Establishment of existence and extent of interrelation between different data groups characterizing branch operation correlation analysis

Boris Kobilyanskiy

PhD.,
Department of health and environmental safety
Teaching and research professional education institute of
Ukrainian engineering and pedagogical academy

Abstract
In this paper there evaluated occupational risk assessment methodology using mathematical methods of regression and correlation allowing to determine the priority areas for the development of measures to prevent and reduce occupational hazards and occupational risk. The ranking of factors of production is particularly important in conditions of limited resources.
Key words: OCCUPATIONAL HAZARDS, OCCUPATIONAL RISK, REGRESSION, CORRELATION.

The human security problem in the production is almost always solved with limited economic opportunities. Hence the need to develop such methods of OSH management, which would allow the management decisions that ensure maximum social impact with limited resources. The implementation of such an approach is possible only through the construction and study of relevant mathematical models that take into account the specific conditions of production.

The problem of human security in the production are almost always solved with limited economic opportunities. Hence the need to develop such methods of OSH management, which would allow the management decisions that ensure maximum social impact with limited resources. The implementation of such an approach is possible only through the construction and study of relevant mathematical models that take into account the specific conditions of production.

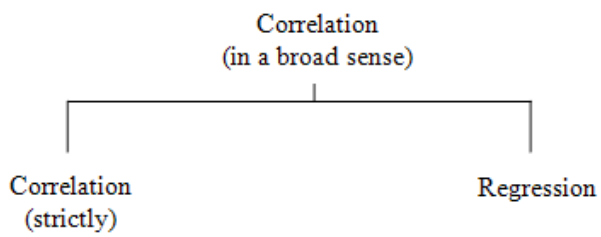
In market conditions may conduct occupational risk assessment in specific organizations, because the employer is obliged to carry out certification of workplaces on working conditions, inform employees of the conditions and safety in the workplace, on the existing health risks. However, the correlation factors of working conditions with workers health indicators studied is not enough.

The concept of regression and correlation is directly linked. While in the correlation analysis estimated the strength of stochastic communication, regression analysis is studied its shape. With the help of importance evaluation there solved a question on the objective existence of real connection. The correlation and regression analyses have many common computing

procedures. Both types of analyses are used to establish causal relationships between phenomena and to determine the presence or absence of communication.

Functional and correlation communication defines the relationship between the phenomena and processes. It should be emphasized that any causal effect can be expressed either by functional or correlation connection. But not every function or every correlation corresponds to a causal relationship between phenomena.

Thus, the relationship between regression and correlation can be presented in the form



For effective study of connections, it is necessary to use aggregates, homogeneous in respect to those signs, which connection is being studied. If the time spent by an employee per unit of product in the enterprises is determined, differing only to the technical level of production, it should be expected that in this case there will be very close connection between the signs. The closer the connection between phenomena, so, consequently, the greater the action of secondary causes is excluded and the less impact of occasional influence. As a result, correlation is close to functional. Therefore, the functional relationship can be seen as a limiting case of correlation. Between technical phenomena there mainly act objectively existing correlations. However, in this case it is necessary to distinguish clearly between correlation and functional relationship.

Correlation between two variables can go into a functional relationship, if some variables connected in a certain way are considered simultaneously.

Thus, W is connected both with X and Y. If we examine the connection between W and X, or W and Y, then the value W is as defined for given values of X and Y. In this case, W can be viewed as a random variable in a statistical sense . Between W and X and between W and Y there is a correlation However, if we look at the same time W, X and Y, then W loses properties of a random variable, and correlations are moving together in functional relationship in the form W = X + Y. The value W is a function of two variables X and Y and are determined by them.

It should be noted that sometimes the true functional relationship is difficult to detect because of the overlapping measurement errors, changes in the conditions of implementation, erroneous or formal consideration of causal relationships. Non-random variables, which are in functional dependence are transformed into the random ones and the connection begins to acquire a stochastic character. For example, the law of free fall is done accurately only in a vacuum. When deviations from the conditions of the law manifests itself in the form of correlation. However, it is clear that the sum of angles in polygons of one species is not a random variable. It seems such measurement errors due to overlapping. In fact, between the number of sides n and the sum of its angles (S) there is a deterministic relationship, described by a function S = (n - 2)*180.

We have already mentioned that the causal effect can be expressed in the form of functional or correlation. But it does not follow the converse that causal relationship is at the back of any correlation or functional link is hidden. Firstly, this is due to the variety of forms of cause-effect relationships; secondly, from the definition of functional and correlation connection one can see that it is a reflection of the quantitative relationship between the phenomena, or the assessment of the communication numerical data. The task research is to find causal relations. Only the knowledge of the true reasonscauses of the phenomena can correctly interpret the observed patterns. However, the correlation as a formal statistical concept itself does not reveal the nature of the causal link. Correlation analysis can not specify how to take the phenomenon as the cause, and which - as a consequence. Correlation only assesses strength, or distress of connection.

The question of a causal relationship between the phenomena in each case is decided on the basis of a logical researcher and professional considerations that must precede the possible correlation analysis. However, the last requirement should not be a prerequisite, as sometimes the explanation of cause and effect can be obtained only after empirical description of communication. There is no doubt that, in any case, the method of mathematical statistics is a very useful tool for opening the connections between phenomena.

In many situations, it is relatively easy, on the basis of logical and professional reasons, to explain which variables are the cause and what is the effect. So, there is a correlation between the growth of labor productivity and wage increase. In general, the growth of labor productivity can be considered as the reason for higher wages. But on the other hand, wage increases could be material stimulus for growth in labor productivity. There is no question, which var-

iables to take as a cause or consequence. However, sometimes it is difficult to figure out the relationship between the variables.

Regarding the nature of the correlations there distinguished:

- positive correlation. It takes place, if increase or decrease of the value of another one variable increases or decreases, respectively. A positive correlation exists, for example, between labor productivity and wages, between height and weight, between the technical level of production and productivity, between the execution of the production plan and the cost of working hours, etc. A positive correlation is also called as direct correlation;

- negative correlation. With the increase or decrease in the value of one variable, the values of another one are reduced or increased respectively. Negative correlation exists for example between productivity and cost of the product, between the volume of production and the cost per unit of product, etc. The negative correlation is also called as feedback.

Regarding the number of variables there are the following types of correlations:

- Simple or pair correlation. This is the correlation between two variables. For example, between income and consumption, between profit and cost,etc.;

- Multiple correlation. This is the correlation between more than two variables. For example, between labor productivity, the level of mechanization of production, skilled workers, the level of use of computer time; between energy consumption, production volume and temperature of the environment. With the help of multiple correlation, we are trying to cover the whole causal complex. This is particularly important where the individual conditions are generally a consequence of not the one but several reasons. Multiple correlation is a reflection of the objectively existing multiple ties. The establishment of these relations followed by their specific explanation reveals the mechanism of the phenomena;

- **Partial correlation. This is the correlat**ion between the two variables during "fixed" influence of the other variables included into analysis. Using partial correlation there fully investigated causal complex and internal structure of relations is revealed. The importance of using partial correlation flows from the fact that, as a rule, multiple causes interact at the same time and have a joint effect on the studied property. If one defines the correlation between the dependent variable (effect) and each explanatory variable (reason) separately, the effect of other variables will affect the connectedness degree of the selected variables. This can lead to erroneous conclusions. For the study, depending on the volume flow rate of steam production at one of the companies producing prefabricated concrete structures under the open sky, there was established negative correlation, i.e., with the increase in production volume there decreased steam consumption. But this is obviously a paradoxical conclusion. Careful analysis has shown that another factor influences significantly on the consumption of steam, namely air temperature. Herein negative correlation between these variables is so strong that the conclusion of an association between consumption of steam and the volume of production cannot be sustained. Therefore, before determining the correlation between the consumption of steam and production volume, one should exclude the effect of temperature on the steam consumption. While determining the correlation between air temperature and steam consumption one should also exclude the impact of production on the steam flow. As a result, we have two partial correlations, each of which indicates a «clean» stochastic relationship between two variables by elimination of the effect of the third.

Regarding the forms of connection there distinguish the following types of correlations:

- Linear correlation. At this type of correlation between the studied variables there exists linear relations;

- Non-linear correlation. At this type of correlation between the studied variables there exists nonlinear relations.

Regarding the type of connection there distinguished:

- Direct correlation. In this case, the studied phenomena are interconnected directly. Explanatory variable has a direct effect on the dependent variable. The direct correlation exists, for example, between labor productivity, technological level of production and production worker skills; between productivity and the cost of the product; between the availability and turnover of working capital; between the loss of working time and the volume of production, etc. So, there is a direct correlation, if one phenomena is logically different, and to explain this correlation there is no need to involve other phenomena;

- Indirect correlation. One may speek about indirect correlations when the studied variables do not have direct causal link, and are determined by their common cause. Logically, such connection can only be explained by other phenomena. If there is a risk of indirect correlation transition to the formal way of research that could lead to spurious correlation. So, one can conclude in different cases profoundly different in their inner meaning. Correct interpretation of

the planned communication is particularly significant when the statistical knowledge is drawn to the justification of vital decisions and practices. Here knowledge of connections remaining without interpretation or misinterpreted, often worse than complete ignorance. Lack of attention to this fact is one of the worst crime statistics.

- False correlation. Under the false correlation (correlation nonsense) is meant purely formal connection between phenomena that are not any logical explanation and are based only on the quantitative relationship between them. Often false correlation arises in the study of time series. This is especially true for economic phenomena. When the location of the material by year or month is easy to detect evolutionary component, showing the main trend of the series. When comparing the series of this type it is necessary (before installing the correlation between the two rows) to excluded from them the regular changes of the level. Coincidence or oppositely of evolution trends without explaining the general and non-community development, can cause an artificial connection, devoid of meaning. This relationship does nothing to investigate the causes that govern phenomena.

Let us assume two features are measured among n objects or variables x and y. The result is a single values x_1, x_2, ... x_j ... x_n and y_1, y_2 ... y_n ... y_j. Initially, both series of observations are considered separately and for each of them we calculate the statistical characteristics. The most important of these characteristics is the average value of x, which is obtained by dividing the sum of individual values in n. The formula is shown in Table 2 line. [5]. To indicate the sum, there used index of summation with corresponding indexes

$$\sum_{j=1}^{n} x_j = x_1 + x_2 + \ldots + x_j + \ldots + x_n$$

Exept the average value, there determined measure of the deviation of value of each variable from this average. To do this, first one should determine so-called sum of squared deviations of individual values from the average (abbreviated as SSD). In Table 1, line 3 there shown the expression of the sum of squared deviations, or designated S_{xx} or S_{yy}, but more convenient for computation formula in line 4. If you divide the sum of squared deviations in *(n – 1)* so-called degrees of freedom, we obtain the dispersion (line 5). Extracting the square root from the dispersion characteristics, we obtain the one, called the standard deviation (line 6). Standard deviation is a measure of the average deviation of the individual values from their average.

Up to this point every variable was considered separately from the values of each one there were calculated mean and standard deviation. Now let us ask how we can draw conclusions about one value according to another one. This task lieing in that how to judge about the value y using tha value x; that is the task of regression calculation. For graphic representation of both variables there used rectangular coordinate system xOy. Any pair of values (x_j, y_j) for each object corresponds a point on the graph. If we consider the accumulation of points on the graph, you can see that with an increase in the value of x, y also increases. Now you need to draw through a cluster of points the line so that on the basis of x "as close as possible" to estimate the value of y. These assessments of y are the most accurate, if the sum of the squares of the vertical deviations from the actual values is the smallest possible. So, we should find the parameters of the line

$$y = bx + a \tag{1}$$

from the condition

$$\sum_j (y_j - \hat{y}_j)^2 = \min. \tag{2}$$

There acceptable formulation of the problem, in which there is no interest in the direction and form of dependence, and would like to know how strong is the connection between the two series of observations relating to the same objects.

It is the task of correlation calculations. The correlation coefficient is a measure of the linear relationship between the two measured values. It can take values between +1 and - 1. If it is zero, a linear connection between x and y is absent. If it is equal to + 1 or - 1, the connection is strictly linear. Fig. 1 shows schematically the possible correlation of the field at different values of the correlation coefficients. Diagram A shows randomly scattered points on the coordinate plane. The value of x cannot show the y value. The connection between x and y is absent, r_{xy} = 0, or does not significantly differ from zero. Diagram B shows all the points lying in one line. To each value of x can be uniquely assigned a value of y. The larger x, the larger y is. If this line corresponds to the regression equation expressing the relationship between the test and productive traits, the equation can be used to find both y according to x, and vice verca. Such an extreme case where the coefficient of correlation is exactly equal to + 1, practically does not occur. On the field of the correlation shown in Graph D, the points spread not accidentaly and have a tendency to stabi-

lize in a particular direction. Such a situation occurs frequently. The larger x, the greater y. The linearity of this relation is expressed by the coefficient of correlation, which in this case is approximately equal to + 0.50. As Compared with the diagram B linear relationship under the influence of non-systematic noise spreads so that the picture seems to gloss over. In such a case, depending on the setting there calculated regression equation with respect to x or y, or x to y. There may be an error when moving from one equation to the other one by interchanging arguments and functions. The magnitude of this error depends on the value of the correlation coefficient.

Diagram C, as well as B reflects strictly linear relationship between x and y. Direct, however, does not pass through the center of coordinates. Moreover, y increases with decreasing x, and vice versa. Therefore, the correlation coefficient is negative. Thus, the negative sign of the correlation coefficient indicates the inverse linear relationship between x and y, and a positive sign - a direct linear relationship, i.e. increases with increasing x and y. The slope of the regression line has no effect on the value of the correlation coefficient, or a sign.

Table 1. Correlation calculation formula

Name	Line	Variable x	Veriable y
Sample size	1	n	n
Average value	2	$\bar{x} = \frac{1}{n}\sum_j x_j$	$\bar{y} = \frac{1}{n}\sum_j y_j$
The sum of squared deviations (SD)	3	$S_{xx} = \sum_j (x_j - \bar{x})^2$	$S_{yy} = \sum_j (y_j - \bar{y})^2$
	4		
	5	$= \sum_j x_j^2 - \frac{(\sum x_j)^2}{n}$	$= \sum_j y_j^2 - \frac{(\sum y_j)^2}{n}$
Variance	6	$s_x^2 = S_{xx}/(n-1)$	$s_y^2 = S_{yy}/(n-1)$
standard deviation		$s_x = \sqrt{S_{xx}/(n-1)}$	$s_y = \sqrt{S_{yy}/(n-1)}$
Sum of the products of deviations	7	$S_{xy} = \sum_j (x_j - \bar{x})(y_j - \bar{y}) = \sum_j x_j y_j - \frac{\sum x_j \sum y_j}{n}$	
Covariance	8	$s_{xy} = S_{xy}/(n-1)$	
Correlation coefficient	9	$r_{xy} = S_{xy} / \sqrt{S_{xx} \cdot S_{yy}} = s_{xy}/(s_x \cdot s_y)$	
The regression equation y to x	10	$y = bx + a$, where $b = S_{xy}/S_{xx}$ and $a = \bar{y} - b \cdot \bar{x}$	
The regression equation x to y	11	$x = b'y + a'$, where $b' = S_{xy}/S_{yy}$ and $a' = \bar{x} - b' \cdot \bar{y}$	

The sign of the correlation coefficient reflects only the direction of the relationship between the two variables. The diagram D also schematically shows the field of correlations during negative correlation coefficient.

Formulas for calculating the correlation coefficient are shown in Table 1. First the sum of products of deviations is defined. We have already met with the sum of squared deviations for each variable. Instead of squaring the deviations, and then summed, as shown in line 3 of Table. 1, a single deviation from the arithmetic mean value of one variable is multiplied by the corresponding rejection of another variable and then summarized. Thus, a sum of the products of deviations S_{xy} (line 7) is obtained.

By analogy with the dispersion, which is obtained by dividing the sum of squared deviations by *n - 1*, we can calculate a so-called covariance S_{xy} for separating *n - 1*. The covariance S_{xy} as well as the correlation coefficient is a measure of the interaction between the two variables. But this figure is not normalized, i.e. the value of the covariance is dependent on the physical dimensions of the variables. The correlation coefficient is a dimensionless quantity. It is a normalized covariance. The covariance between height in inches and weight in pounds is numerically different from the covariance, estimated between height in centimeters and weight in kilograms for the same individual. However, the correlation coefficient in both cases is the same. The magnitude of the correlation coefficient does not affect the linear transformation of the measuring scale, i.e., if the measurement results on a constant increase and multiply it, then the value of the coefficient does not change.

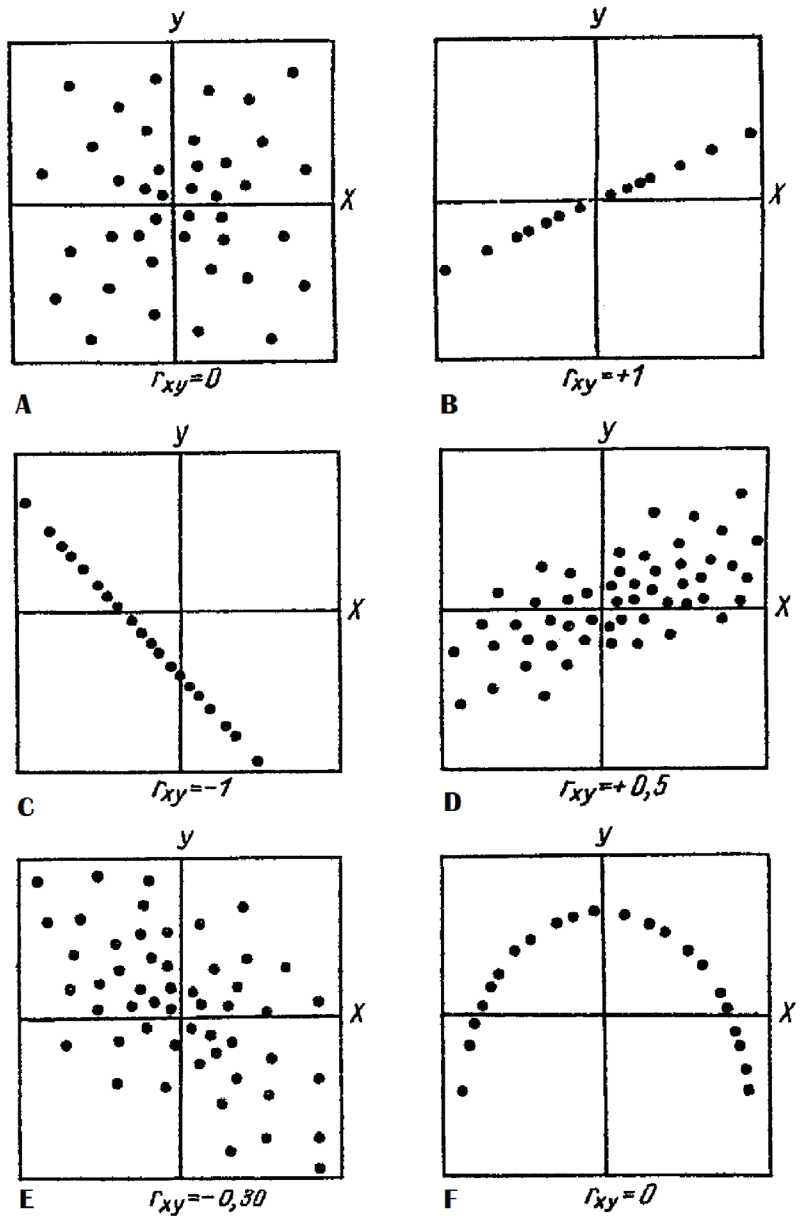


Figure. 1. Schematic representation of the different types of dependencies with the corresponding values of the linear correlation coefficient

In the literature on regression analysis of the total variance is decomposed into two components: a variable dispersion caused by regression and residual dispersion caused by errors of observations. It is known that the distance $y-\bar{y}$ consists of segments $\hat{y}-\bar{y}$ and $y-\hat{y}$. Therefore, the equality is $(y-\bar{y})=(\hat{y}-y)+(y-\hat{y})$. If both sides of this equation squared and summed over all points, we get

$$(y-\bar{y})^2=(\hat{y}-\bar{y})^2+2(\hat{y}-\bar{y})(y-\hat{y})+(y-\hat{y})^2; \quad (3)$$

$$\sum_j (y_j-\bar{y})^2=\sum_j (\hat{y}_j-\bar{y})^2+2\sum_j (\hat{y}_j-\bar{y})(y_j-\hat{y}_j)+\sum_j (y_j-\hat{y}_j)^2. \quad (4)$$

The second term on the right-hand side is twice the product of systematic and random components and the summation is zero if $(\hat{y}-y)$ and $(y-\hat{y})$ are uncorrelated. The independence of these components is a necessary condition for the model. So,

$$\sum_j (y_j-\bar{y})^2=\sum_j (\hat{y}_j-\bar{y})^2+\sum_j (y_j-\hat{y}_j)^2, \quad (5)$$

or

$$\frac{\sum(y_j-\bar{y})^2}{n-1}=\frac{\sum(\hat{y}_j-\bar{y})^2}{n-1}+\frac{\sum(y_j-\hat{y}_j)^2}{n-1} \quad (6)$$

The left side is called a complete dispersion of the variable y. The first term on the right is the variance associated with the regression.

This dispersion is characterized by the dispersion of the test factor, i.e., is the so-called "explainable" dispersion. The second member of the right-hand side is the "unexplained" variance, known as the residual variance. The origin of these names is explained as follows. Deviations ($\hat{y} - \bar{y}$) dependend on regression equation, therefore they represent the effect of the regression communication. Thus, this part of the variation is explained by the regression model. In contrast, the deviation (y - $\hat{y}$) can vary randomly and cannot be explained by the model, in this case - a linear, i.e. these variations reflect the influence of random factors. The quotient of the variance due to regression to the total variance is called the coefficient of determination. The coefficient of determination is used as a characteristic variation in the proportion of total dispersion due to the influence of the factor of x in the case of linear regression.

$$r_{xy}^2 = \frac{\text{the variance due to regression}}{\text{complete regression}} = \frac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\bar{y})^2}. \quad (7)$$

The coefficient of determination is changed from 0 to 1. Remove the square root of this factor, we obtain the correlation coefficient r_{xy}.

Before calculating the correlation coefficient, one should test the hypothesis of normality of both distributions and linear relationship between them. In general, quite carefully scrutinize in the correlation. Testing the hypothesis of normal distribution produced by a test χ^2. In conclusion it should be pointed out that the interpretation of the correlation coefficient need to be as careful.

There are many examples where there is a high correlation coefficient in the absence of a causal link between the phenomena only by the heterogeneity of the material.

Thus, the occupational risk assessment methodology using mathematical methods of regression and correlation will determine the priority areas for the development of measures to prevent and reduce occupational hazards and occupational risk. The ranking of factors of production is particularly important to prioritize limited resources.

References

1. Sharikova L. *Labor protection. A course of lectures for managers and specialists of labor protection*. Nizhny Novgorod, Biota-plus, 2007, p.172.
2. Burkov V., Javakhadze G. *Economic-mathematical models of management of development of industrial production.* Institute of Control Sciences, 1997, p.64.
3. Eliseeva I.. *Econometrics*. Moscow, Prospekt, 2009, p.288.
4. Gmurman V. *Probability theory and mathematical statistics. A manual for schools. 10th edition*. Moscow, Higher School, 2004, p.479
5. Shmoilova R. *General Theory of Statistic. Textbook. 3rd edition, revised*. Moscow, Finance and Statistics, 2002, p.560.
6. Asaev R., Akhmetov K., Imasheva A., Shalgynbaeva G. *Econometrics.* Almaty, Agro University, 2007, p.231.