# Lecture.12

## Correlation – definition – Scatter diagram -Pearson's correlation co-efficient – properties of correlation coefficient

### Correlation

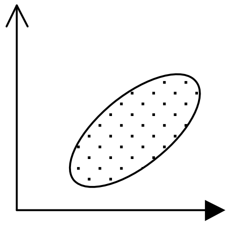
Correlation is the study of relationship between two or more variables. Whenever we conduct any experiment we gather information on more related variables. When there are two related variables their joint distribution is known as bivariate normal distribution and if there are more than two variables their joint distribution is known as multivariate normal distribution.

In case of bi-variate or multivariate normal distribution, we are interested in discovering and measuring the magnitude and direction of relationship between 2 or more variables. For this we use the tool known as correlation.

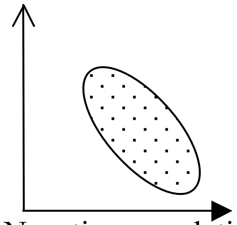
Suppose we have two continuous variables X and Y and if the change in X affects Y, the variables are said to be correlated. In other words, the systematic relationship between the variables is termed as correlation. When only 2 variables are involved the correlation is known as simple correlation and when more than 2 variables are involved the correlation is known as multiple correlation. When the variables move in the same direction, these variables are said to be correlated positively and if they move in the opposite direction they are said to be negatively correlated.

### Scatter Diagram

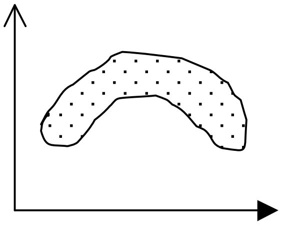
To investigate whether there is any relation between the variables X and Y we use scatter diagram. Let $(x_1,y_1)$, $(x_2,y_2)\ldots.(x_n,y_n)$ be n pairs of observations. If the variables X and Y are plotted along the X-axis and Y-axis respectively in the x-y plane of a graph sheet the resultant diagram of dots is known as scatter diagram. From the scatter diagram we can say whether there is any correlation between x and y and whether it is positive or negative or the correlation is linear or curvilinear.



Positive Correlation

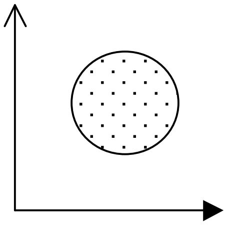


Negative correlation



Curvilinear

(or) non linear



no correlation

**Pearsons Correlation coefficient**

The measures of the degree of relationship between two continuous variables is called correlation coefficient. It is denoted by r.( in case of sample )and ρ (in case of population). The correlation coefficient r is known as Pearson's correlation coefficient as it was discovered by Karl Pearson. It is also called as product moment correlation.

The correlation coefficient r is given as the ratio of covariance of the variables X and Y to the product of the standard deviation of X and Y.

Symbolically,

$$r = \frac{\frac{1}{n-1}\left(\sum (x-\bar{x})(y-\bar{y})\right)}{\sqrt{\frac{1}{n-1}\sum (x-\bar{x})^2}\sqrt{\frac{1}{n-1}\sum (y-\bar{y})^2}}$$

which can be simplified as

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}\sqrt{\sum y^2 - \frac{\left(\sum y\right)^2}{n}}}$$

This correlation coefficient r is known as Pearson's Correlation coefficient. The numerator is termed as sum of product of X and Y and abbreviated as SP(XY). In the denominator the first term is called sum off squares of X (i.e) SS(X) and second term is called sum of squares of Y (i.e) SS(Y)

$$\therefore r = \frac{SP(XY)}{\sqrt{SS(X)}\sqrt{SS(Y)}}$$

The denominator in the above formula is always positive. The numerator may be positive or negative making r to be either positive or negative.

**Assumptions in correlation analysis:**

Correlation coefficient r is used under certain assumptions, they are

1. The variables under study are continuous random variables and they are normally distributed
2. The relationship between the variables is linear
3. Each pair of observations is unconnected with other pair (independent)

**Properties**

1. The correlation coefficient value ranges between –1 and +1.
2. The correlation coefficient is not affected by change of origin or scale or both.
3. If r > 0 it denotes positive correlation
   r< 0 it denotes negative correlation between the two variables x and y.
   r = 0 then the two variables x and y are not linearly correlated.(i.e)two variables are independent.
   r = +1 then the correlation is perfect positive
   r = -1 then the correlation is perfect negative.

**Testing the significance of r**

The significance of r can be tested by Student's t test. The test statistics is given by

$$t = \frac{|r|}{\sqrt{\frac{1-r^2}{n-2}}}$$

This t is distributed as Student's t distribution with (n-2) degrees of freedom.

The relationship between the variables is interpreted by the square of the correlation coefficient ($r^2$) which is called coefficient of determination. The value 1-$r^2$ is called as coefficient of alienation. If $r^2$ is 0.72, it implies that on the basis of the samples 72% of the variation in one variable is caused by the variation of the other variable. The coefficient of determination is used to compare 2 correlation coefficients.

**Problem**

Compute Pearsons coefficient of correlation between plant height (cm) and yield (Kgs) as per the data given below:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Plant Height (cm) | 39 | 65 | 62 | 90 | 82 | 75 | 25 | 98 | 36 | 78 |
| Yield in Kgs | 47 | 53 | 58 | 86 | 62 | 68 | 60 | 91 | 51 | 84 |

**Solution**

$H_o$: The correlation coefficient r is not significant

$H_1$: The correlation coefficient r is significant.

Level of significance 5%

From the data

n = 10

$\sum x = 650$ $\sum y = 660$ $\sum xy = 45604$ $\sum x^2 = 47648$ $\sum y^2 = 45784$

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

$$= \frac{45604 - \frac{(650)(660)}{10}}{\sqrt{47648 - \frac{(650)^2}{10}}\sqrt{45784 - \frac{(660)^2}{10}}}$$

$$= \frac{45604 - 42900}{(73.47)(47.1)} = 0.7804$$

Correlation coefficient is positively correlated.

**Test Statistic**

$$t \;=\; \frac{|r|}{\sqrt{\frac{1-r^2}{n-2}}} \sim (\text{n - 2})\ \text{d.f.}$$

$$t \;=\; \frac{0.7804}{\sqrt{\frac{1-(0.7804)^2}{10-2}}} = 3.530$$

$t_{tab}=t_{\mathbf{(10\text{-}2,\ 5\%los)}}=2.306$

**Inference**

$t > t_{tab}$, we reject null hypothesis.

$\therefore$The correlation coefficient r is significant. (i.e) there is a relation between plant height and yield.

**Questions**

**1.** Limits for correlation coefficient.

(a) $-1 \le r \le 1$ (b) $0 \le r \le 1$

(c) $-1 \le r \le 0$ (d) $1 \le r \le 2$

**Ans: $-1 \le r \le 1$**

**2.** The correlation coefficient is unaffected by change of

(a) Origin (b) scale

(c) Scale & origin (d) None of these

**Ans: scale & origin**

3. When r = +1, there is Perfect positive correlation.

**Ans: True**

4. Karl pearsons correlation coefficient is calculated only when the two variables are continuous.

**Ans: True**

5. The correlation between two variables is symmetric

**Ans: True**

6. The correlation between two variables is known as multiple correlation.
**Ans: False**

**7.** What is a scatter diagram? Mention its uses

8. Define correlation.

9. Explain the method how to calculate the Karl pearsons correlation coefficient?

10. Mention the properties of the correlation coefficient?