Correlation Analysis

Dr. Mohammad Rafiqul IslamAssociate
Professor of StatisticsDept. of Mathematics
and Natural Sciences (MNS)
Brac University, Dhaka,Bangladesh
mrafiq@bracu.ac.bd

Introduction

- Correlation analysis is one of the most widely used statistical methods to find the relationship between the variables.
- In all sciences, natural, social or biological as well as in business, we are largely concerned with the study of interrelationships among variables.
- For example, we may be intersected to know the relationship between the age and weight of the students, between family income and expenditure, or the relationship between price of a commodity and amount demand.
- Correlation analysis enables us to measure or quantify the relationship between the pairs of variables.

Definition

- The statistical tool for discovering and measuring the relationship between two or more variables is studied is called correlation analysis.

- Correlation shows whether and how strongly pairs of variables are related.

Measurement of relationship

- The relationship between the variables is measured by the Pearson's correlation coefficient is denoted by 'r'.
- It was introduced by Galton in 1877 and developed later by Karl Pearson in 1895, is most widely used in practice
- Thus r is also knon as Karl Pearson or simply Pearson's correlation coefficient
- the relationship between two variables x and y, r is defined as

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

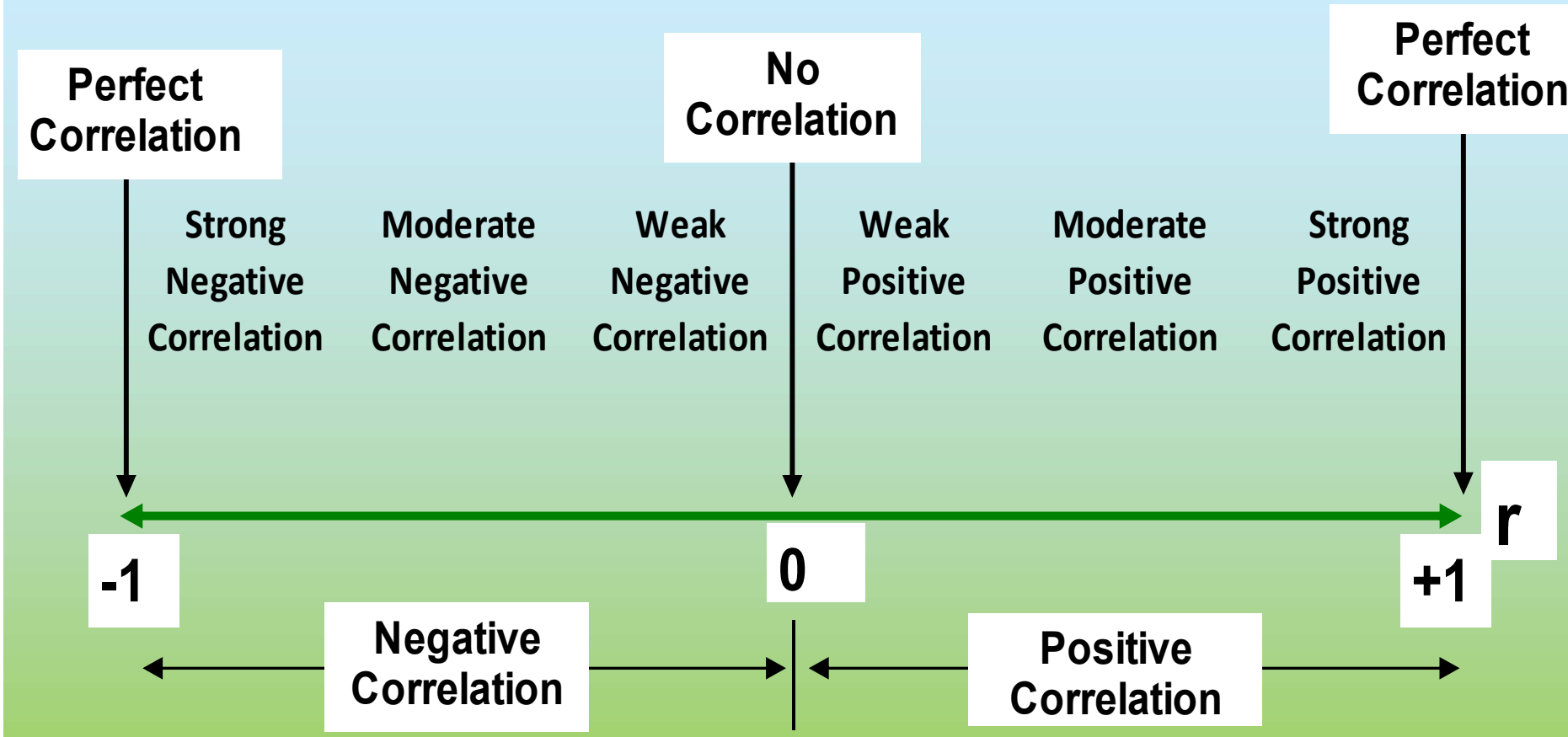
- The formula can be written for the calculation as:

$$r = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum x_i^2 - n\bar{x}^2)(\sum y_i^2 - n\bar{y}^2)}}$$

- The correlation coefficient, r is used for describing the degree and direction of relationship between two variables.

Interpretation of r

- The value of the correlation coefficient r lies between -1 and +1;
- The closer the value of r to either +1 or -1, the more strongly the two variables is related.
- When r is positive, one variable tends to increase as the other increases or vice versa;
- When r is negative, one variable tends to decrease as the other increases or vice versa;
- When r = +1, it means there is perfect positive correlation between the variables;
- When r = -1, it means there is perfect negative correlation between the variables;
- When r = 0, the variables are said to be uncorrelated.


Perfect Correlation
No Correlation
Perfect Correlation
Strong Negative Correlation
Moderate Negative Correlation
Weak Negative Correlation
Weak Positive Correlation
Moderate Positive Correlation
Strong Positive Correlation
-1
0
+1
r
Negative Correlation
Positive Correlation

Strength of correlation

Perfect	+1	−1
Strong	+0.9	−0.9
	+0.8	−0.8
	+0.7	−0.7
Moderate	+0.6	−0.6
	+0.5	−0.5
	+0.4	−0.4
Weak	+0.3	−0.3
	+0.2	−0.2
	+0.1	−0.1
Zero	0	

Practical Example

Find the correlation between age and weight and comment

Age (years)	Height (inches)
6	40
7	42
8	45
9	47
10	49
11	52
12	55
13	59
14	62
15	65

Age (years) x	Height (inches) ,y	xy	x^2	y^2
6	40	6×40=240	6×6=36	40×40=1600
7	42	7×42=294	7×7=49	42×42=1764
8	45	360	64	2025
9	47	423	81	2209
10	49	490	100	2401
11	52	572	121	2704
12	55	660	144	3025
13	59	767	169	3481
14	62	868	196	3844
15	65	975	225	4225
∑x=105	∑y=516	∑xy =5649	$\sum x^2$ =1185	$\sum y^2$= 27278

n =10

$$\bar{x} = \frac{\sum x}{n} = \frac{105}{10} = 10.5 \ \ and \ \ \bar{y} = \frac{\sum y}{n} = \frac{516}{10} = 51.6$$

r

We calculate the correlation coefficient r as follows

$$r = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum x_i^2 - n\bar{x}^2)(\sum y_i^2 - n\bar{y}^2)}}$$

$$r = 0.93$$

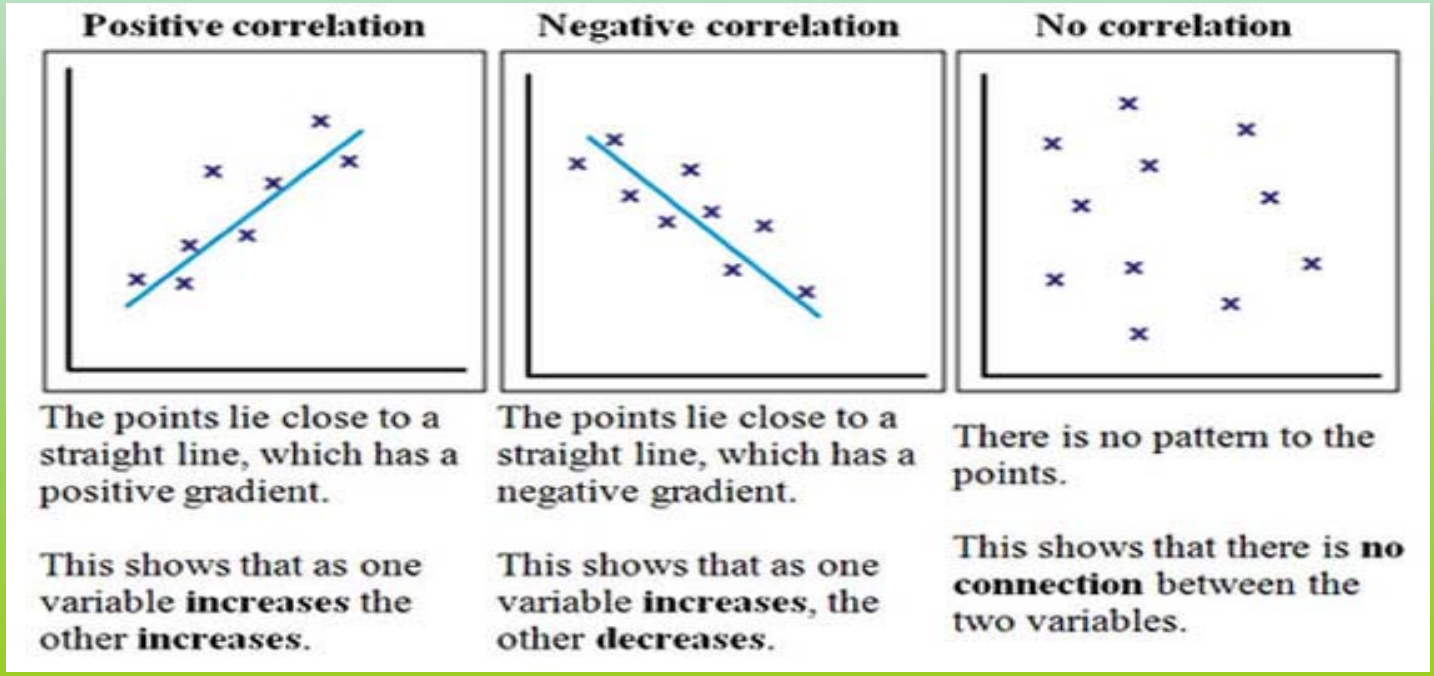
So there is positive and strong correlation between age and height. It means that when age increases, height also increases; or when age decreases, height also decreases.

Scatter Diagram

Scatter diagram or scatter plot, is a helpful way to visualize a relationship and to identify patterns between two variables.

In scatter diagram the independent variable x scaled on the horizontal axis and the dependent variable y on the vertical axis

Interpretation



Example 1*:** Let ***x and ***y*** represent the two variables. Examine the relationship by scatter diagram.

Income ('000' tk) X	Expenditure('000' tk) y
5	4
10	8
15	12
20	18
30	20
50	45

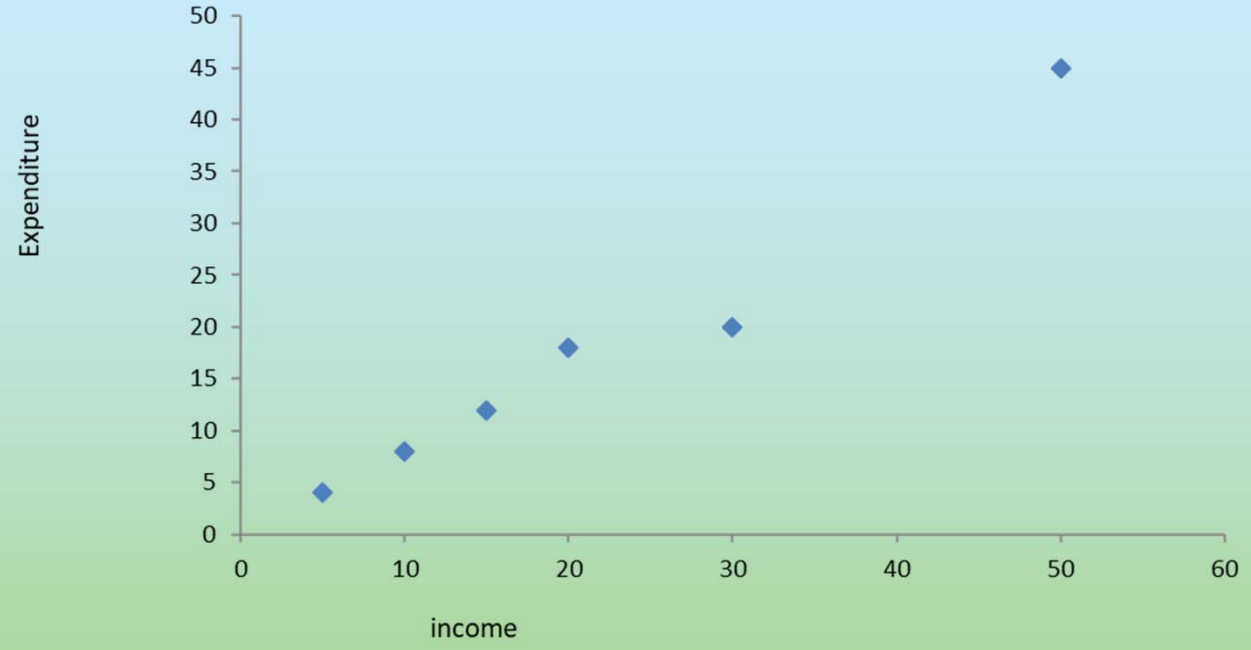


Figure: Scatter diagram between income and expenditure

The scatte diagram shows that there exists a ***positive*** correlation. We say that there is positive relationship between income, ***x*** and the expenditure ***y*** which means for higher income expenditure will also be higher or vice-versa.

Example 2: The price and amount of demand is given below. Draws scatter diagram and comment

Price ($)	Demand (Kg)
20	4
19	4
18	5
16	6
14	6
12	8
10	9
10	9
9	10
8	10

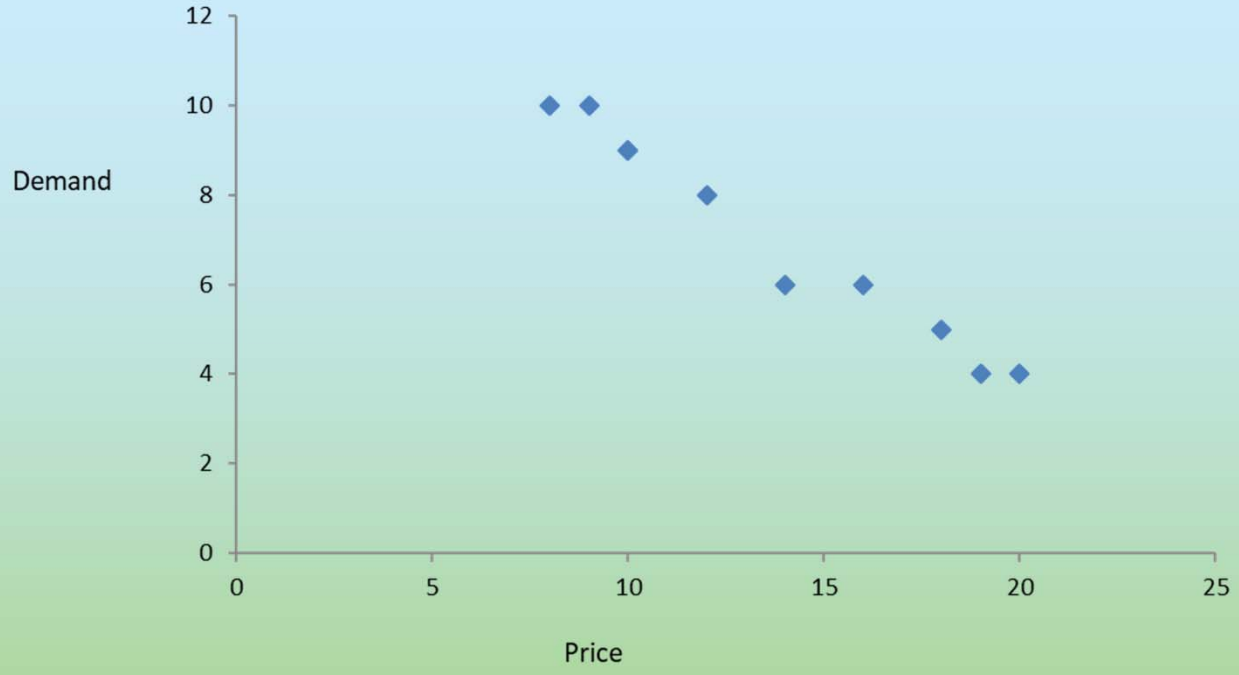


Figure: Scatter diagram between price and demand

In this case, the correlation is ***negative***. There is negative relationship between the price, ***x*** and the amount demand, ***y***, when price increase, demand decreases