table 2, but the differences are insignificant. For example, for direct correlation the following meanings and terms are used: weak (or very weak) connection - from 0.1 to 0.3 (or from 0 to 0.3); moderate (or weak) connection - from 0.3 to 0.5; noticeable (or average) connection - from 0.5 to 0.7; high (or withstrongly expressed)connection- from 0.7 to 0.9; very high (very high, strong) connection - from 0.9 to 1.0 (or from 0.9 to 0.99). Therefore, when using the Chaddock scale, it is advisable to indicate which of the variants of this scale is meant (see also[50]).

Example 3.For the data in Example 1, the Spearman correlation coefficient is 0.88 and the Kendall correlation coefficient is 0.78. According to the Chaddock scale (Table 5), these correlation coefficients are described as “strong.” Thus, the relationship between the variables is close to monotonic. MIt can be said that changes in the values of one variable follow changes in the second variable.

Many scales of this type have been developed with various names [51], usually in honor of the researchers who proposed them.

6. Statistical significance of correlation coefficients

Using the Chaddock scale can be misleading. The fact is that in the case when the theoretical coefficient is equal to 0 (for example, when the coordinates of the random vector are independent), its sample value, due to purely random reasons, may turn out to be quite far from 0 and, on the Chaddock scale, the correlation will turn out to be, for example, significant.

Let's consider two hypotheses:

N_0(null hypothesis): the theoretical correlation coefficient is 0 (in other words, the sample correlation coefficient is not statistically significant, i.e., due to random reasons, it differs from 0);

N_1(alternative hypothesis): The theoretical correlation coefficient is not equal to 0 (in other words, the sample correlation coefficient is significantly different from 0).

Further reasoning is carried out in the same way for all three correlation coefficients under consideration - Pearson, Spearman and Kendall. Let us denote any of these coefficients as qn. The following reasoning is valid for all three cases: qn = rn (linear pairwise Pearson correlation coefficient), $q_n = \rho_n$ (non-parametric Spearman rank correlation coefficient) and, finally, $q_n = \tau_{1n}$ (normalized non-parametric Kendall rank correlation coefficient), i.e.

$$\tau_{1n} = \frac{\tau_n}{\sqrt{D(\tau_n)}} = \tau_n \sqrt{\frac{9n(n-1)}{2(2n+5)}} = \frac{P-Q}{\sqrt{\frac{n(n-1)(2n+5)}{18}}}.$$

The need for normalization is due to the fact that the variance of the Kendall correlation coefficient according to [12, 13] is equal to

$$D(\tau_n) = \frac{2(2n+5)}{9n(n-1)} = \frac{4}{9n}\left(1+O\left(\frac{1}{n}\right)\right),$$

while in similar formulas for the Pearson and Spearman correlation coefficients, instead of 4/9 there is 1.

We have proven the following theorem regarding the asymptotic distribution of the correlation coefficients under consideration.

Theorem. If the null hypothesis (H0) is true that the theoretical correlation coefficient is equal to 0, the distribution of the corresponding sample correlation coefficient is asymptotically normal with a mathematical expectation of 0 and a variance of 1/n, i.e. for all x the limit relation is valid

$$\lim_{n\to+\infty} P(\sqrt{n}q_n < x) = \Phi(x),$$

Where $\Phi(x)$ - function of standard normal distribution with mathematical expectation 0 and variance 1.

*Proof*carried out using limit theorems of applied mathematical statistics[5, chapter 4].

This theorem allows one to construct (asymptotic) decision rules. As usual when testing statistical hypotheses, the critical values in the decision rules are determined by the significance level α, which is specified by the researcher.

The two-sided alternative hypothesis is that the theoretical correlation coefficient differs from 0, and it is unknown whether it is positive or negative. Then the decisive rule is this.If $\sqrt{n}|q_n| < C(\alpha)$, then the null hypothesis is accepted (at the significance levelα), i.e. there is no reason to assert that the corresponding correlation coefficient differs from 0. If $\sqrt{n}|q_n| \geq C(\alpha)$, then the alternative hypothesis is accepted, i.e. the correlation coefficient is significantly different from 0. Here for the critical value $C(\alpha)$ equality is true

$$P(\sqrt{n}\,|\,q_n\,| \geq C(\alpha)) = P(\sqrt{n}q_n \leq -C(\alpha)) + P(\sqrt{n}q_n \geq C(\alpha)) = \Phi(-C(\alpha)) + 1 - \Phi(C(\alpha)) = \alpha.$$

Hence,

$$2 - 2\Phi(C(\alpha)) = \alpha,\ \Phi(C(\alpha)) = 1 - \frac{\alpha}{2},\ C(\alpha)) = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right).$$

For the most commonly used significance level $\alpha = 0,05$ critical value $C(0,05) = \Phi^{-1}(0,975) = 1,96.$

Under a one-sided alternative hypothesis, it is known in which direction the theoretical correlation coefficient deviates from 0, but it is not known by how much it deviates. For definiteness, let the theoretical correlation coefficient be positive under the alternative hypothesis. Then the decisive rule is this.If $\sqrt{n}q_n < D(\alpha)$, then the null hypothesis is accepted (at the significance levelα), i.e. there is no reason to claim that the corresponding correlation coefficient is positive. If $\sqrt{n}q_n \geq D(\alpha)$, then the alternative hypothesis is accepted, i.e. the correlation coefficient is positive. Here for the critical value $D(\alpha)$ equality is true

$$P(\sqrt{n}q_n \geq D(\alpha)) = 1 - \Phi(D(\alpha)) = \alpha.$$

Hence,

$$\Phi(D(\alpha)) = 1 - \alpha,\ D(\alpha)) = \Phi^{-1}(1-\alpha).$$

For the most commonly used significance level $\alpha = 0,05$ critical value $D(0,05) = \Phi^{-1}(0,95) = 1,64.$

Thus, the result of testing a statistical hypothesis depends on the value $\sqrt{n}q_n$. Let the value of the correlation coefficient calculated from the sample be 0.7. According to the Chaddock scale (Table 2), this degree of correlation is “withstrongly expressed." However, suchthe sample correlation coefficient differs significantly from 0 only in the case $0,7\sqrt{n} \geq 1,96,\ n \geq 8.$ Similarly, a moderate correlation can differ significantly from 0 only in the case $0,3\sqrt{n} \geq 1,96,\ n \geq 43.$ If n = 4, then any sample value of the correlation coefficient will not differ significantly from 0. From the above it is clear that the assessment of the degree of correlation on the Chaddock scale up to sample sizes of several tens of units must necessarily be considered together with the results of the significance test.

Above we considered correlation coefficients in the nonparametric case. Let us compare the above with the well-known results regarding the bivariate normal distribution. As already mentioned, to test hypotheses about the Pearson correlation coefficient in the general case, strictly speaking, you cannot use tables calculated under the very particular assumption of normal distribution of sample elements. If this assumption is true, the random variable

$$t = r_n\sqrt{\frac{n-2}{1-r_n^2}}$$

has a Student distribution with (n - 2) degrees of freedom (see, for example, [12, p. 76]). It is known that as the sample size increases, the Student distribution converges to the standard normal distribution. Replacing the quantile of the Student distribution with the limiting normal quantile, we obtain a decision rule for testing the hypothesis that the theoretical correlation coefficient is equal to 0

in the form: |t| < 1/96 - we accept the hypothesis, otherwise we reject it (at a significance level of 0.05). We have

$$|r_n|\sqrt{\frac{n-2}{1-r_n^2}} \le 1{,}96, \quad \frac{|r_n|}{\sqrt{1-r_n^2}} \le \frac{1{,}96}{\sqrt{n-2}},$$

and therefore, with a sufficiently large sample size, n can be replaced by (n - 2) by n. If the hypothesis of equality 0 is true, the theoretical correlation coefficient can be replaced$1-r_2^n$ by 1 (since the sample correlation coefficient tends to the theoretical correlation coefficient). We obtain the decision rule in the form

$$|r_n| \le \frac{1{,}96}{\sqrt{n}}, \quad |r_n|\sqrt{n} \le 1{,}96,$$

those. in the form that was considered in the general case above.

Example 4.For example data 1 sample size n =10, the Spearman correlation coefficient is 0.88, and the Kendall correlation coefficient is 0.78. The test statistics for testing the hypothesis that the Spearman coefficient is equal to 0 is equal to $0{,}88\sqrt{10} = 2{,}78$. Since 2.78 > 1.96, the alternative hypothesis is accepted, i.e. the theoretical correlation coefficient is not equal to 0. Therefore,the correlation is reliably true,variables are dependent. Nthe standardized nonparametric Kendall rank correlation coefficient is equal to

$$\tau_{1n=} = \tau_n\sqrt{\frac{9n(n-1)}{2(2n+5)}} = 0{,}78\sqrt{\frac{9\times10\times(10-1)}{2(2\times10+5)}} = 0{,}78\sqrt{\frac{810}{50}} = 0{,}78\times4{,}025 = 3{,}14.$$

Since 3.14 > 1.96, the null hypothesis is rejected, the Kendall correlation coefficient is statistically significant,variables are dependent.

7. Concluding remarks

The article analyzes the main issues related to correlation analysis that arise during statistical processing of real data. Let us discuss some directions for further research.

The recommendations obtained are asymptotic. Their errors require study. In further studies, it is advisable to study the rate of convergence in the obtained asymptotic expressions, in other words, to consider finite sample sizes. To

assess the errors of the limit theorems obtained in the article, the method of statistical tests can be applied according to [52]. Perhaps amendments will be useful. Thus, the nonparametric Spearman rank correlation coefficient is asymptotically normal with the parameters:

$$M\rho_n = 0,\ D(\rho_n) = \frac{1}{n-1}$$

(see, for example, [12]). Perhaps, in the above decisive rule for testing the hypothesis that the theoretical correlation coefficient is equal to 0, it is advisable to replace $\sqrt{n}\rho_n$ on $\sqrt{n-1}\rho_n$.

Another example is the parametric t statistic, designed to test the equality of the Pearson correlation coefficient to 0 under the assumption of normality, and it may be advisable to use it in a nonparametric setting.

It seems natural to build decision rules based on the asymptotic normality of the sample Pearson correlation coefficient. However, formula (27.8.1) for the variance of the sample correlation coefficient given in the classic monograph by G. Cramer [53] on p. 393 is incorrect. This is clear from the fact that it gives zero variance if the hypothesis that the theoretical correlation coefficient is equal to 0 is true. According to what was said above, the dispersion should not be equal to 0, but 1/ n (in asymptotics). Formula (27.8.2) in [53] is also incorrect - there is no need to square the numerator on the right.

It is important to study the impact of elections on the values of correlation coefficients. Let's pay attention to Bernstein's example. Academician USSR Academy of Sciences S.N. Bernstein, back in 1932, considered [54] the following problem: “Determine the smallest possible value of the Pearson correlation coefficient R between the quantities X and Y, if it is known that their mathematical expectations are equal to 0 and that there are two constants L and λ such that always

$$0 \le \lambda \le \frac{Y}{X} \le L$$

."

Having solved this problem in the general case and considered a numerical example, S.N. Bernstein ends the article [54] this way: "... it is enough that only one out of 701 individuals disobeys the prevailing law of proportionality Y = 0.1X for the correlation coefficient to drop to a value of 0.198" (see also [37]).

This article is adjoined by works on indicators of the relationship of characteristics and testing of independence in contingency tables. We are talking about variables measured in name scales (nominal). They are also called categorized variables. As stated in the classic monograph [55, p. 720], "historically, much of the literature on categorized variables has been devoted to the task of testing for the presence and measurement of interdependence between two such variables." Chapter 33 of the monograph [55] is devoted to these problems (see also [56]).

Literature

1. Shamsuvaleeva A.M., Orlov A.I.The use of correlation and concordance coefficients // Thirteenth Charnov Readings. Collection of proceedings of the XIII All-Russian Scientific Conference on Organization of Production. – M.: MSTU im. N. E. Bauman, NP "Association of Controllers", 2023. – P.171-180.

2. Orlov AI Basic requirements for statistical methods of data analysis / Polythematic Online Scientific Journal of Kuban State Agrarian University. 2022. N 181. P. 316-343. – DOI 10.21515/1990-4665-181-026. – EDN OKGBOS.

3. Orlov A.I. Controlling of statistical methods / Controlling. 2022. No. 4(86). pp. 2-11. – EDN JGCVTT.

4. Orlov A.I. On the requirements for statistical methods of data analysis (general article) // Factory Laboratory. Diagnostics of materials. 2023. T. 89, No. 11. P. 98-106. – DOI 10.26896/1028-6861-2023-89-11-98-106. – EDN VEWJXD.

5. Orlov A.I. Applied statistics. - M.: Exam, 2006. - 671 p.

6. Kendal M. Rank correlations. - M.: Statistics, 1975. - 216 p.

7. Orlov A.I. Sustainability in socio-economic models. - M.: Nauka, 1979. - 296 s.

8. Orlov A.I. Sustainable economic and mathematical methods and models. - M.: IP Ar Media, 2022. - 337 p.

9. Agalarov Z. S., Orlov A. I. Econometrics: textbook. - M.: Dashkov and K, 2021. - 380 p.

10.Orlov A.I. Distributions of real statistical data are not normal / Scientific journal of KubSAU. 2016. No. 117. pp. 71–90.

11. Orlov A.I. Coincidence analysis model for calculating nonparametric rank statistics / Factory laboratory. Diagnostics of materials. 2017. T.83. No. 11. pp. 66-72.

12.Bolshev L.N., Smirnov N.V. Tables of mathematical statistics - M.: Nauka, 1983. - 416 p.

13. Hollender M., Wolf D. Nonparametric methods of statistics. - M.: Finance and Statistics, 1983. – 518 p.

14. Eliseeva I. I., Yuzbashev M. M. General theory of statistics: Textbook. 5th ed., revised. and additional - M.: Finance and Statistics, 2004. - 656 p.

15. YarkinaN. N. Statistics: textbook. – Kerch: FSBEI HE “KGMTU”, 2021. – 229 p.

16. Borkovskaya I. M. [et al.]. Econometrics and economic-mathematical methods and models: educational method. manual for students of economic specialties of correspondence education. – Minsk: BSTU, 2018. – 129 p. Corr-reg. Analysis

17.Baraz V.R. Correlation and regression analysis of the relationship between commercial activity indicators using Excel: a textbook. – Ekaterinburg: State Educational Institution of Higher Professional Education “USTU-UPI”, 2005. – 102 p. Corr-reg. Analysis

18. Lukyanova N. Yu. (comp.) Statistics: Correlation and regression analysis of statistical relationships on a personal computer: Guidelines for practical classes for students of all forms of study in the specialty “Management”. - Kaliningrad: Kaliningr. univ. 1999. - 35 p.

19.Bakeeva L. V., Pastukhova E. V.(comp.)Mathematics. Elements of mathematical statistics. Correlation and regression analysis: Guidelines for performing calculation tasks. – SPb.: St. Petersburg Mining University. 2019. - 42 p.

20. Grabovets O. V., Sidorchukova E. V. Correlation and regression analysis as a method for substantiating management decisions / Economics and Society. 2015. No. 6-2 (19). pp. 633-638.

21. Correlation and regression analysis // Educational portal “Handbook”. — URL: https://spravochnick.ru/ekonomicheskiy_analiz/korrelyacionno-regressionnyy_analiz/ (access date: 07/31/2024).

22. Bolshakova L. V., Litvinenko A. N. Methodology for using a statistical analysis package for conducting correlation and regression analysis in the course of economic research / Bulletin of Economic Security. 2021. No. 3. P. 259-265. doi:10.24412/2414-3995-2021-3-259-265

23. Chernichenko A. N., Naumenko S. M., Chernichenko L. L. Correlation and regression analysis of the relationship between socio-economic indicators of the organization’s activities / University Science. 2021. No. 1(11). pp. 111-120. – EDN ATVCCX.

24. Gaifullina M. M., Nizamova G. Z. Correlation and regression analysis of the investment attractiveness of the oil refining industry / Management. 2021. T. 9, no. 3. pp. 27-38. – DOI 10.26425/2309-3633-2021-9-3-27-38. – EDN GYITKK.

25. Nizamova G. Z., Gaifullina M. M. Correlation and regression analysis of the efficiency of using investment resources of an oil company / Bulletin of USPTU. Science, education, economics. Series: Economics. 2021. No. 1(35). pp. 15-23. – DOI 10.17122/2541-8904-2021-1-35-15-23. – EDN TIWUYV.

26. Mityakov S. N., Mityakov E. S., Mityakova O. I., Yakovleva G. N. Tools for assessing innovative activity in the regions: correlation and regression analysis / Innovations. 2021. No. 1(267). pp. 60-67. – DOI 10.26310/2071-3010.2021.267.1.009. – EDN EYPPMA.

27. Malakhova O. S., Kuznetsova O. A., Kuznetsov A. V. Correlation and regression analysis of the dependence of an enterprise’s revenue on factors of foreign economic activity / Applied mathematics and management issues. 2019. No. 1. P. 113-123. – DOI 10.15593/2499-9873/2019.1.08. – EDN CZLQHG.

28. Velitskaya S.V. Correlation and regression analysis of the influence of indicators on the company’s revenue (using the example of PJSC Gazprom) / Vector of Economics. 2019. No. 6(36). P. 45. – EDN KLWQTV.

29. Parkhomenko A. V., Shebunyaeva E. A., Parkhomenko V. L. Correlation and regression analysis in assessing the impact of small business development on the standard of living of the population / Application of mathematics in economic and technical research. 2019. No. 1(9). pp. 63-68. – EDN BDQWVD.

30. Mannanova M. A. Correlation and regression analysis of the impact of inflation on the level of wages in the Russian Federation / Economics and management of innovative technologies. 2022. No. 8(113). – EDN IGYTSX.

31. Leonova L. B. Correlation and regression analysis as a tool for predicting the impact of the functioning of a socio-economic cluster in the housing and communal services sector on the economy of the region / Bulletin of USPTU. Science, education, economics. Series: Economics. 2020. – No. 3(33). – pp. 57-65. – DOI 10.17122/2541-8904-2020-3-33-57-42. – EDN COJQRL.

32. Shlyushenkova K. A. Correlation and regression analysis of the dependence of the revenue of the largest TNCs of the BRICS countries in the oil and gas industry on factors of economic activity using the example of the Oil and Natural Gas Corporation / Innovative Science. 2023. No. 8-1. pp. 32-37. – EDN BILGUR.

33. Mayurova A. S., Roy V., Kustikova M. A. Correlation and regression analysis of the degree of influence of various socio-economic factors on the dynamics of the incidence of opisthorchiasis in the territory of Khanty-Mansi Autonomous Okrug-Yugra / Life Safety. 2021. No. 5(245). pp. 50-55. – EDN PHHEYU.

34. Orlov A.I. The variety of regression analysis models (summarizing article) / Factory laboratory. Diagnostics of materials. 2018. T.84. No. 5. pp. 63-73.

35. Szekely G. Paradoxes in probability theory and mathematical statistics. – M.: Mir, 1990 – 240 p.

36. Orlov A. I. Errors when using correlation and determination coefficients/ Factory laboratory. Diagnostics of materials. 2018. T.84. No. 3. P. 68-72.https://doi.org/10.26896/1028-6861-2018-84-3-68-72

37. Orlov A. I. Probabilistic-statistical models of correlation and regression / Scientific journal of KubSAU. 2020. No. 160. pp. 130–162.

38.Nalimov V.V. Theory of experiment. - M.: Nauka, 1971. - 208 p.

39. Ermakov S.M., Brodsky V.Z., Zhiglyavsky A.A. and others. Mathematical theory of experiment planning. - M.: Fizmatlit, 1983. - 392 p.

40. Emelyanov O. V., Shakhmaeva K. E. Planning an engineering experiment: educational manual. Ed. 2nd. – Magnitogorsk: Magnitogorsk State Technical University named after. G.I. Nosova, 2022. – 78 p. – ISBN 978-5-9967-2346-1. – EDN QMFJCR.

41. Ereshchenko T.V., Dushko O.V., A.A. Churakov A.A. Planning an experiment: Educational and practical guide. – Volgograd: Volgograd State Technical University, 2021. – 80 p. – ISBN 978-5-9948-4091-7. – EDN UHKSBX.

42. Asanova N.V., Solovyova O.Yu., Kozhanova T.E. Planning an experiment: Textbook. allowance. – Volgograd: Volgograd State Technical University, 2018. – 96 p. – ISBN 978-5-9948-2948-6. – EDN RQQWJJ.

43. Ermakov A. S. Planning and organization of experiment. – M.: National Research Moscow State University of Civil Engineering, 2014. – 83 p. – ISBN 978-5-7264-0889-7. – EDN VIPTKB.

44. Grigoriev Yu. D. Experimental plans for spline-type regression models / Factory laboratory. Diagnostics of materials. 2013. T.79. No. 11. P. 60-66.

45. Grigoriev Yu. D. Q-optimal and similar experimental plans for polynomial regression on a segment / Factory Laboratory. Diagnostics of materials. 2020. T.86. No. 5. P. 65-72.

46. Ordinartseva N.P. Planning an experiment in measurements / Factory laboratory. Diagnostics of materials. 2013. T.79. No. 3. P. 72-76.

47. Shamsuvaleeva A.M., Prokhorov S.Yu., Orlov A.I., Pivkin A.L., Leus N.A. Formation of an integral indicator - an index of countries' readiness for space activities / EconomistAspace. 2024. No. 1(7). pp. 28-42.

48.Bavrina A. P., Borisov I. B. Modern rules for using correlation analysis / Medical almanac. 2021. No. 3. P. 70-79.

49. Chaddock RE Principles and methods of statistics. – Boston, New York [etc.]: Houghton Mifflin. 1925. 471 p.

50. Kadochnikova E.I., Varlamova Yu.A. Statistical analysis of spatial data: textbook. – Kazan: Kazan University Publishing House, 2023. – 140 p.

51. Koterov A. N., Ushenkova L. N., Zubenkova E. S., Kalinina M. V., Biryukov A. P., Lastochkina E. M., Molodtsova D. V., Vainson A. A. Sila communications. Message 2. Gradations of the correlation value / Medical radiology and radiation safety. 2019. T. 64. No. 6. P. 12–24.

52. Orlov A.I. Method of statistical testing in applied statistics / Factory laboratory. Diagnostics of materials. 2019. T.85. No. 5. pp. 67-79.

53. KramerG. Mathematical methods of statistics. - M.: Mir, 1975. - 648 p.

54. Bernshtein S.N. On one elementary property of the correlation coefficient / Zap. Khark. math. Comrade 1932. T. 5. P. 65-66.

55. Kendall M.J., Stewart A. Statistical inference and connections. – M.: Nauka, 1973. - 896 p.

56.Muravyova V. S., Orlov A. I. Statistical analysis of tables of four fields / Polythematic network electronic scientific journal of the Kuban State Agrarian University. 2021. No. 174. pp. 285-314.