1. Introduction

Welcome to the series of E-learning modules on scatter diagram. In this module, we will learn about the scatter diagram, correlation, its utility and types, construction of the scatter diagram, merits and demerits of the scatter diagram.

By the end of this session, you will be able to:

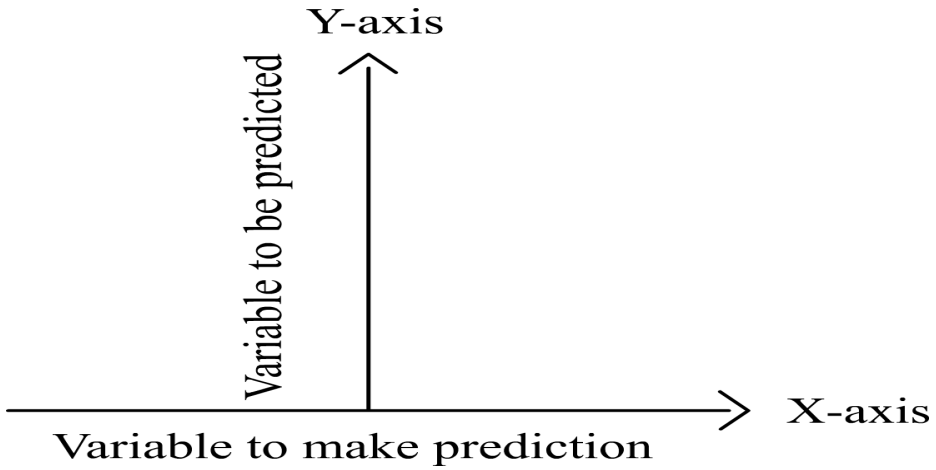
- Explain scatter diagram
- Explain correlation
- Describe the utility of correlation
- Describe the types of correlation
- Describe the linear relationship between the variables
- Explain the construction of scatter diagram
- List the merits and demerits of scatter diagram

In earlier modules, we have studied about the bivariate data and fitting different curves to the given bivariate data using the method of least squares. The type of the curve to be fitted is decided by plotting the data and identifying the behavior of the curve.
Here, we consider one variable as independent and the other as the dependent variable. We presume that there is a relation between the two variables. The type of relationship can be studied using scatter diagram.

Scatter diagram gives an idea about the relationship between paired data and can provide information that is more useful. It is a graphical technique used to analyze the relationship between two variables.

Two sets of data are plotted on a graph with the Y-axis being used for the variable to be predicted and X-axis being used for the variable to make prediction.

Figure 1



A Scatter Diagram shows a correlation between two items for the following three reasons:

1. There is a cause and effect relationship between the two measured items, where one is the cause and the other is an effect.
2. The two measured items are caused by a third item. For example, a Scatter Diagram that shows a correlation between cracks and transparency of glass utensils, where changes in both are caused by changes in furnace temperature.
3. Complete coincidence. It is possible to find high correlation of unrelated items, such as the number of ants crossing a path and newspaper sales.

Scatter Diagrams may thus be used to give evidence for a cause and effect relationship.

2. Degree of Scatter Affected by Several Factors

Wherever there is a cause-effect relationship, the degree of scatter in the diagram may be affected by several factors:

1. The proximity of the cause and effect: There is better chance of a high correlation if the cause is directly connected to the effect than if it is at the end of a chain of causes. Thus, a root cause may not have a clear relationship with the end effect.
2. Multiple causes of the effect: When measuring one cause, other causes are making the effect vary in an unrelated way. Other causes may also be having a greater effect, swamping the actual effect of the cause in the question.
3. Natural variation in the system: The effect may not react in the same way each time, even to a close major cause.

Consider the diagram that gives the multiple causes for a particular phenomenon.
There is not even one clear degree of correlation, above which a clear relationship can be said to exist. Instead, as the degree of correlation increases, the probability of that relationship also increases.
If there is sufficient correlation, then the shape of the scatter diagram will indicate the type of correlation.

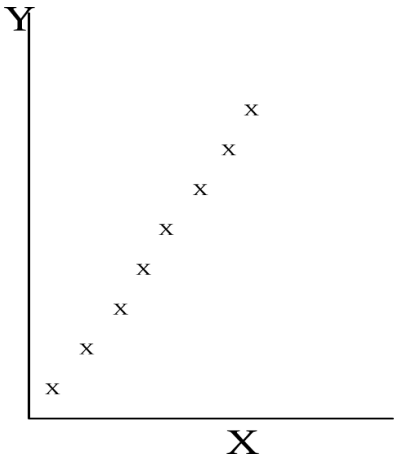
For example, there is a relationship between an ingredient and hardness of the product; between the cutting speed of a blade and the variations observed in length of parts; or the relationship between the illumination levels on the production floor and the mistakes made in quality inspection of product produced.

To illustrate this relationship, below are the few examples of scatter diagrams indicating the relationships between paired data. We will discuss how to interpret these charts, and then we will learn how to make one with paper and pencil.
In the following examples, you can see that the dots, which are actually data points, have various relationships.

In these diagrams, you would be able to determine that there is a relationship between the two variables.

Figure 2

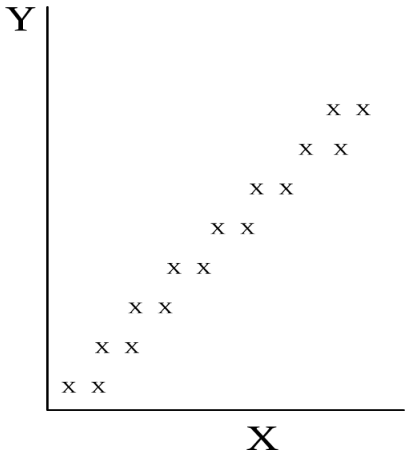
Perfect Correlation
(1)



The first figure shows that there is perfect relation between two variables. As X increases, Y also increases or as X decreases, Y also decreases. Therefore, there is perfect positive relation between the two variables.

Figure 3

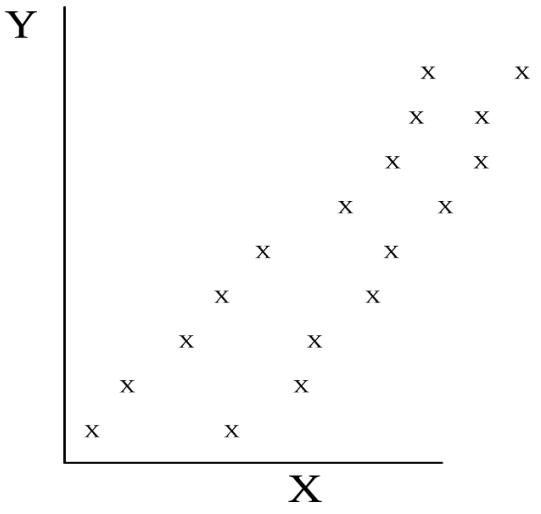
High degree positive
Correlation
(2)



In the second figure, the relation is not like the first one. However, there is high positive relation between two variables as the points are scattered around the straight line.

Figure 4

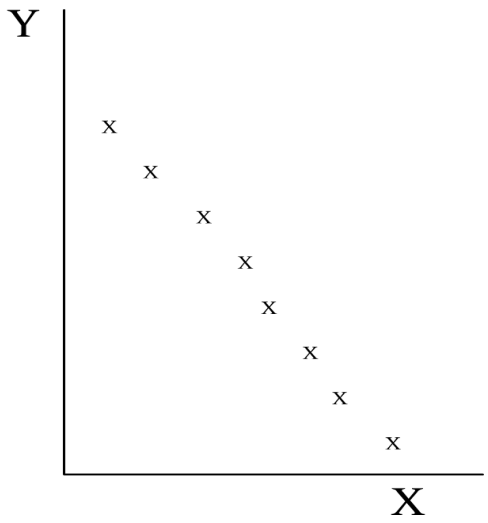
Low degree positive
correlation
(3)



In the third figure, observe that the points are scattered widely around the straight line. Hence, there is low positive correlation between the variables.

Figure 5

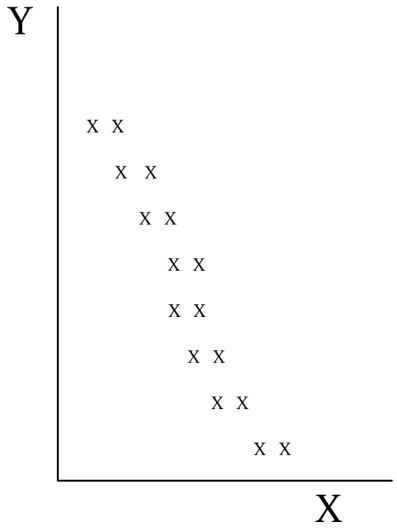
Perfect Negative Correlation (1)



In the above figures, the first shows there is perfect linear relationship between the two variables. Observe that in the figure, as X increases, Y decreases or as Y increases, X decreases. It shows perfect negative relationship between two variables.

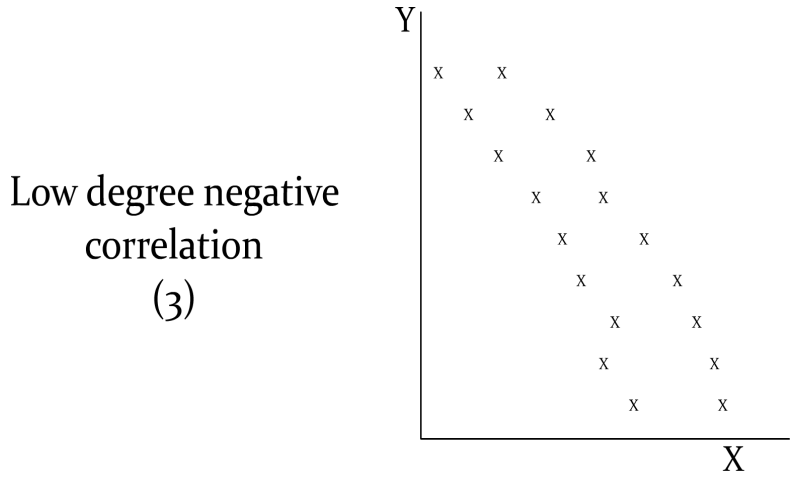
Figure 6

High degree negative Correlation (2)



In the second figure, the points are clustered around the straight line. Hence, there is high negative relationship between the two variables.

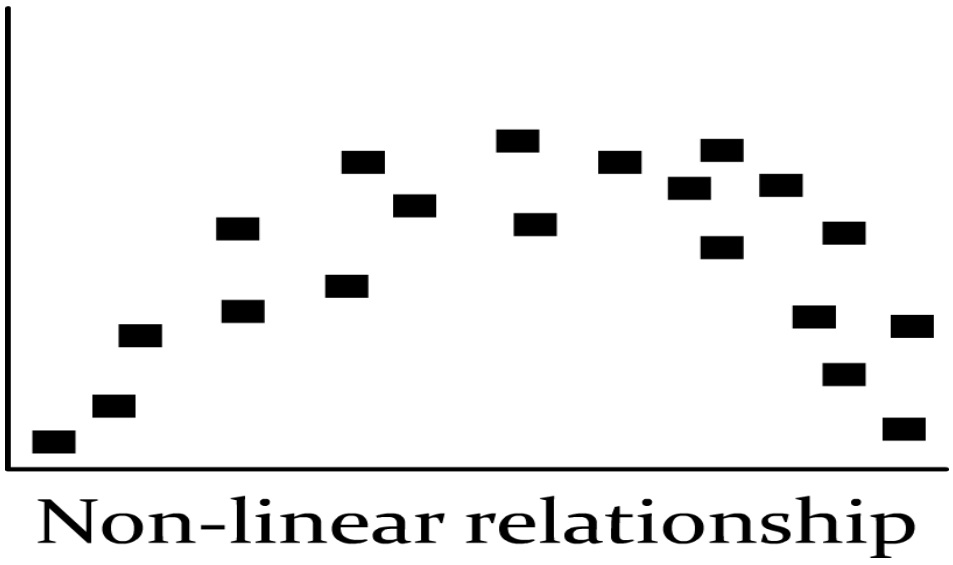
Figure 7



In the third figure, observe that the points are spread widely around the straight line. Hence, there is low negative relationship between the two variables.

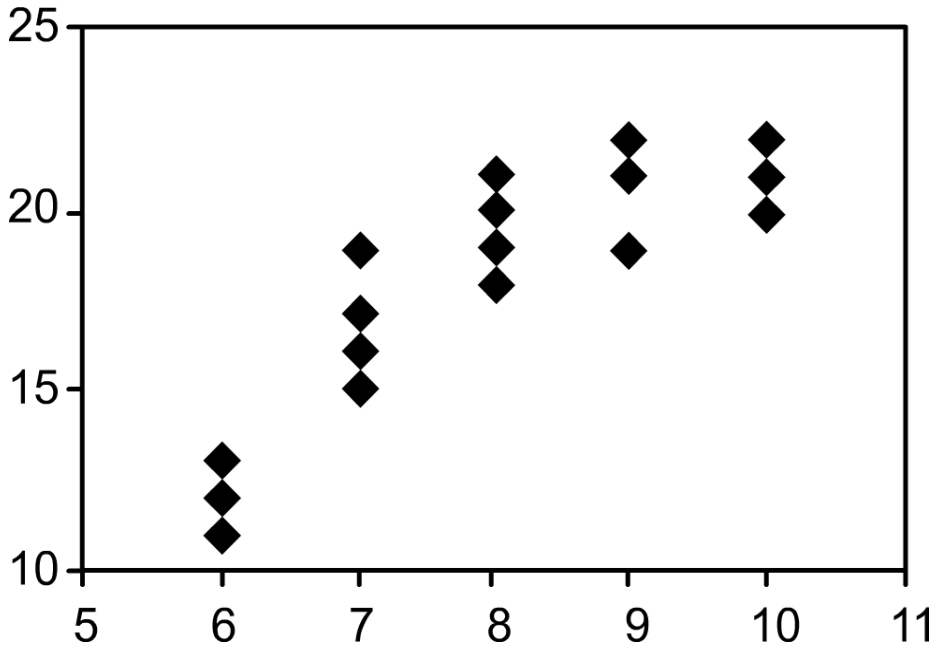
These diagrams show a non-linear relationship between the two variables.

Figure 8



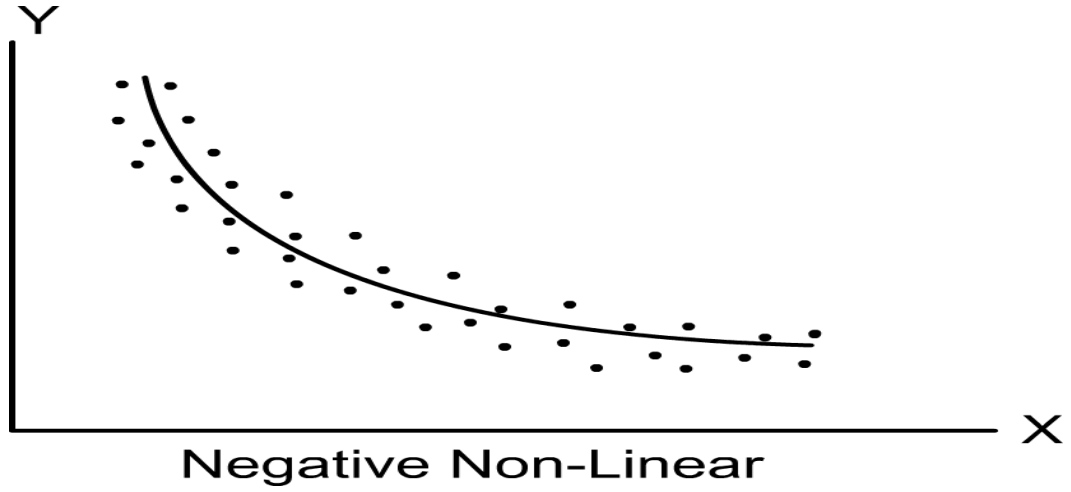
The first figure shows that there is a non-linear relationship between the two variables.

Figure 9



The second figure shows that there is a positive non-linear relationship between the two variables.

Figure 10



The third figure shows that there is a negative non-linear relationship between the two variables.

In this diagram, you would be able to determine that there is absolutely no relationship between the two items. You need to review the "Cause-and-Effect" Diagram or "brain-storming" session to find another item so that your primary item measured might have a relationship too.

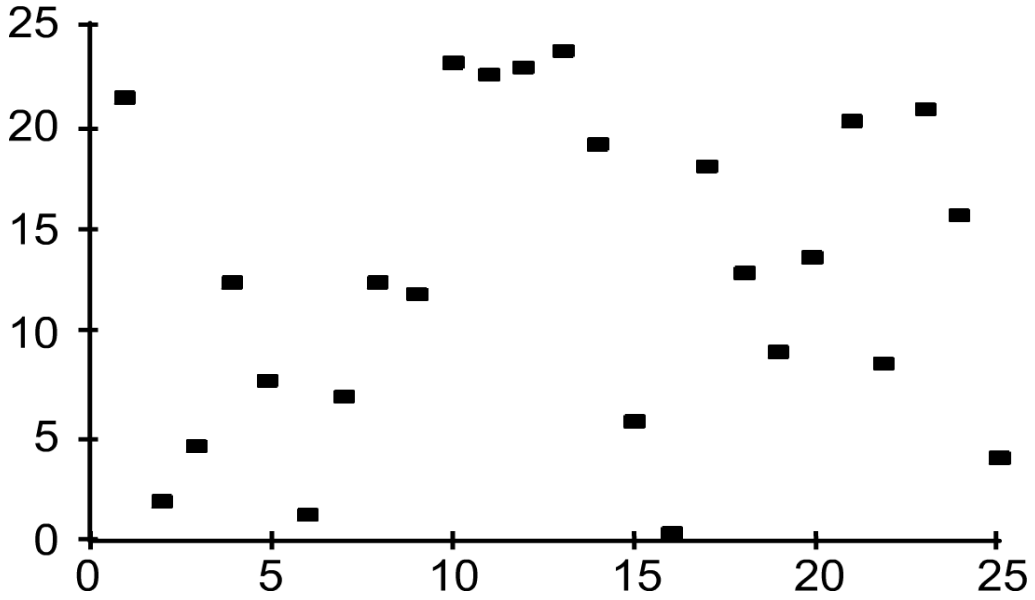
Points, which appear well outside a visible trend region, may be due to special causes of variation, and should be investigated.

In addition to visual interpretation, several calculations may be made around the scatter diagrams.

3. Meaning and Different Types of Correlation

Now, let us see the meaning of correlation and different types of correlation. The term correlation or co-variation indicates the relationship between two variables, where change in the value of one variable leads to the change in the value of another variable.

Figure 11



Let us state some important definitions of correlation.
L. R. Connor defines correlation as:
“If two or more quantities vary in sympathy, so that movements in the one tend to be accompanied by corresponding movements in the other, then they are said to be correlated.”

Croxton and Cowden define correlation as:
“When the relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation.”

A. M. Tuttle defines correlation as:
“Correlation is an analysis of the co-variation between two or more variables.”

According to Tippet, “The effect of correlation is to reduce the range of uncertainty of one’s prediction.”
Finally, according Ya Lun Chou,
“Correlation analysis attempts to determine the degree of relationship between the variables.”

The utility of study of correlation is immense both in physical as well as in social sciences.

- The study of correlation reduces the range of uncertainty associated with decision-

making. Particularly in business world, forecasting is an important phenomenon, and correlation studies help us to make relatively more dependable forecasts.

- Correlation analysis is very helpful in understanding economic behavior. This is helpful in studying factors by which economic events are affected.

- Correlation studies help us to estimate the likely change in a variable with a particular amount of change in related variable. For example, correlation study can help in finding out the change in demand with certain amount of change in price.

- Inter-relationship studies between different variables are very helpful tools in promoting research and opening new frontiers of knowledge.

Let us discuss the different types of correlation.

Correlation can be positive or negative. When values of the variables move in the same direction, that is, when an increase in the value of one variable is associated with an increase in the value of the other variable, or a decrease in the value of one variable is associated with the decrease in the value of the other variable, then the correlation is said to be positive.

On the other hand, if the values of two variables move in opposite direction, so that with an increase in the value of one variable, the value of the other variable decreases, or with a decrease in the value of one variable the value of the other variable increases, then the correlation is said to be negative.

Correlation can be, simple, multiple and partial.
In simple correlation, we study only two variables – say price and demand.

In multiple correlation, we study together the relationship between three or more factors like production, rainfall and use of fertilizers.

In partial correlation, though more than two factors are involved, correlation is studied only between two factors and the other factors are assumed to be constant.

Correlation can be linear or non-linear.
The correlation between two variables is said to be linear if there is a corresponding unit change in the value of one variable then there is a constant change in the value of the other variable. That is in case of linear correlation, the relation between the variable x and y is of the type y is equal to a plus bx.

The correlation between two variables is said to be non-linear or curvilinear, if there is a corresponding unit change in the value of one variable then the other variable does not change at a constant rate but at a fluctuating rate.

When we say that there is linear relationship, there are 4 parameters for linearity.

1. Correlation - Measures how well the data is lined up. The more the data resembles a straight line, the higher the correlation to each other.
2. Slope - Measures the steepness of the data assuming the correlation is good. The

steeper the data slope is, the greater is the importance of the relationship. A change in the "X" variable will have a larger impact on the "Y" variable, and you will begin to see a pattern that represents the Moderate Correlation.

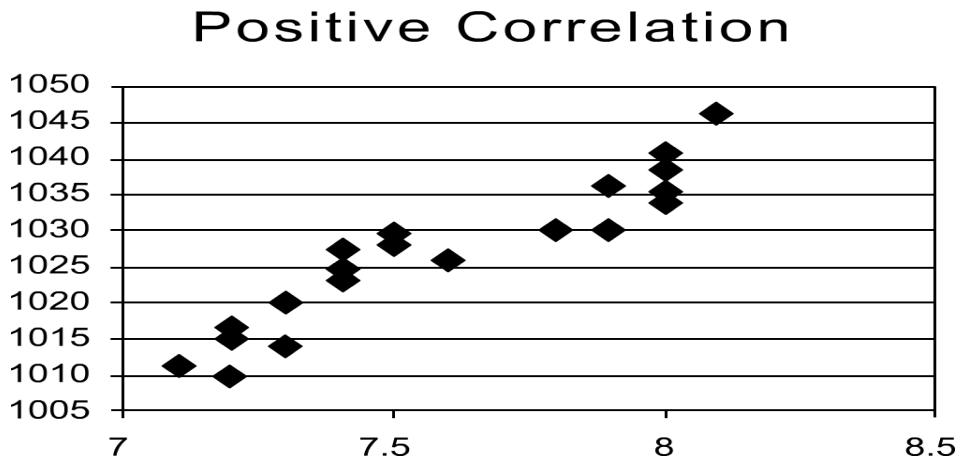
3. Direction - The "X" variable can have a positive or a negative impact on the "Y" variable. As one factor goes up, the other goes down. In a positive correlation, both factors will move in the same direction. In the graph examples below, you can see that the positive correlation moves from the lower left, towards the upward right. The negative correlation moves from the lower right, towards the upward left.

Y Intercept – It is a line drawn through the data crosses the "Y" axis. For a positive correlation, it represents the minimum "Y" value. For a negative correlation, it presents the maximum "Y" value.

4. Example of Correlation

In the this figure, you can see that the data pattern is moving from the bottom left upward to the top right indicating a positive correlation between the data. This is an upward sloping data grouping.

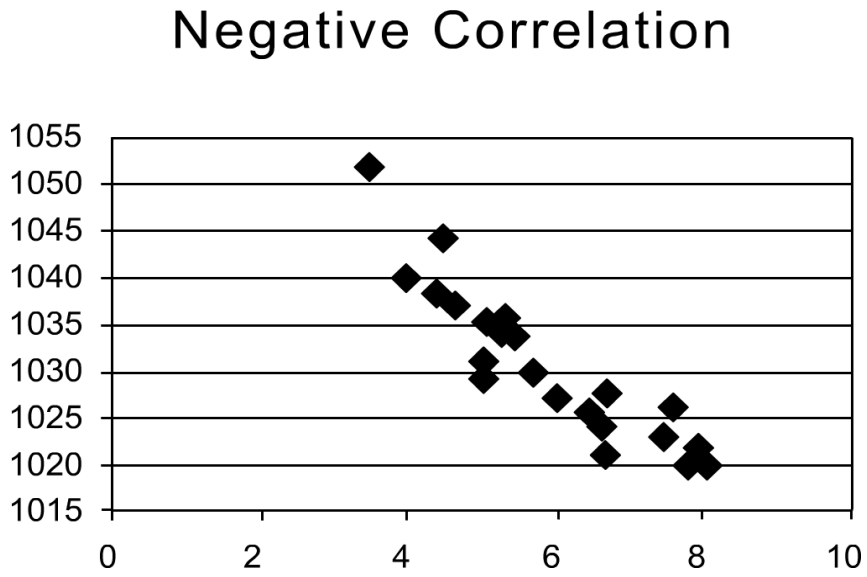
Figure 12



For example,

- Age of husband and age of wife
- Demand of the commodity and its supply
- Increase in rainfall (up to a point) and production of rice
- Increase in heat and temperature
- Increase in advertisement cost and volume of sales
- Increase in height and weight

Figure 13



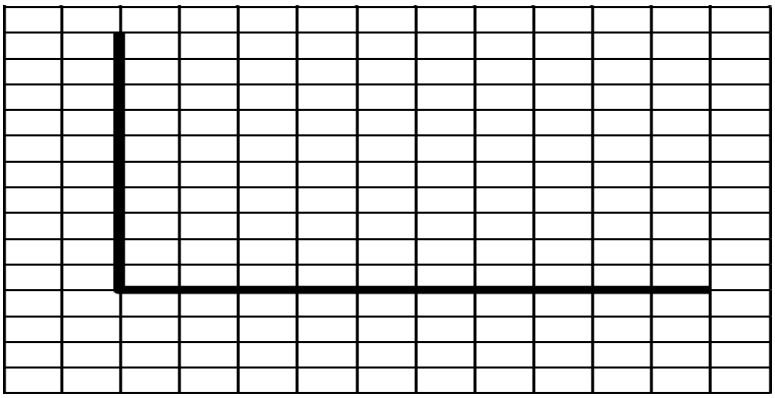
Conversely, in this figure, the data pattern moving from the top left downward to the bottom right indicating a negative correlation between the data. This is a downward sloping data grouping.

For example,

- Demand of a commodity may go down as a result of rise in prices
- Increase in the number of television sets and the number of cinemagoers
- Sales of woolen garments and the day temperature

Now, let us discuss how to make a scatter diagram. That is the basic scatter diagram layout.

Figure 14



Step 1 - Draw an "L" form just as you did for the pareto diagram (see the below figure). Make your scale units at even multiples, such as 10, 20, etc. so as to have an even scale system.

Step 2 - On the Horizontal axis (Known as the "X" axis, from Left to Right) you place the Independent or "cause" variable.

Step 3 - On the Vertical axis (Known as the "Y" axis, from Bottom to Top) you place the Dependent or "effect" variable.

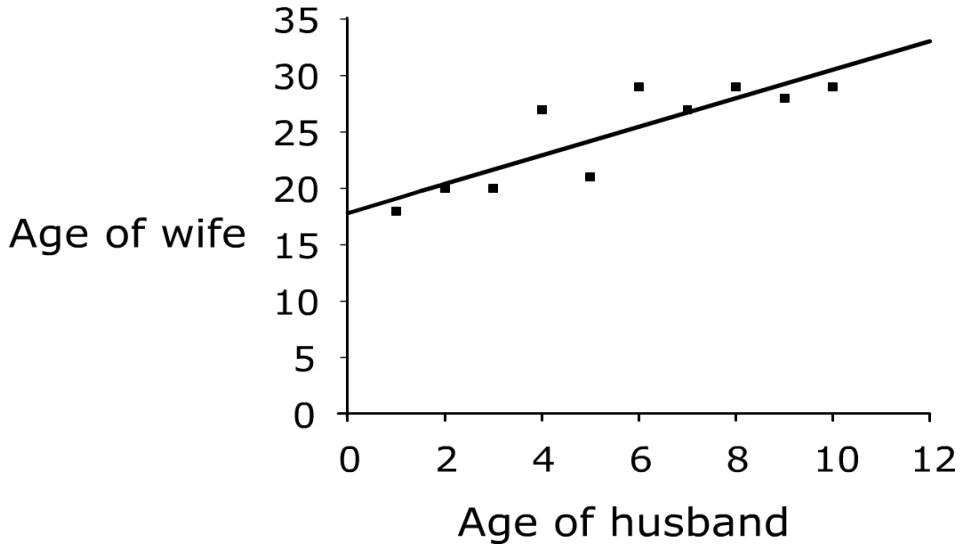
Step 4 - Plot your data points at the intersection of your data plots of the X and Y values. For Example, if X = 5, Y = 2. Go right 5 spaces, and then go up 2 spaces to plot the point.

Now, let us consider the following illustrations. Here, we construct the scatter diagrams for the given data and try to identify the type of relationship between the two variables.

Illustration 1:
Estimate the coefficient of correlation for the following data regarding ages of husbands and wives with the help of scatter diagram.

Figure 15

Age of Husband	23	27	28	28	29	30	31	33	35	36
Age of Wife	18	20	20	27	21	29	27	29	28	29



Observe that there is high and positive correlation between the ages of husbands and the ages of wives.

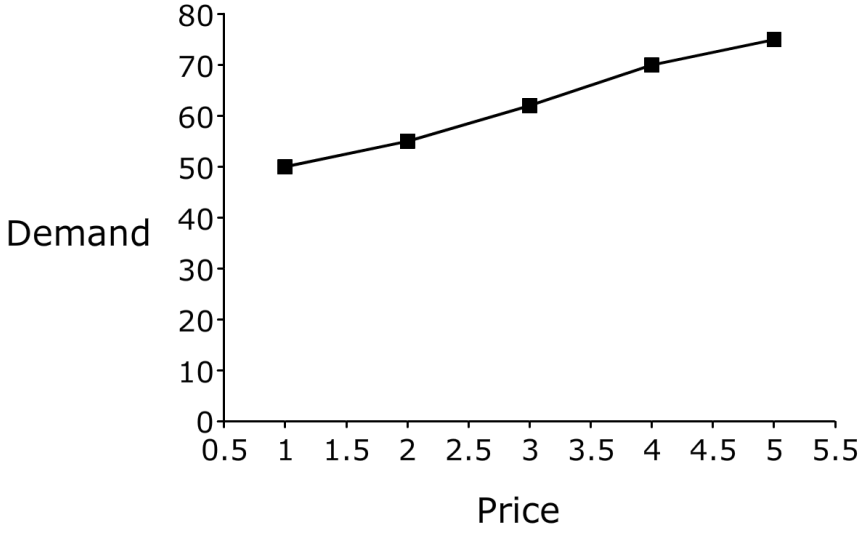
Illustration 2:
For the following data regarding price and demand, estimate the coefficient of correlation with the help of a scatter diagram.

Figure 17

Price	10	12	14	16	18
Demand	25	20	16	10	3

Let us construct the scatter diagram for the above data by taking price on X-axis and demand on Y-axis. Let us draw an imaginary line to identify the type of relation between the variables.

Figure 18



Observe the high degree or we can say perfectly negative correlation between the price and demand for a particular commodity. That is as price of the commodity increase, demand decreases.

5. Illustration

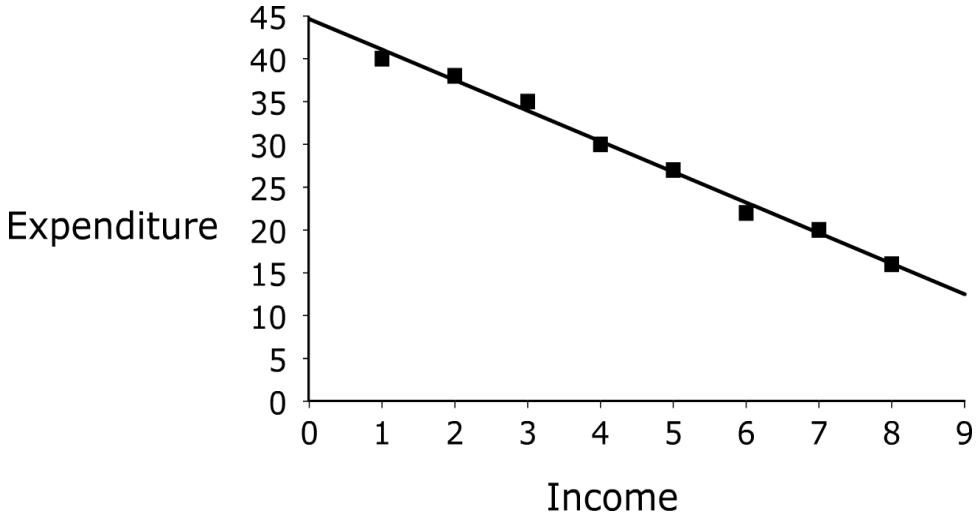
Illustration 3:
From the following data, ascertain with the help of scatter diagram, whether the income and expenditure of the workers of a factory are correlated. The table gives the average income and average expenditure per month in thousand for 8 years.

Figure 19

Year	2001	2002	2003	2004	2005	2006	2007	2008
Income (Rs.)	210	215	218	222	230	236	245	255
Expenditure (Rs.)	205	208	212	218	225	230	237	250

Solution:
Let us construct the scatter diagram for the above data by taking income along the X-axis and expenditure along the Y-axis. Once we plot all the points, we draw an imaginary line to show the relationship around the plotted points.

Figure 20



Observe that above graph shows that there is a very strong positive relationship between the variables income and expenditure of the workers in the factory. Therefore, as the average income increases, the average expenditure also increases.

Illustration 4:
The following table shows the change in the price for every change in the supply of a particular commodity. Table shows supply and price in rupees per unit over the years from 2 thousand 2 to 2 thousand 9.

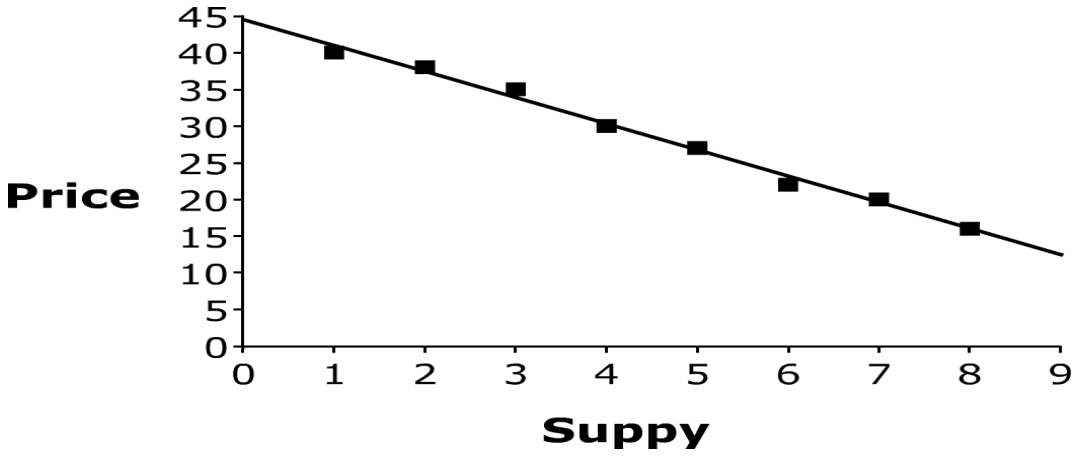
Figure 21

Year	2002	2003	2004	2005	2006	2007	2008	2009
Supply	50	55	62	70	75	78	80	82
Price	40	38	35	30	27	22	20	16

Solution:

Let us construct the scatter diagram for the above data by taking supply along the y-axis and price along the X-axis. An imaginary line is drawn through the plotted points in the scatter diagram. It shows the line moving from up side to down. That is there is a negative correlation between the two variables price and supply of the commodities. That is as supply increases, the price of the commodities decreases.

Figure 22



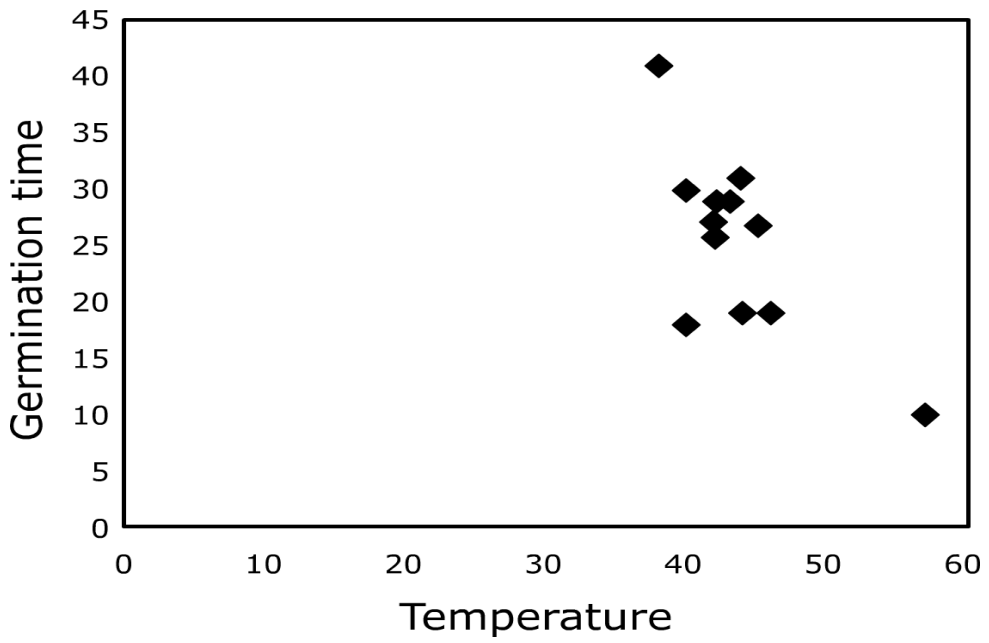
The table (in the right) shows the soil temperature and the germination time required for the seeds at various places. With the help of scatter diagram, verify whether there is any association between the two variables.

Figure 23

Temperature	Germination Time
57	10
42	26
40	30
38	41
42	29
45	27
42	27
44	19
40	18
46	19
44	31
43	29

We construct the scatter diagram for the data in the above table. To plot the scatter diagram, we take temperature along the X-axis and germination time along the Y-axis.

Figure 24



From the scatter diagram, we can observe that there is slightly negative relationship between the two variables soil temperature and germination time for the seeds. That is as temperature increases, the germination time decreases for the seeds and vice versa.

Let us list the merits and demerits of the scatter diagram.
Following are the merits of the scatter diagram:

- It is very easy to draw a scatter diagram.
- It can be easily understood and interpreted as we can easily identify whether the scattering of the points are linear or non-linear.
- Values of extreme items do not affect this method. Such points are always isolated in the diagram.

Following are the demerits of the scatter diagram:

- The major limitation of this method is that it only gives a visual picture of the relationship of two variables. It only tells us whether there is correlation between the variables, and if so, then in which direction, positive or negative.
- It does not give an idea about the precise degree of relationship, as it is not amenable to mathematical treatment.

Here is a summary of our learning of this session, where we have understood:

- The scatter diagram
- Correlation
- Utility of correlation
- Types of correlation
- About linear relationship between the variables

- The construction of scatter diagram
- Merits and demerits of scatter diagram