Journal of Natural Sciences and Mathematics of UT, Vol. 4, No. 7-8, 2019

UDC: 519.676
Conference paper

STUDY FOR TWO DIMENSIONAL DATA AND APPLICATION

Lazim Kamberi[1*], Shpresim Kameri[2]

[1] *Department of Mathematics, Faculty of Natural Sciences, University of Tetova, RNM*

[*] *Corresponding author e-mail: lazim.kamberi@unite.edu.mk*

Abstract

The purpose of examining two-dimensional data is to show a kind of relationship or connection between two variables quantitative and qualitative. This link can be made through statistical tables, graphs, and using a "cloud points or scattergram" to show what is the connection, or link, or dependence, that exists between the two variables if any exists. This connection can be identified as either strong or weak, if the points appear as collected or scattered, respectively. Also, we use regression analysis to digest the relationship between the two variables. Usually, one of the variables, the explanatory variable (independent) can be identified as an impact on the value of other variables, the variance variable (dependent). During the applications to show the dependence between the two variables we will also use the chi-square test, which also enables us to evaluate how closely a choice matches the expected distribution.

Keywords: *variables, linear dependence, regression, scattergram, hi-squares.*

1. Introduction

Two-dimensional data simply means data where each individual is characterized by two quantitative or qualitative variable variables, in this case, for each individual i, $(1 \le i \le N)$ two variables are measured X and Y and observations are couples (x_i, y_i) where x_i (respectively y_i) is the value found of X (respectively Y) of the individual i. But what if we notice that two variables seem to be related? We may note that the values of two variables, such as, for example, the result of the average of the test scores of three state Matura courses of a student who is used to admission to the university and the average grade of secondary school, behave in the same way and that students who have a high score of the oral test scores also tend to have a high average score (see Table 1). In this case, we would like to study the nature of the connection between the two variables.

Table 1: A table of variable scores of the average scores of the three state maturity test scores of a student who is used to admission to the university and the average grade of secondary school grades

Student	Average of the state mature	Average grade in the school
1	2.67	2.93
2	3.33	4.65
3	3.67	3.20
4	3.00	2.21
5	3.67	3.13
6	3.67	3.18
7	3.67	3.46

These types of studies are quite common and we can use the concept of correlation and regression analysis to describe the dependence between two constant variables. The purpose of this paper is to show a kind of relation of dependence between the two variables and to evaluate by what power a category might affect the other.

2. Scatterplots (or cloud points)

The purpose of examining two-variable data is usually to show a kind of connection between two variables. Just reviewing the data table is not enough to detect this relationship. Therefore, a graph can be constructed, placing one of the variables on the axis of the abyss and the other variable in the axis of the order.

When the variables are quantitative, each preview (x_i, y_i) is a couple of real numbers and is therefore a point in the selected system of the above axes.

By presenting each pair of observations with the respective point, a point pool is obtained in the plane. This graphic representation is called "new points" (scattergram).

The two-variable data can be represented using a "point pointer" to show what the relation between two variables (if any) is.

If there is a relation between two variables, it can be identified as strong if the items appear more accidentally assembled (Figure 1). It can be identified as weak if the points appear more randomly distributed (Figure 2).

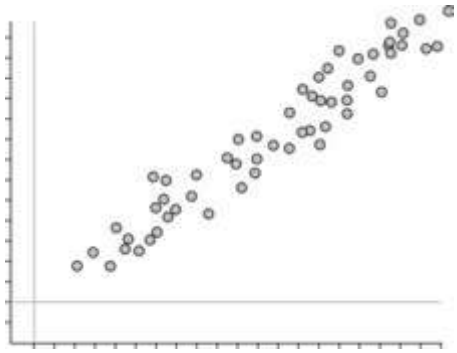

Figure 1. Point accidentally assembled

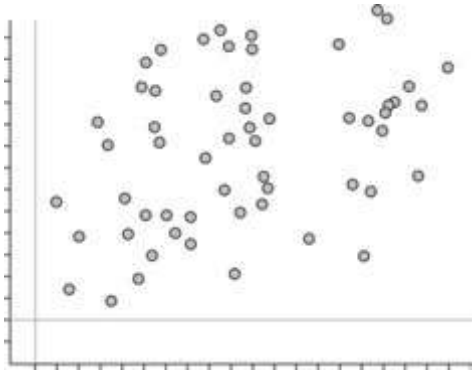

Figure 2. Points randomly distributed

The easiest way to describe a dependency on two-dimensional data is to draw an ellipse as a point cloud (Figure 3).

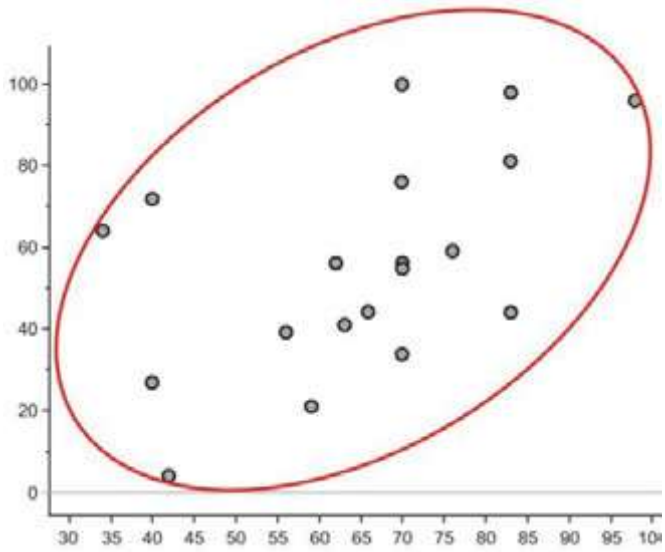

Figure 3. An ellipse as a point cloud

The Ellipse also gives us some information about the linear relation strength. If there were a strong linear data relation, the new data cloud would be much longer than it will be wider. Long and narrow ellipses imply a strong linear bond, while short and broad ellipses show a weaker linear relationship.

3. Correlation between values and the use of scatterplots

Correlation also measures the ratio between two-dimensional data. Two-dimensional data is a set of data in which each subject (individual) has two related observations. If we carefully examine the data in Table 1, we note that those students with high average grades of state Matura used to admit to university tend to have high average grades of middle school and those with lower average grades of state graduates tend to have average lower grades of secondary school. In this case, there is a tendency for students to score similar results in both variables and the performance between the variables seems to be dependent.

3.1 Correlation Patterns in Scatterplot Graphs

Examining a dot plot chart allows us to get an idea about the connection between two variables when the points on the score bar graph produce a "**lower left-to-right**" model (figure 4). Then we say there is a positive correlation between the two variables. This model means that when the result of a survey is high, then we expect the result of other observations to be high and vice versa.

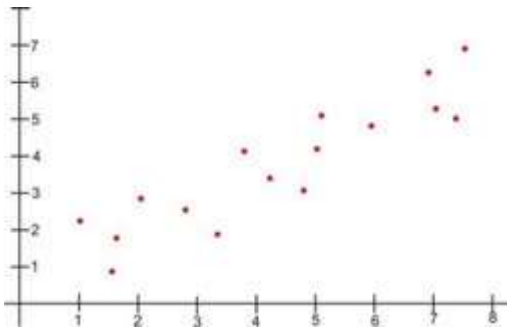

Figure 4. Lower left-to-right" model

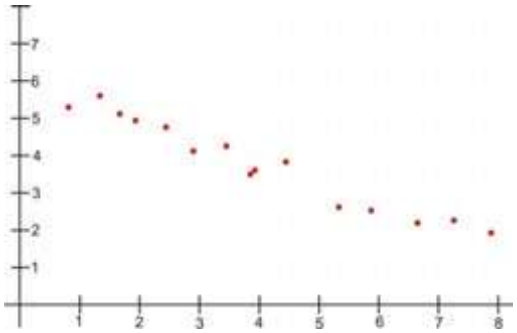

Figure 5. Upper left-bottom-right" pattern

When points on a dot plot chart produce an "**upper left-bottom-right**" pattern (Figure 5), then we say there is a negative correlation between the two variables. This model means that when the result of a survey is high, then we expect the result of other surveys to be low and conversely.

When all the points in a dot point lie in a straight line, then we have that what is called the perfect (true) correlation between the two variables (Figure 6).

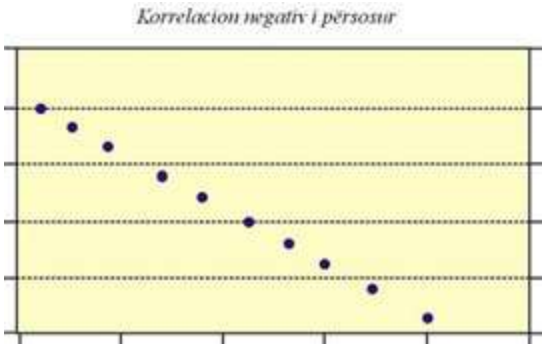

Figure 6. Correlation between the two variables

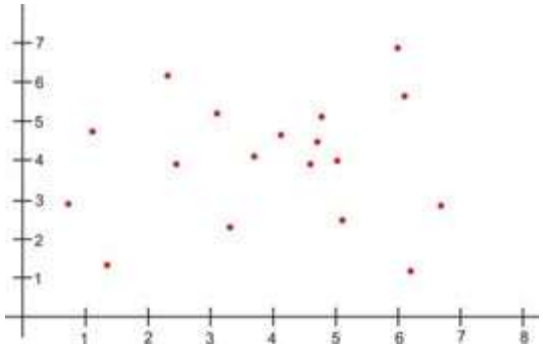

Figure 7. Close-zero correlation

A Scatter Plots in which the points do not have a linear (positive or negative) direction is called zero correlation or close-zero correlation (Figure 7).

When we are talking about point cloud we want to see not just the direction of dependence positive or negative=, but we also want to see the magnitude of the dependence. If we pull an ellipse around all the points in the Scatter Plots, then we would be able to see the extent or the magnitude of the dependence between the variables.

If the points are close to each other and the width of the ellipse is small, this means that there is a strong correlation between the variables (Figure 8).

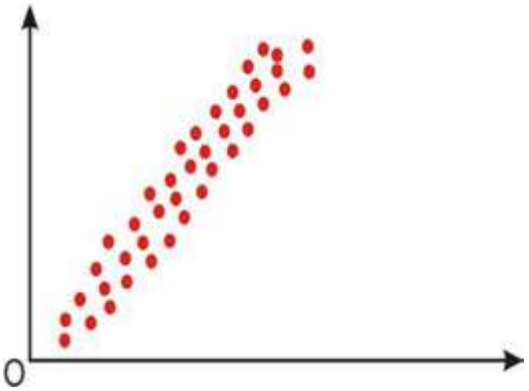

Figure 8. Correlation between the variables

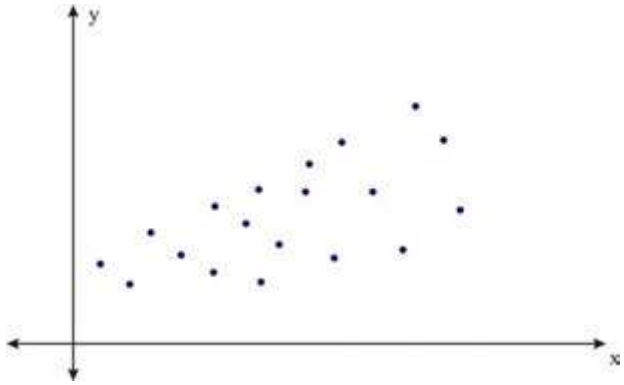

Figure 9. correlation between the variables

However, if the points are far apart and the ellipse is very wide, this means that there is a weak correlation between the variables (figure 9).

3.2 Correlation coefficients

Correlation measures the direction and magnitude of linear dependence between two-dimensional data. When we look at the graphic of Scatter Plots, we can determine if the correlation is positive, negative, perfect, or zero. A correlation is strong when the points on the dot-point graphic are close at the same time. The correlation coefficient is an accurate measure of the dependence between the two variables. This indicator can take values between -1 and 1 as well as including -1 and 1.

The closer to table1 is the absolute value of the correlation coefficient, the stronger is the dependence between the variables. For example, a correlation coefficient of 0.20 indicates that there is a weak linear dependence between variables, and a coefficient of 0.90 indicates that there is a strong linear dependence between variables. To calculate the correlation coefficient, we use mostly the formula:

$$r_{XY} = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right]}\sqrt{\left[n\sum y^2 - \left(\sum y\right)^2\right]}} \qquad (3.2.1)$$

Let's use the above example from the entry to demonstrate how to calculate the correlation coefficient using the score value formula.

Example: What is the Pearson correlation coefficient for the two variables shown in Table 1?

To calculate the correlation coefficient, we need to calculate some pieces of information, including xy, x^2 and y^2 . Therefore, the values of xy, x^2 and y^2 are given in Table 2.

Table 2: The values of xy, x^2 and y^2

Students	**avarage (X) of statematura**	**average (Y) of school**	xy	x^2	y^2
1	2.67	2.63	7.02	7.13	6.97
2	3.33	4.65	15.48	11.09	21.62
3	3.67	3.20	11.74	13.45	10.24
4	3.00	2.21	6.63	9.00	4.89
5	3.67	3.13	11.49	13.45	9.70
6	3.67	3.18	11.67	13.45	10.11
7	3.67	3.46	12.70	13.45	11.98
Sum	23.68	22.76	76.74	81.02	75.51

Applying the formula (3.2.1) for those data, we find the calculated value:

$$r_{XY} = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right]}\sqrt{\left[n\sum y^2 - \left(\sum y\right)^2\right]}}$$

$$= \frac{7(76.73) - (23.68)(22.76)}{\sqrt{\left[7(81.02) - (23.68)^2\right]}\sqrt{\left[7(75.51) - (22.76)^2\right]}} = \frac{\text{-0.85}}{8.22} \approx -0.10 \qquad (3.2.2)$$

The correlation coefficient not only provides a measurement of the dependence between the variables but gives us an idea about how the dependence of a variable can generally relate to the dependence of the other. For example, the -0.10 correlation coefficient we calculated above shows a poor dependence on variables X and Y, since its absolute value is away from 1. That means that you have a cloud points where points do not have a linear (positive or negative) direction.

4. Conclusions

- If there is a relation between the two variables, it can be represented using a "scatterplots".
- Studying the graphic of scatterplots allows us to get an idea about the connection between the two variables.
- The connection between the two variables can be identified as "strong" if the points appear more accidentally accumulated and can be identified as "weak" if the points appear more randomly distributed.
- The easiest way to describe dependency is to draw an ellipse as a cloud of points: Long and narrow ellipses imply a strong linear bond, while short and broad ellipses show a weaker linear relationship.
- The linear dependence between the two-dimensional data is also measured by the correlation coefficient that is an accurate measurement of the dependence between the two variables. For example, a correlation coefficient of 0.20 indicates that there is a

weak linear dependence between variables, and a coefficient of 0.90 indicates that there is a strong linear dependence on variables.

- For example, we have: $|r_{XY}| = 0.1$ that shows a poor dependence on variables X and Y. This means that you have scatterplots in which the points do not have a linear (positive or negative) direction.

References

[1]. Lluka Puka, Hyrje në statistikën e zbatuar, Mediaprint, 2015, Tirana, Albania.

[2]. Ross Sh. (2012), A First Course in Probability. Upper Saddle River: Prentice Hall.

[3]. Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. Proceedings of the Royal Society, London, Ser. A, 160, 268-282.

[4]. Feller, W. (1968). An Introduction to Probability Theory and Its Applications, Volume I. New York: Wiley.

[5]. Feller, W. (1971). An Introduction to Probability Theory and Its Applications, Volume II. New York: Wiley.