UDC:	519.233.5
	Original scientific paper
Receved	August 15, 2022.
Accepted:	August 29, 2022.
Corresponding author:	larix.mb@gmail.com

STATISTICAL ANALYSIS OF THE RELATIONSHIP BETWEEN OBSERVED CHARACTERISTICS

Marković Branka

Faculty of Ecology; Independent University, Banja Luka, Bosnia and Herzegovina,
e-mail: larix.mb@gmail.com

Ružica Đervida

Faculty of Economics; Independent University, Banja Luka, Bosnia and Herzegovina,
e mail: ruzica.djervida@nubl.org.ba

Abstract: *All statistical surveys are conducted using statistical methods, where one phenomenon can be analyzed independently of other phenomena, where in that case we consider such an analysis as one-dimensional. On the basis of the results obtained during the analysis, a conclusion is made, ie. a decision is made.However, individual phenomena in nature and society cannot always be presented and observed individually because they are intertwined.There is a certain interdependence between them.Such a relationship between phenomena is simply a consequence of a certain legality, regularity.In statistical research of time series, one basic feature is that based on data from the past and present, future events can be predicted, while at the same time, the existence of relationships between phenomena is determined.This analysis does not represent a cause and effect analysis, but determines the relationship between observed phenomenaIt gives an answer to the question of whether changes in one phenomenon result in a change in another phenomenon, i.e. They determine the interconnectedness of phenomena and their strength, or some would emphasize whether there is a certain agreement in the movement of variations..These interdependencies are determined by statistical methods based on the combination of variations. It is on this basis that the research presented in the paper is based.The aim is to examine a certain statistical relationship between total electricity production in BiH and exported electricity from BiH.There is much research in the literature on the association of phenomena using linear regression.The notion of correlation is briefly explained, represented by the correlation coefficient Pearson's coefficient, Spearman's coefficient.*

Key words: *correlation, regression, correlation coefficient, interrelationshi, coefficient*

INTRODUCTION

There is a certain interdependence of phenomena in nature and society. This interdependence can be analyzed using statistical methods. Using the χ2 test, it is examined whether there is a certain dependence between two phenomena (characteristics). to serve to determine the strength of the relationship between two characteristics. Such an analysis brings with it a series of shortcomings that carry with it a series of shortcomings that carry certain limitations such as: it is not possible to determine the form of the relationship that exists, it is not possible to determine the value of one variable in relation to the second variable, one can only analyze the ratio of two characteristics, no more, and analyze the ratio of nominal characteristics.

From this it can be concluded that the analysis of two phenomena using the χ2 test is applied only to nominal characteristics. Since statistics studies and analyzes many phenomena that have a quantitative character, there are statistical methods that are used to determine the interdependence of quantitative characteristics.

Interdependence ie. the connection between these characteristics can be functional or stochastic (statistical). A functional, exact or deterministic connection is a connection between variables in which any change in the independent variable causes the same change in the dependent variable. factors cannot be determined exactly by how much the dependent variable changes with a change in the independent variable, and in this relationship it is valid that one value of the independent variable corresponds to a series of possible values of the dependent variable. This is characterized by a series of economic phenomena and thus emphasizes a certain degree of uncertainty in the realization of the expected value.

The statistical method used to study the relationship between two or more quantitative variables is called the correlation and regression method. The existence of the strength of the relationship between two variables can be studied, and in the course of the case, simple regression and correlation analysis are discussed. Although it is called simple, it refers to the number of variables and not to the complexity of the relationship. If it is a relationship between three or more variables, in that case it is called multiple regression and correlation. This type of statistical analysis can be carried out within descriptive and inferential statistics. Descriptive analysis represents the process of determining numerical indicators and indicators that enable the final decision to be made. The basic task of regression is to determine the regression model that will best describe the relationship that exists between the variables and to use that model to evaluate and predict the values of the dependent variable y for a precisely determined value of the variable x.

1. CORRELATION AND REGRESSION

As already mentioned, descriptive analysis includes the analysis of a statistical or stochastic connection between two or more variables. If the goal is to analytically express the existence of a relationship between two or more phenomena, a regression model is used, while the degree of statistical connection between phenomena is measured by the method of correlation analysis. 1

The basic goal of statistical analysis is to determine the relationship between variables, to numerically express the degree of their connection or to present it with the help of algebraic means. regression model. With the help of a mathematical function, the interdependence of variables can be shown. The general form of that function is:

$$Y=f(x)+u$$

Since it is already known that there is no functional but statistical connection between the variables. It is the deviation of the relationship from the functional one that is represented by the variable u , which gives this relationship a statistical character. In addition to using a mathematical function, the relationship between variables can also be shown using a scatter diagram. The value of one variable is displayed on the x-axis, while the value of another variable is displayed on the y-axis. At the same time, the scatterplot serves as an aid to determine what type of function is involved in the model. The points in the diagram are obtained based on n pairs of values of certain variables and they have coordinates T(x;y). Based on the distribution of points, the first description of the shape of the connection or the shape of the function is given. The regression model represents the analysis of the

[1] Šošić I; Serdar V.(1994) Uvod u statistiku; Školska knjiga Zagreb

assessment of unknown parameters, the calculation of dispersion measures and the method used to determine the quality of the results. The regression model also shows the average agreement of changes in the investigated phenomena.

To determine the strength of the relationship between variables, correlation analysis is used, the first answer to the strength of the relationship is obtained by observing the scatter diagram, if the points are close to the direction, the correlation is higher, and as the points move away (greater dispersion), the correlation is lower. Between the variables there may be linear and non-linear correlation. With linear correlation, all points are grouped around a straight line, while with nonlinear correlation, they are grouped around a curved line. When talking about a correlation relationship, that relationship can have different relationships: if the value of one variable corresponds to the same value of another variable (if the value of one variable is small, it corresponds to the small value of another variable), there is a positive correlation: If a large value of one corresponds to a small value of another variable we are talking about a negative correlation. 2 In some cases, the correlation can be non-monotonic, if the value of one variable in a certain interval corresponds to a much smaller value of another variable. If the relationship between variables changes from positive to negative more than once, it is called cyclical correlation. If it is not possible to draw a conclusion about the value of another variable based on the value of one variable, it can be concluded that there is no correlation.

1.1. The correlation coefficient

Unlike regression, with simple linear correlation, there is no difference between the dependent and independent variables, because both variables have identical status. In this case, both variables are treated as random variables, which is why the research of one variable as a function of the other is never set. And as a consequence of random errors, a certain arrangement of points in the coordinate system appears. The dispersion in some cases can be so large that it is very difficult to determine the tendency, i.e. the trend. Based on the diagram showing the arrangement of points in the coordinate system, it can be said that it is a strong

[2] Lovrić, M., (2008) Osnovi statistike, Ekonomski fakultet, Kragujevac

connection (if the points are concentrated around one direction or a curved line). If the points are far from the imagined direction or curved line, it is said that the connection is weak (it is far from the functional relationship).

Measures used to determine the degree of statistical relationship are called correlation coefficients. They represent values that allow the existence of a straight-line agreement (functional relationship) between two observed variables. The two most commonly used correlation coefficients are: Pearson's simple linear correlation coefficient or simply the correlation coefficient, this coefficient shows the strength of the relationship between the quantitative agreement of two numerical characteristics with a linear model. Spearman's coefficient is used to determine the strength of the relationship between two non-quantitative variables that are not in a linear relationship.

1.2. Pearson's coefficient (r_{xy})

If there is a linear relationship between the two variables being analyzed, the Pearson coefficient is used to determine the degree of association. As a relative measure, the simple linear correlation coefficient has an interval value of +1 to -1. If the sign is negative, it indicates a negative direction of correlation, while it is a positive correlation if the sign is positive. If the Pearson coefficient has values closer to 1, there is a stronger correlation between the variables. There are two extreme cases, if the simple linear correlation coefficient is equal to 0 and 1. If it is 0, then there is no linear connection, and if it is 1, it is a perfect correlation. The coefficient of simple linear correlation is based on the comparison of the actual influence of the analyzed variables on each other in relation to the maximum influence of both variables. The basis for calculating the Pearson coefficient lies in the covariance, which represents the first mixed moment around the mean. Since the initial basis for calculating the variance is the deviation of the value of the variable from its arithmetic mean, it is concluded that the covariance of two variables is always equal to 0 if one variance is equal to 0, and therefore in such cases there is no correlation between the variables. This means that if the variance has a value of 0, all the values of the variables are mutually equal, such a case does not happen in practice. .It means that the measure for determining the strength of the connection between phenomena cannot be measured by covariance. That is why it is necessary to transform the variables z, that is, to express the deviation of the values of the variables from the arithmetic mean by the number of standard deviations. During the transformation, the variables retain their arrangement, but the arithmetic mean is 0, while the standard deviation is 1

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

The formula above represents the Pearson linear correlation coefficient or as it is also called the product moment. If regression analysis is used as a starting point for statistical analysis, the results obtained during that analysis are used to calculate the linear correlation coefficient. In that case, the coefficient of linear correlation is obtained using the coefficient of determination or with the help of the product of the regression coefficient and the ratio between the standard deviation of the dependent and independent variable. As already stated, the coefficient of determination varies within the interval -1 to +1. Depending on its size, the strength of the connection between the variables is also interpreted. The table below shows the values of the correlation coefficient in absolute terms and their interpretation

Table 1 Values of the correlation coefficient

Linear correlation coefficient	Interpretation
0	There is no correlation
0-0,5	Weak correlation
0,50-0,80	Medium strong correlation
0,80-1,00	Strong correlation
1,00	Perfect correlation

(Source: I.šošić,V.Serdar (1994) Uvod u statistiku, školska knjiga Zagreb)

However, if there is not all the data, but a smaller part, the sample, in that case the correlation coefficient of the set is estimated based on the correlation coefficient of the sample. In order to be able to calculate the confidence interval of the estimate of the Pearson coefficient of a statistical set, the transformation procedure is used, whereby a linear correlation coefficient is obtained, whose sample distribution has the form of a normal distribution. Even with small samples, the transformed z-distribution is approximately the same as the normal one, so the tabular values of the areas under the normal curve are taken as the estimation coefficient. The tabular values represent a function of the significance level α and the number of samples n. If the obtained values are greater than the tabular values, in that case there will be a correlation with significance α-1. In most cases, the significance interval is taken to be 95%.

In the previous part, we talked about a relationship that is linear, which does not mean that the relationship between phenomena can be curvilinear. When using the Pearson coefficient, it is necessary to remove all values in the set or sample that are extreme, because in this way the influence on the final result of the coefficient is eliminated.

1.3. Spearman's coefficient (r_s)

Spearman's correlation coefficient is a type of coefficient that is exclusively used for those variables where Pearson's coefficient cannot be used. If it is a matter of establishing the relationships between variables that are given in the form of rank, Pearson's coefficient cannot be used because rank variables do not have metric properties.

The starting point for determining the Spearman coefficient is that the rank modalities must take the given values of the first n numbers (related), and the size for measuring the correlation relationship, the difference in ranks is taken as the starting point. while unlike the Pearson coefficient, the form of the connection itself is not absolutely important.

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n}$$

The value of the Spearman coefficient ranges in the interval $-1 \leq r_s \leq 1$, where d is the difference between the values of the ranks of the two analyzed variables. If the coefficient takes the value +1, it means a complete positive rank correlation, and it will take the value of -1 if the order of modality of one variable is inversely proportional to the order of modality of the other variable. Only making conclusions about the strength can be made by comparing the values obtained with the limit value of the Spearman coefficient.

2. ELECTRICITY INDUSTRY OF B&H

To determine the relationship between the variables, we took the relationship between the production of electricity in Bosnia and Herzegovina and the export of electricity from Bosnia and Herzegovina. Before the actual analysis, the way of functioning of Electricity industry of B&H should be explained.

Electricity industry of B&H consists of four separated companies: JP Elektoprivreda BiH; Elektro privreda Hrvatske zajednice Herceg Bosne; Elektroprivreda Republike Srpske.

JP Elektroprivreda BiH is the largest electric utility company in BiH. Within JP Elektroprivreda BiH there are production facilities: HPP on the Neretva (Jablanica; Grabovica; Salakovac), Thermal Power Plants (TE Kakanj; TE Tuzla). In 2021, the Podveležje-Mostar wind plant was put into operation.

Elektroprivreda of the Croatian Community of Herceg Bosna deals with production, maintenance and research. The production plants are located on the basins of the rivers: Neretva and Vrbasa (HE Jajce I i II; HE Rama; HE Mostar; HE Peć; HE Mlini; HE Mostarsko blato; CHE Čapljina.

Elektroprivreda of Republika Srpska, due to its characteristics, the most important hydropower potential is located in Vrbas; Trebišnica and Drina, and includes two thermal power plants, TE Gacko and TE Ugljevik.

Based on data from the economy, the basic assumption is that between certain variables, which are analyzed, i.e. that between the production of electricity in BiH in the period from 2017-2021 and its export in GW/h there is a linear relationship.

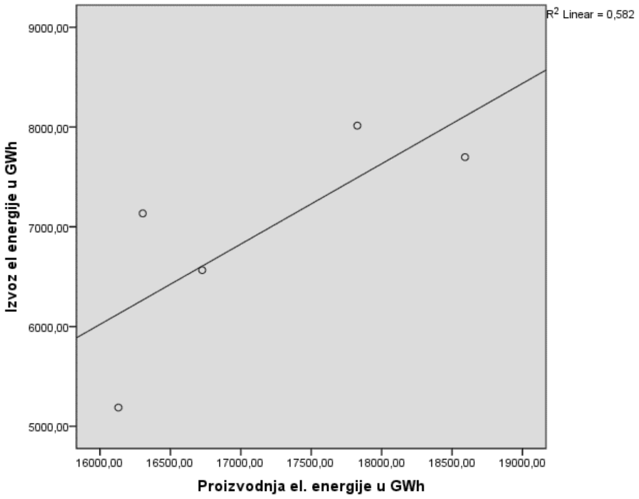
In order to determine the linear relationship, a simple linear regression model is used. It is assumed that the independent variable x is the total production of electricity in BiH in the period 2017-2021, while the dependent variable y is the export of electricity in GW/h

Table 2 Production, export, import of electricity in Bosnia and Herzegovina (GW/h)

Year	Production	Export	Import
2017	16.131	5.188	3.345
2018	18.592	7.698	3.092
2019	16.726	6.565	2.824
2020	16.304	7.314	3.266
2021	17.827	8.014	3.259

(Source: Statistical Agency of Bosnia and Herzegovina)

The first step in the analysis is a graphical display - a scatter diagram of the relationship between the production of electricity in B&H and the export of electricity from B&H



Graph 1 Scatter diagram
(Source: Author's calculation)

The scatter diagram can confirm the original assumption that there is a linear relationship between the observed variables.

The SPSS program is used to determine the relationship between variables and to draw conclusions. This program allows the use of data from standard databases.

Table 3. Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,763a	,582	,443	832,01381

a. Predictors: (Constant), Proizvodnja el. energije u GWh

b. Dependent Variable: Izvoz el energije u GWh

(Source: Author's calculation)

In the SPSS analysis in the Model Summary table, we found that the value of the Pearson coefficient is R=0.763, which indicates the existence of a strong linear relationship between the production of electrical energy and the export of electrical energy. The indicator R Square=0.582 represents the coefficient of determination (relative measure of the representativeness of the regression line) is used to determine the percentage participation of the explained variability in the total variability, which means that the independent variable electricity production explains 58.2% of the variability of the dependent variable, i.e. 58 .2% of the dispersion of electricity production.

To determine statistical significance, ie. testing the null hypothesis that the coefficient of determination in the analyzed relationship is equal to zero, i.e. that linearity does not exist between variables, ANOVA is used. ANOVA is used to determine whether there is a statistically significant difference between the means of multiple statistical sets.

Table 4 Analysis of variance (ANOVA)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2897023,855	1	2897023,855	4,185	,133a
	Residual	2076740,945	3	692246,982		
	Total	4973764,800	4			

(Source: Author's calculation)

The probability that the random variable F -distribution, which has d_f1=1 degree of freedom in the numerator and d_f2=3 degrees of freedom in the denominator, exceeds the value 4.185 is equal to the significance i(Sig.) or written P($F_{1;3}$) >4,185=0,133 (significance can also be described as a probability P-value). If the empirical level of significance and the theoretical level of significance (α) are compared, a decision can be made. Since in this case the empirical level of significance is higher than the theoretical level (0.133>0.05), the null hypothesis is not rejected. From this we conclude that there is no statistically significant difference in the average amount of electricity produced and the average amount of electricity exported, at the level of significance α=0.05

Table.5 Coefficients

Model		Unstandardized Coefficients		StandardizCoefficient	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-6865,948	6749,100		-1,017	,384	-28344,596	14612,700
	Proizvodnja el. energije u GWh	,805	,394	,763	2,046	,133	-,448	2,058

a. Dependent Variable: Izvoz el energije u GWh

(Source: Author's calculation)

The parameters of the simple linear regression equation can be read from table no. 5 so that the equation can be written as follows: y=-6865,948+0,805x

CONCLUSION

Statistics and its field of statistical modeling enables easier access and processing of statistical data. A statistical model is nothing more than a mathematical model that establishes a mathematical relationship between one or more random variables. When applying statistical models, the first step is data collection, followed by model selection. One of the popular statistical models, which depends on the number of parameters, the type of variables, the relationship between variables, is the regression model, a time series analysis model.

A regression model is a type of predictive statistical model that determines the existence of a relationship between characteristics, while the strength of these relationships is determined using correlation, i.e. the correlation coefficient as a relative measure. The method used to determine the strength of the relationship is determined by the type of features being analyzed. The strength of the relationship between the data measured by the rank scale and the nominal scale is determined by the correlation coefficient. In most cases, the chosen statistical model is flexible, new parameters can be included in it, and it adapts to the introduction of new data. The strength of the connection ie. the connection of certain characteristics is measured and used in a large number of scientific disciplines.

Correlation analysis is used today in marketing, medicine, economics, etc. Due to its simplicity, which is the main advantage compared to other statistical methods, it is widely used. The advantage of the model is reflected in the identification of the presence or absence of a connection between the variables and thus is more relevant for everyday use, then correlation analysis is in many cases the starting point for research, where the direction and strength of the connection is determined, and the findings can then be narrowed down in further research.

REFERENCES

1. Agencija za statistiku BiH
2. Anderson, D., Sweeney, D., Williams, T.(2011): Statistics for business and economics, 11- th Edition, Cenage Learning
3. Cassel C.M, Särnadal,C.E, & Wretman, J.H. (1977). Foundations of Inference in Survey Sampling. New York: Wiley
4. FIPA (Foreign Investment Promotion Agency); BiH
5. Kovačić, Z.J. (1995). Analiza vremenskih serija. Univerzitet u Beogradu, Ekonomski fakultet, Beograd
6. Lovrić, M., (2008) Osnovi statistike, Ekonomski fakultet, Kragujevac
7. Mills, T. C. (2019). Applied Time Series Analysis - A Practical Guide to Modeling and Forecasting. U.K.: Academic Press
8. Šošić I; Serdar V.(1994) Uvod u statistiku; Školska knjiga Zagreb