Chapter: Correlation -6

Md. Abdul Alim
2-11-2010

Bivariate data: Bivariate data is data for which there are two variables for each observation. As an example, the following bivariate data show the ages of husbands and wives of 10 married couples.

Husband	36	72	37	36	51	50	47	50	37	41
Wife	35	67	33	35	50	46	47	42	36	41

Correlation: Correlation is a statistical technique which measure and analyses the degree or extent to which two or more variables fluctuate with reference to one another.

Correlation thus denotes the interdependence amongst variates. The degrees are expressed by a coefficient which ranges between -1 and1. The direction of change is indicated by + or - signs.

If the **increase (decrease)** in one variable results in the corresponding **increase (decrease)** in the others i.e. if the changes are in the same directions the variables are **positively correlated**. For example, the heights and weights of a group of persons are positively correlated, advertising and sales.

If the **increase (decrease)** in one variable results in the corresponding **decrease (increase)** in the others i.e. if the changes are in the opposite directions the variables are **negatively correlated**. For example, T.V registration and cinema attendance is negatively correlated.

An absence of correlation is indicated by zero.

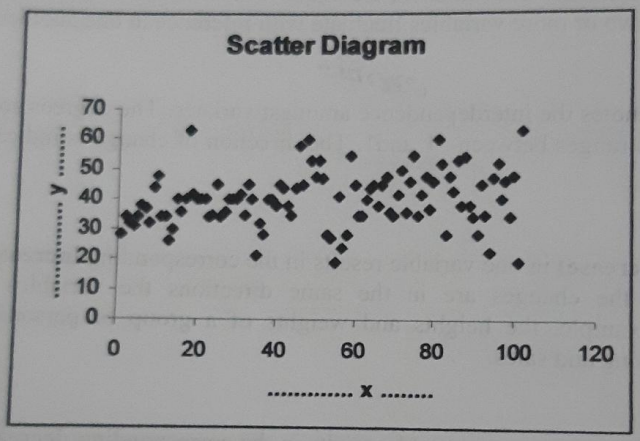
Correlation thus expresses the relationship through a relative measure of change and it has nothing to do with the units in which the variables are expressed.

Uses

- Economic theory and business studies relationships between variables like price and quantity demanded, advertising, expenditure scales promotion measure etc. The correlation analysis helps in deriving precisely the degree and direction of such relationships.
- The concepts of regression are also based upon the measure of correlation.

Scatter Diagram: Scatter diagram (or Dotogram or Scattergram) is a simple and attractive method of diagrammatic represent of bivariate distribution for ascertaining the nature of correlation between the variables. Thus for the bivariate distribution (x_i, y_i), $i = 1, 2, ..., n$ if the values of the variables X and Y be plotted along the X-axis and Y-axis respectively in the XY plane, the diagram of dots so obtained is known as scatter diagram.

On the other hand, a scatter plot of two variables shows the values of one variable on the Y-axis and the values of the other variable on the X-axis. Scatter plots are well suited for revealing the relationship between two variables.



Types of Correlation: Correlation is described or classified in several different ways. Three of the most important are:

- Positive and negative Correlation
- Simple, partial and multiple Correlation
- Linear and non-linear Correlation

Positive and negative correlation: If two variables changes in the same direction (i.e. if one increases the other also increase or if one decreases the other also decreases) then this is called a positive correlation. For example:

Positive Correlation	
X	Y
10	15
12	20
14	22
18	25
20	37

Positive Correlation	
X	Y
80	50
70	45
60	30
40	20
30	10

If two variables change in the opposite direction (i.e. if one increases, the other decreases and vice versa), then the correlation is called a negative correlation. For example: T.V registrations and cinema attendance.

Negative Correlation	
X	Y
20	40
30	30
40	22
60	15
80	12

Negative Correlation	
X	Y
100	10
90	20
60	30
40	40
30	50

2. Simple, Partial and Multiple Correlation

- When only two variables are studied it is a problem of simple correlation.
- When three or more variables are studied it is a problem of either multiple or partial correlation.

In multiple correlation three or more variables are studied simultaneously. For example, when we study the relationship between the yield of rice per acre and both the amount of rainfall and the amount of fertilizers used, it is problem of multiple correlation. Similarly the relationship of plastic hardness, temperature and pressure is multivariate.

In partial correlation we recognize more than two variables. But consider only two variables to be influencing variable being kept constant. For example, in the rice problem taken above if we limit our correlation analysis of yield and rainfall to periods when a certain average daily temperature existed, it becomes a problem of partial correlation.

3. Linear and non-linear correlation: The nature of the graph gives us the idea of the linear type of correlation between two variables. If the graph is in a straight line, the correlation is called a "linear correlation" and if the graph is not in a straight line, the correlation is non-linear and curve-linear.

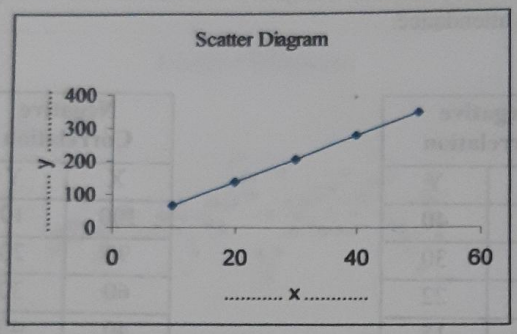
[illegible] monojit_math09@yahoo.com

Md. Moyazzem Hossain
Lecture, Dept. of Statistics
Islamic University, Kushtia.

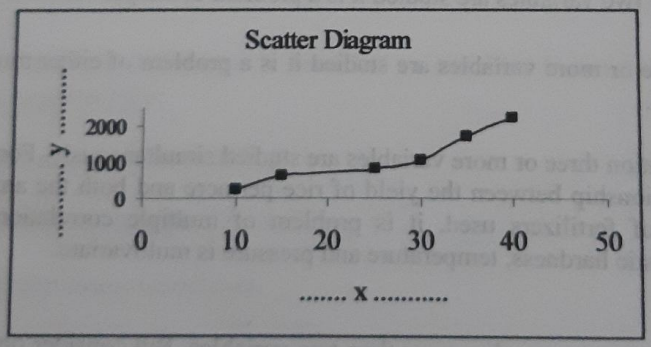
The distinction between linear and non-linear correlation is based upon the constancy of the ratio of change between the variables. If the amount of change in one variable tends to bear a constant ratio to the amount of change in the other variable then the correlation is said to be linear. For example, observe the following two variables X and Y:

X:	10	20	30	40	50
Y:	70	140	210	280	350

It is clear that the ratio of change between the two variables is the same. If such variables are plotted on a graph paper all the plotted points would fall on a straight line.



Correlation would be called non-linear or curvilinear if the amount of change in one variable doesn't bear a constant ratio to the amount of change in the other variable. For example, if we double the amount of rainfall, the production of rice or wheat etc. would not necessarily be doubled.



Properties of the Coefficient of Correlation: The following are the important properties of the coefficient of correlation, r:

- The coefficient of correlation lies between -1 and +1, $-1 \le r \le +1$.

মুহাম্মদ মোয়াজ্জেম হোসেন, লেকচারার, পরিসংখ্যান বিভাগ, ইবি। সহযোগিতায় মোছাঃ মনোজিত কুমার মন্ডল, গণিত বিভাগ, ২০০৯-১০(ইবি)।
মোবাইল ০১৯১৩৫৫৪০২৫, ০১৯৮৬২০৫১২ ইমেইল : monojit_math09@yahoo.com

Md. Moyazzem Hossain
Lecture, Dept. of Statistics

- The coefficient of correlation is the geometric mean of the two regression coefficients. Symbolically: $r = \sqrt{b_{xy} \times b_{yx}}$
- If X and Y are independent variables then coefficient of correlation is zero. However, the converse is not true.

Degrees of Correlation: Through the coefficient of correlation, we can measure the degree or extent of the correlation between two variables. On the basis of the coefficient of correlation we can also determine whether the correlation is positive or negative and also its degree or extent.

⇨**Perfect correlation:** If two variables changes in the same direction and in the same proportion, the correlation between the two is **perfect positive**. According to Karl Pearson the coefficient of correlation in this case is +1. On the other hand, if the variables change in the opposite direction and in the same proportion, the correlation is **perfect negative**. Its coefficient of correlation is -1. In practice we rarely come across these types of correlations.

⇨**Absence of correlation:** If two series of two variables exhibit no relations between them or change in variable does not lead to a change in the other variable, then we can firmly say that there is **no correlation** or **absurd correlation** between the two variables. In such a case the coefficient of correlation is 0.

⇨**Limited degrees of correlation:** If two variables are not perfectly correlated or is there a perfect absence of correlation, then we term the correlation as Limited correlation. It may be positive, negative or zero but lies with the limits ±1.

High degree, moderate degree or low degrees are the three categories of this kind of correlation. The following table reveals the effect (or degree) of coefficient or correlation.

Degrees	Positive	Negative
Absence of correlation →	Zero	0
Perfect correlation →	+ 1	-1
High degree →	+ 0.75 to + 1	- 0.75 to -1
Moderate degree →	+ 0.25 to + 0.75	- 0.25 to - 0.75
Low degree →	0 to 0.25	0 to - 0.25

মুহাম্মদ : মনোজিৎ কুমার সাহা,পরিসংখ্যান বিভাগ, ইবি। সহযোগিতায় সহযোগিতায় মুহাম্মদ মঈয়ুজ হোসেন, প্রভাষক, পরিসংখ্যান বিভাগ,ইবি।
মোবাইল ০১৯১৩৫৪০৮২৩,০১৯৮৩৯২০৫১২ ইমেইল : monojit_math09@yahoo.com

Md. Moyazzem Hossain
Lecture, Dept. of Statistics
Islamic University, Kushtia.

Methods of Determining Correlation: We shall consider the following most commonly used methods.

- Scatter Plot.
- Karl Pearson's coefficient of correlation.
- Spearman's Rank-correlation coefficient.
- Method of Least Squares.

Karl Pearson's Coefficient of Correlation: Of the several mathematical methods of measuring correlation, the Karl Pearson's method, popularly known as Pearsonian coefficient of correlation, is most widely used in practice. The coefficient of correlation is denoted by the symbol r. If the two variables under study are X and Y, the following formula suggested by Karl Pearson can be used for measuring the degree of relationship.

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left[\left\{\sum X^2 - \frac{(\sum X)^2}{N}\right\}\left\{\sum Y^2 - \frac{(\sum Y)^2}{N}\right\}\right]}}$$

The value of the coefficient of correlation as obtained by the above formula shall always lie between ± 1.

When $r = +1$, it means there is perfect positive correlation between the variables.

When $r = -1$, it means there is a perfect negative correlation between the variables.

When $r = 0$, it means there is no relationship between the variables.

Example 1: Calculate the coefficient of correlation between the heights of father and his son for the following data.

Height of father (cm):	165	166	167	168	167	169	170	172
Height of son (cm):	167	168	165	172	168	172	169	171

Solution: We know that. Correlation of coefficient

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left\{\sum X^2 - \frac{(\sum X)^2}{N}\right\}\left\{\sum Y^2 - \frac{(\sum Y)^2}{N}\right\}}}$$

Let us consider the height of father is X and height of son is Y.

By using calculator we get,

$\sum X^2 = 225828$ $\sum X = 1344$ $N = 8$

$\sum Y^2 = 228532$ $\sum Y = 1352$ $\sum XY = 227160$

$$r = \frac{227160 - \frac{1344 \times 1352}{8}}{\sqrt{\left\{225828 - \frac{(1344)^2}{8}\right\}\left\{228532 - \frac{(1352)^2}{8}\right\}}}$$

$= 0.603022689 = 0.603$

So the relation between the height of father and height of son is moderate degree positive.

Example 2: The following data consist of observations for the weights of 10 different automobiles (in 1000 pounds) and the corresponding fuel consumptions (gallons per 100 miles).

Weight (x)	Fuel Consumption (y)
3.4	5.5
3.8	5.9
4.1	6.5
2.2	3.3
2.6	3.6
2.9	4.6
2.0	2.9
2.7	3.6
1.9	3.1
3.4	4.9

We would like to find out how y is correlated to x.

Solution: We know that. Correlation of coefficient

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left\{\sum X^2 - \frac{(\sum X)^2}{N}\right\}\left\{\sum Y^2 - \frac{(\sum Y)^2}{N}\right\}}}$$

By using calculator we get,

$\sum x^2 = 89.29$ $\sum X = 29$ $N = 8$

$\sum Y^2 = 207.31$ $\sum Y = 43.9$ $\sum XY = 135.8$

Md. Mayazzem Hossain
Lecture, Dept. of Statistics
Islamic University, Kushtia.

$$r = \frac{135.8 - \frac{29 \times 43.9}{10}}{\sqrt{\left\{89.29 - \frac{(29)^2}{10}\right\}\left\{207.31 - \frac{(43.9)^2}{10}\right\}}} = 0.976629971 = 0.976$$

So the relation between weight and Fuel Consumption is high degree positive.

(3) Spearman's Rank Correlation: The association between two series of rank is called rank correlation. The method of ascertaining the coefficient of correlation by ranks was devised by Charles Edwards Spearman in 1904.This method is especially useful in case when the actual magnitudes or item values are not given and simply their ranks in the series are known. Spearman's rank correlation coefficient, usually denoted by ρ (Rho) is given by the formula:

$$\rho = 1 - \frac{6\sum d_i^2}{n^3 - n} = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

where d stands for the difference between the pair of ranks and n the number of paired observations.

The value of Spearman's rank correlation coefficient ranges between -1 and +1.When ρ is +1, the concordance between rankings is perfect and the ranks are in the same direction. When ρ is -1, there is also perfect concordance between rankings but the ranks in opposite direction.

In rank correlation we may have two types of problems:

A. Where actual ranks are given.

B. Where ranks are not given.

A. Where Actual Ranks are given: Where Actual Ranks are given the steps required for computing rank correlation are:

- Take the difference of the two ranks i.e $(R_1 - R_2)$ and denote these differences by d.
- Square these differences and obtain the total $\sum d_i^2$
- Apply the formula

$$\rho = 1 - \frac{6\sum d_i^2}{n^3 - n}$$

Example 1: Two managers are asked to rank a group of employees in order of potential for eventually becoming top managers .The rankings are as follows:

[illegible] : [illegible], [illegible], [illegible] : [illegible], [illegible], [illegible] : monojit_math09@yahoo.com

Md. Moyazzem Hossain
Lecture, Dept. of Statistics
Islamic University, Kushtia

Employee	Ranked by manager I	Ranked by Manager II
A	10	9
B	2	4
C	1	2
D	4	3
E	3	1
F	6	5
G	5	6
H	8	8
I	7	7
J	9	10

Compute the coefficient of rank correlation and comment on the value.

Solution: Calculation of Rank Correlation Coefficient

Employee	Ranked by manager I (R_1)	Ranked by Manager II (R_2)	$d^2=(R_1-R_2)^2$
A	10	9	
B	2	4	By using Calculator
C	1	2	
D	4	3	
E	3	1	
F	6	5	
G	5	6	
H	8	8	
I	7	7	
J	9	10	
Total			$\sum d_i^2=14$

We know that, $\rho=1-\frac{6\sum d_i^2}{n^3-n}=1-\frac{6\times 14}{10^3-10}=0.915$

Thus we find that there is a high degree of positive correlation in the ranks assigned by the two managers.

B. Where Ranks are not given: When we are given the actual data and not the ranks it will be necessary to assigns the ranks .Ranks can be assigned by taking either the highest value as 1 or the lowest value as 1. But whether we start with the lowest value or the highest value we must follow the same method in case of all the variables.

Md. Moyazzem Hossain
Lecture, Dept. of Statistics
Islamic University, Kushtia.

Example1: Calculate the rank correlation coefficient for the following data of marks of 2 tests given to candidates for a clerical job:

Preliminary test	92	89	87	86	83	77	71	63	53	50
Final test	86	83	91	77	68	85	52	82	37	57

Solutions: Calculation of Rank Correlation Coefficient

Preliminary test	R_1	Final test	R_2	$d^2 = (R_1 - R_2)^2$
92	10	86	9	
89	9	83	7	
87	8	91	10	
86	7	77	5	By using
83	6	68	4	Calculator
77	5	85	8	
71	4	52	2	
63	3	82	6	
53	2	37	1	
50	1	57	3	
Total				$\sum d_i^2 = 44$

We know that, $\rho = 1 - \frac{6\sum d_i^2}{n^3 - n} = 1 - \frac{6 \times 44}{10^3 - 10} = 1 - 0.267 = 0.733$

Thus there is a high degree of positive correlation between preliminary and final test.

Merits

- This method is simpler to understand and easier to apply compared to the Karl Pearson's method.
- Where the data are of a qualitative nature like honesty, efficiency, intelligence etc., this method can be used with great advantage. For example the workers of two factories can be ranked in order of efficiency and the degree of correlation established by applying the method.
- This is the only method that can be used where we are given the ranks and not the actual data.
- Even where actual data are given rank method can be applied for ascertaining rough degree of correlation.

Limitations:

- This method cannot be used for finding out correlation in a grouped frequency distribution.

[illegible] : [illegible], [illegible] : monojit_math09@yahoo.com

Md. Moyazzem Hussain
Lecture, Dept. of Statistics
Islamic University, Kushtia.

- Where the number of observations exceed 30 the calculations becomes quite tedious and require a lot of time. Therefore this method should not applied where n exceeding 30 unless we are given the ranks and not the actual values of the variable.

(4) Method of Least Squares

For finding out correlation by the coefficient method of least squares we have to calculate the values of two regression coefficients that of x on y and y on x. The correlation coefficient is the square root of the product of two regression coefficients. Symbolically,

$$r = \sqrt{b_{xy} \times b_{yx}}$$

Coefficient of Determination: One very convenient and useful way of interpreting the value of coefficient of correlation between two variables is to use the square of coefficient of correlation, which is called coefficient of determination. The coefficient of determination thus equals r^2.

*** If the value of $r = 0.9$, r^2 will be 0.81 and this would mean that 81% of the variation in the dependent variable has been explained by the independent variable.