PGDISM - SEMESTER-II
[e-Content]

Course/Paper Code – 06(201)

SAFETY STATISTICS AND ACCIDENT INSPECTION

UNIT-4

MR. RAJEEV KUMAR
[GUEST FACULTY]
PMIR DEPARTMENT
PATNA UNIVERSITY, PATNA

Email:-
rajeevk.patna@gmail.com

Mobile No.:
+91-6287858781

CONTENTS

UNIT- 4

UNIT-4

1. Introduction

So far we have discussed problem relating to one variables only, but now we will study the method of problem solving with the use of two or more variables which is used in large organization or businesses.

The statistical tool which tools which helps in solving the problems or relationship between two or more than two variables is called **correlation.**

1.1 Correlation

Correlation is statistical Analysis which measures and analyses the degree or extent to which the two variables fluctuate with reference to each other. It indicates that there is some connection between the variables. It measures the closeness of the relationship. Correlation does not indicate cause and effect relationship.

Correlation expresses the **inter-dependence** of two sets of variables upon each other. One variable may be called as **independent** and other relative variable is called as **dependent**.

The measure of correlation is called the coefficient of correlation denoted by the symbol 'r'. Thus correlation analysis refers to the techniques used in measuring the closeness of the relationship between between the variables.

A very simple definition of correlation is given by **A. M.Tuttle** as : " An analysis of the covariation of two or more variables is usually called **correlation**."

1.2 Uses of correlation:

1. It is used in physical and social sciences.
2. It is useful for economists to study the relationship between variables like price, quantity etc. Businessmen estimates costs, sales, price etc. using correlation.

3. It is helpful in measuring the degree of relationship between the variables like income and expenditure, price and supply, supply and demand etc.
4. Sampling error can be calculated.
5. It is the basis for the concept of regression.

Merits:

1. It is a simplest and attractive method of finding the nature of correlation between the two variables.
2. It is a non-mathematical method of studying correlation. It is easy to understand.
3. It is not affected by extreme items.
4. It is the first step in finding out the relation between the two variables.
5. We can have a rough idea at a glance whether it is a positive correlation or negative correlation.

Demerits:

By this method we cannot get the exact degree or correlation between the two variables.

1.3 Types of Correlation:

Correlation is classified into various types.The most important ones are:

i) Positive and negative.
ii) Linear and non-linear.
iii) Partial and total.
iv) Simple and Multiple

Positive and Negative Correlation

If both the variables move in the same direction, we say that there is a positive correlation, *i.e.,* if one variable increases, the other variable also increases on an average or if one variable decreases, the other variable also decreases on an average.

On the other hand, if the variables are varying in opposite direction, we say that it is a case of negative

correlation; *e.g.,* movements of demand and supply.

Linear and Non-linear correlation:

If the ratio of change between the two variables is a constant then there will be linear correlation between them.

Consider the following.

X	2	4	6	8	10	12
Y	3	6	9	12	15	18

Here the ratio of change between the two variables is the same. If we plot these points on a graph we get a straight line.

If the amount of change in one variable does not bear a constant ratio of the amount of change in the other. Then the relation is called Curvi-linear (or) non-linear correlation. The graph will be a curve.

Simple and Multiple correlation:

When we study only two variables, the relationship is simple correlation. For example, quantity of money and price level, demand and price. But in a multiple correlation we study more than two variables simultaneously. The relationship of price, demand and supply of a commodity are an example for multiple correlation.

Partial and total correlation:

The study of two variables excluding some other variable is called **Partial correlation**. For example, we study price and demand eliminating supply side. In total correlation all facts are taken into account.

Computation of correlation:

When there exists some relationship between two variables, we have to measure the degree of relationship. This measure is called the measure of correlation (or) correlation coefficient and it is denoted by ‘ r’ .

The commonly used methods for studying linear relationship between two variables involve both graphic and algebraic methods. Some of the widely used methods include:

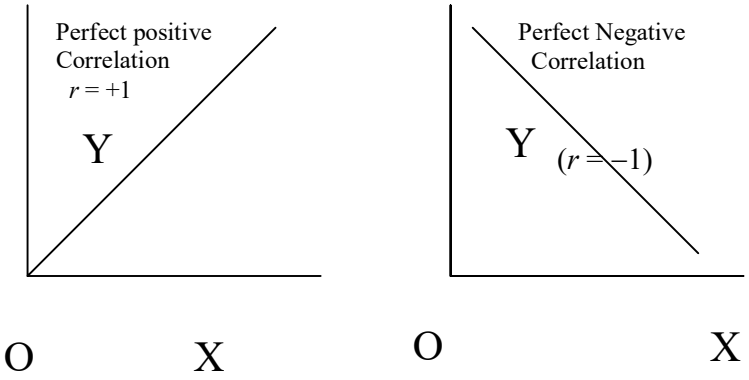
1. Scatter Diagram
2. Correlation Graph
3. Pearson's Coefficient of Correlation
4. Spearman's Rank Correlation
5. Concurrent Deviation Method

1.4 Scatter Diagram:

It is the simplest method of studying the relationship between two variables diagrammatically. One variable is represented along the horizontal axis and the second variable along the vertical axis. For each pair of observations of two variables, we put a dot in the plane. There are as many dots in the plane as the number of paired observations of two variables. The direction of dots shows the scatter or concentration of various points. This will show the type of correlation.

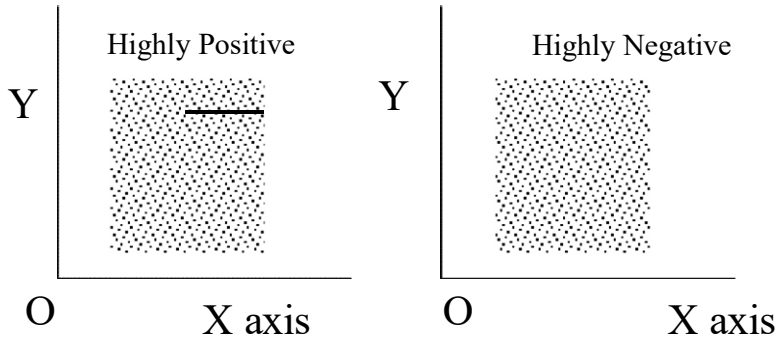
1. If all the plotted points form a straight line from lower left hand corner to the upper right hand corner then there is Perfect positive correlation.

We denote this as $r = +1$



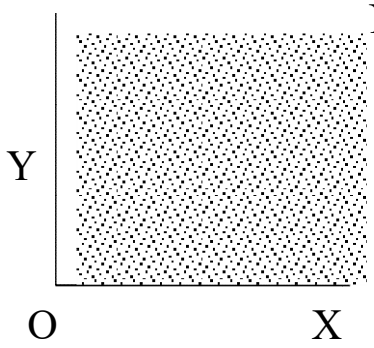
1. If all the plotted dots lie on a straight line falling from upper left hand corner to lower right hand corner, there is a perfect negative correlation between the two variables. In this case the coefficient of correlation takes the value $r = -1$.

2. If the plotted points in the plane form a band and they show a rising trend from the lower left hand corner to the upper right hand corner the two variables are highly positively correlated.



1. If the points fall in a narrow band from the upper left hand corner to the lower right hand corner, there will be a high degree of negative correlation.
2. If the plotted points in the plane are spread all over the diagram there is no correlation between the two variables.

No correlation ($r = 0$)



1.5 Merits and Limitations of Scatter Diagram:

Merits:

1. It is a simple and non mathematical method of studying correlation between the variables.
   It can be easily understood and a rough idea can be very quickly be formed as to whether or not the variables are related.
2. It is not influenced by the size of extreme values whether most of the mathematical methods of finding correlation are influenced by extreme value.
3. Making a scatter diagram usually the first step in investigating the relationship between the variables.

Limitations:

By applying this method we can get an idea about the direction of correlation and also whether it is high or low, but we cannot establish the exact degree of correlation between the variables as is possible by applying the mathematical method.

1.6 Rank Correlation:

It is studied when no assumption about the parameters of the population is made. This method is based on ranks. It is useful to study the qualitative measure of attributes like honesty, colour, beauty, intelligence, character, morality etc.The individuals in the group can be arranged in order and there on, obtaining for each individual a number showing his/her rank in the group. This method was developed by Edward Spearman in 1904. It is defined as $r = 1 - \frac{6\Sigma D^2}{n^3 - n}$ r = rank correlation coefficient.

Note: Some authors use the symbol ρ for rank correlation.
ΣD^2 = sum of squares of differences between the pairs of ranks. n = number of pairs of observations.

The value of r lies between –1 and +1. If r = +1, there is complete agreement in order of ranks and the direction of ranks is also same. If r = -1, then there is complete disagreement in order of ranks and they are in opposite directions.

Computation for tied observations: There may be two or more items having equal values. In such case the same rank is to be given. The ranking is said to be tied. In such circumstances an average rank is to be given to each individual item. For example if the value so is repeated twice at the 5th rank, the common rank to be assigned to each item is $\frac{5+6}{2} = 5.5$ which is the average of 5 and 6 given as 5.5, appeared twice.

If the ranks are tied, it is required to apply a correction factor which is $\frac{1}{12}(m^3-m)$. A slightly different formula is used when there is more than one item having the same value.

The formula is

$$r = 1 - \frac{6\left[\Sigma D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \quad \right]}{n^3 - n}$$

Where m is the number of items whose ranks are common and should be repeated as many times as there are tied observations.

Example :

In a marketing survey the price of tea and coffee in a town based on quality was found as shown below. Could you find any relation between and tea and coffee price.

Price of tea	**88**	**90**	**95**	**70**	**60**	**75**	**50**
Price of coffee	**120**	**134**	**150**	**115**	**110**	**140**	**100**

Price of tea	**Rank**	**Price of coffee**	**Rank**	**D**	**D^2**
88	**3**	**120**	**4**	**1**	**1**
90	**2**	**134**	**3**	**1**	**1**
95	**1**	**150**	**1**	**0**	**0**
70	**5**	**115**	**5**	**0**	**0**
60	**6**	**110**	**6**	**0**	**0**
75	**4**	**140**	**2**	**2**	**4**
50	**7**	**100**	**7**	**0**	**0**
					$\Sigma D^2 = 6$

$$\frac{6\Sigma D^2}{n^3 - n} = 1 - \frac{6 \times 6}{7^3 - 7}$$

r = 1 – = 1 – 0.1071 = 1 – $\frac{3}{6}$ $\frac{}{33}$ 6

= 0.8929

The relation between price of tea and coffee is positive at 0.89. Based on quality the association between price of tea and price of coffee is highly positive.

1.7 Merits and Limitations of Rank Method :

Merits :

1. This method is simpler to understand and easier to apply compared to the Karl Pearson's Method.
2. Whether the data are of a qualitative nature like honesty , efficiency , intelligence , etc – this method can be used with great advantage.
3. This is the only method that can be used where we are given the ranks and not the actual data.
4. Even where actual data are given , rank method can be applied for ascertaining rough degree of correlation.

Limitation:

1. This method can be used for finding out correlation in a grouped frequency distribution .
2. Where the number of observations exceeds 30 , the calculations became quite tedious and require a lot of time.

Therefore this method should not be applied where N is exceeding 30 unless we are given the ranks and not the actual values of the variable.

Example

Calculate the rank coefficient of correlation from the following data:

***X*:**	**75**	**88**	**95**	**70**	**60**	**80**	**81**	**50**
***Y*:**	**120**	**134**	**150**	**115**	**110**	**140**	**142**	**100**

Solution:

Calculations for Coefficient of Rank Correlation

X	Ranks R_X	Y	Ranks R_Y	$d = R_X - R_Y$	d^2
75	5	120	5	0	0
88	2	134	4	-2	4
95	1	150	1	0	0
70	6	115	6	0	0
60	7	110	7	0	0
80	4	140	3	+1	1
81	3	142	2	+1	1
50	8	100	8	0	0

$$\sum d^2 = 6$$

$$\rho = 1 - \frac{6\sum d^2}{N(N^2 - 1)}$$

$$= 1 - \frac{6 \times 6}{8(8^2 - 1)}$$

$$= 1 - \frac{36}{504}$$

$= 1 - 0.07$

$= + 0.93$

Hence, there is a high degree of positive correlation between X and Y

2. Definition:

Regression is the measure of the average relationship between two or more variables in terms of the original units of the data.

2.1 Types Of Regression:

The regression analysis can be classified into:

a) Simple and Multiple
b) Linear and Non –Linear
c) Total and Partial

a) Simple and Multiple:

In case of simple relationship only two variables are considered, for example, the influence of advertising expenditure on sales turnover. In the case of multiple relationship, more than two variables are involved. On this while one variable is a dependent variable the remaining variables are independent ones.

For example, the turnover (y) may depend on advertising expenditure (x) and the income of the people (z). Then the functional relationship can be expressed as $y = f(x,z)$.

b) Linear and Non-linear:

The linear relationships are based on straight-line trend, the equation of which has no-power higher than one. But, remember a linear relationship can be both simple and multiple. Normally a linear relationship is taken into account because besides its simplicity, it has a better predective value, a linear trend can be easily projected into the future. In the case of non-linear relationship curved trend lines are derived. The equations of these are parabolic.

c) Total and Partial:

In the case of total relationships all the important variables are considered. Normally, they take the form of a multiple relationships

because most economic and business phenomena are affected by multiplicity of cases. In the case of partial relationship one or more variables are considered, but not all, thus excluding the influence of those not found relevant for a given purpose.

2.2 Linear Regression Equation:

If two variables have linear relationship then as the independent variable (X) changes, the dependent variable (Y) also changes. If the different values of X and Y are plotted, then the two straight lines of best fit can be made to pass through the plotted points. These two lines are known as regression lines. Again, these regression lines are based on two equations known as regression equations. These equations show best estimate of one variable for the known value of the other. The equations are linear.

Linear regression equation of Y on X is $Y = a + bX$ ….. (1)

And X on Y is $X = a + bY$ ….. (2)

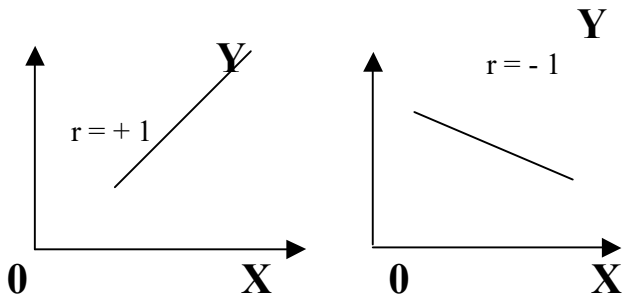
a, b are constants.

From (1) We can estimate Y for known value of X.
(2) We can estimate X for known value of Y.

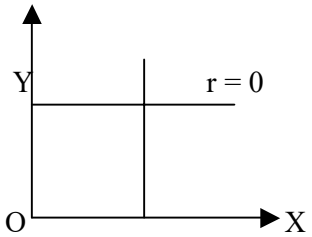
2.3 Regression Lines:

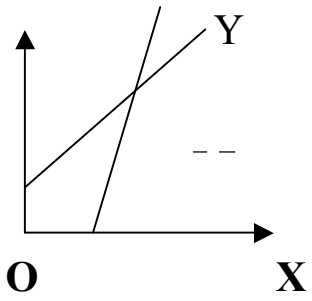
For regression analysis of two variables there are two regression lines, namely Y on X and X on Y. The two regression lines show the average relationship between the two variables.

For perfect correlation, positive or negative i.e., $r = \pm 1$, the two lines coincide i.e., we will find only one straight line. If r = 0, i.e., both the variables are independent then the two lines will cut each other at right angle. In this case the two lines will be parallel to X and Y-axes.



Lastly the two lines intersect at the point of means of X and Y. From this point of intersection, if a straight line is drawn on X- axis, it will touch at the mean value of x. Similarly, a perpendicular drawn from the point of intersection of two regression lines on Y- axis will touch the mean value of Y.





$(\bar{x}, \bar{y})$

2.4 Principle of ' Least Squares' :

Regression shows an average relationship between two variables, which is expressed by a line of regression drawn by the method of "least squares". This line of regression can be derived graphically or algebraically. Before we discuss the various methods let us understand the meaning of least squares.

A line fitted by the method of least squares is known as the line of best fit. The line adapts to the following rules:

(i) The algebraic sum of deviation in the individual observations with reference to the regression line may be equal to zero. i.e.,
$\sum(X - Xc) = 0$ or $\sum (Y - Yc) = 0$

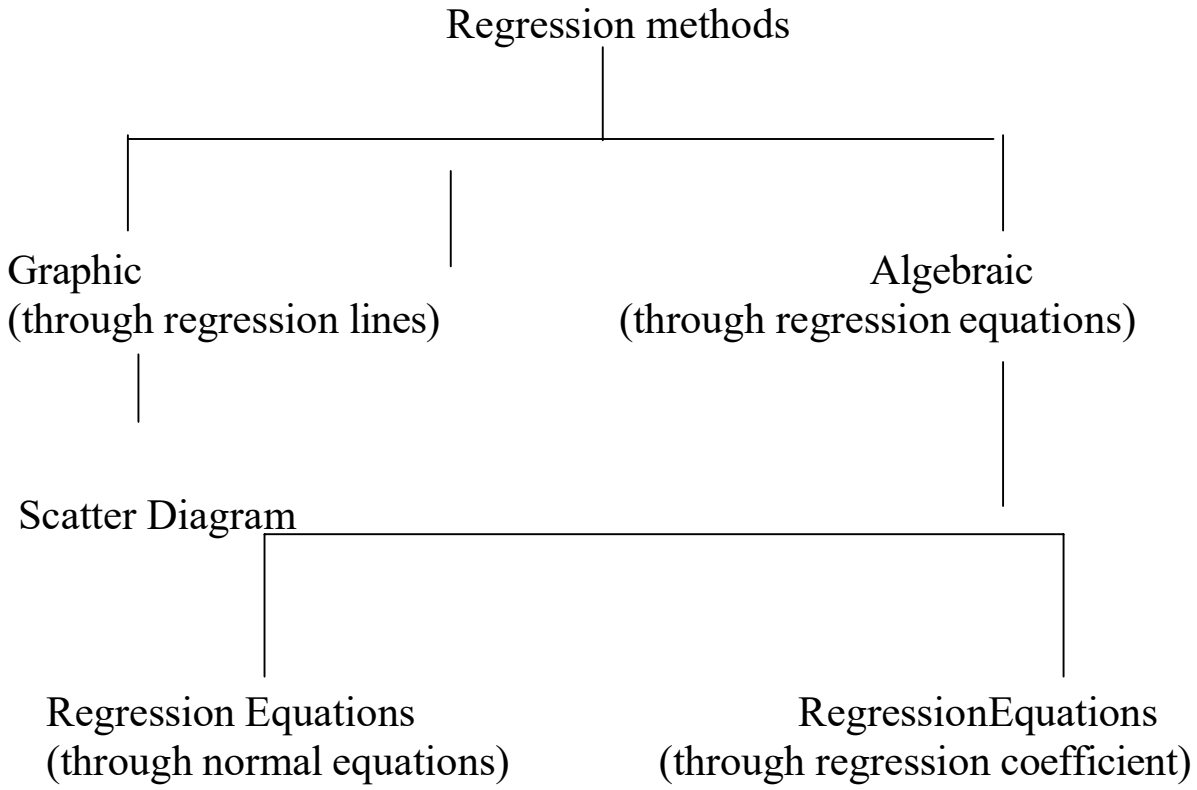
Where Xc and Yc are the values obtained by regression analysis.

(ii) The sum of the squares of these deviations is less than the sum of squares of deviations from any other line. i.e.,
$\sum(Y - Yc)^2 < \sum (Y - Ai)^2$
Where Ai = corresponding values of any other straight line.

(iii) The lines of regression (best fit) intersect at the mean values of the variables X and Y, i.e., intersecting point is $\bar{x}, \bar{y}$.

2.5 Methods of Regression Analysis:

The various methods can be represented in the form of chart given below:



2.6 Graphic Method:

Scatter Diagram:

Under this method the points are plotted on a graph paper representing various parts of values of the concerned variables. These points give a picture of a scatter diagram with several points spread over. A regression line may be drawn in between these points either by free hand or by a scale rule in such a way that the squares of the vertical or the horizontal distances (as the case may be) between the points and the line of regression so drawn is the least. In other words, it should be drawn faithfully as the line of best fit leaving equal number of points on both sides in such a manner that the sum of the squares of the distances is the best.

Algebraic Methods:

(i) Regression Equation.

The two regression equations for X on Y; $X = a + bY$

And for Y on X; $Y = a + bX$

Where X, Y are variables, and a,b are constants whose values are to be determined

For the equation, $X = a + bY$ The normal equations are

$$\sum X = na + b \sum Y \text{ and}$$
$$\sum XY = a\sum Y + b\sum Y^2$$

For the equation, $Y = a + bX$, the normal equations are

$$\sum Y = na + b\sum X \text{ and}$$
$$\sum XY = a\sum X + b\sum X^2$$

From these normal equations the values of a and b can be determined.

Example:

Find the two regression equations from the following data:

X:	**6**	**2**	**10**	**4**	**8**
Y:	**9**	**11**	**5**	**8**	**7**

Solution:

X	Y	X^2	Y^2	XY
6	9	36	81	54
2	11	4	121	22
10	5	100	25	50
4	8	16	64	32
8	7	64	49	56
30	40	220	340	214

Regression equation of Y on X is $Y = a + bX$ and the normal equations are

$$\sum Y = na + b\sum X$$
$$\sum XY = a\sum X + b\sum X^2$$

Substituting the values, we get

$$40 = 5a + 30b \ldots\ldots (1)$$
$$214 = 30a + 220b \ldots\ldots (2)$$

Multiplying (1) by 6

$$240 = 30a + 180b \ldots\ldots (3)$$

(2) – (3) $-26 = 40b$

or $b = \frac{-26}{40} = -0.65$

Now, substituting the value of 'b' in equation (1) $40 = 5a - 19.5$

$5a = 59.5$

$a = \frac{59.5}{5} = 11.9$

Hence, required regression line Y on X is $Y = 11.9 - 0.65\ X$. Again, regression equation of X on Y is

$X = a + bY$ and

The normal equations are

$$\sum X = na + b\sum Y \text{ and}$$
$$\sum XY = a\sum Y + b\sum Y^2$$

Now, substituting the corresponding values from the above table, we get

$30 = 5a + 40b$ … (3)

$214 = 40a + 340b$ … (4)

Multiplying (3) by 8, we get

$240 = 40a + 320\ b$ … (5)

(4) – (5) gives

$-26 = 20b$

$b = -\frac{26}{20} = -1.3$

Substituting $b = -1.3$ in equation (3) gives $30 = 5a - 52$

$5a = 82$

$a = \frac{82}{5} = 16.4$

Hence, Required regression line of X on Y is

X = 16.4 – 1.3Y

(ii)Regression Co-efficents:

The regression equation of Y on X is $y_e = y + \frac{r}{\sigma}\sigma_y (x - x)$

x

Here, the regression Co.efficient of Y on X is

$b_1 = = r\frac{\sigma_y}{\sigma}$

b_{yx}

$y_e = \bar{y} + b_1 (x - \bar{x})$

The regression equation of X on Y is

$X_e = x + \frac{r}{\sigma_y}\sigma_x (y - y)$

Here, the regression Co-efficient of X on Y

$b_2 = b_{xy} = r\frac{\sigma_x}{\sigma}$

y

$X_e = X + b_2 (y - \bar{y})$

If the deviation are taken from respective means of x and y

$b_1 = b_{yx} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2} = \frac{\sum xy}{\sum x^2}$ and

$b_2 = b_{xy} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(Y-\bar{Y})^2} = \frac{\sum xy}{\sum y^2}$

Where

$\bar{x} = X - \bar{X}, y = Y - \bar{Y}$

If the deviations are taken from any arbitrary values of x and y (short – cut method)

$b_1 = b_{yx} = \frac{n\sum uv - \sum u \sum v}{n\sum u^2 - (\sum u)^2}$

$b_2 = b_{xy} = \frac{n\sum uv - \sum u \sum v}{n\sum v^2 - (\sum v)^2}$

where u = x – A : v = Y-B

A = any value in X B = any value in Y

2.7 Properties of Regression Co-efficient:

1. Both regression coefficients must have the same sign, ie either they will be positive or negative.
2. correlation coefficient is the geometric mean of the regression coefficien $r = \pm\sqrt{b_1 b_2}$ ts ie,
3. The correlation coefficient will have the same sign as that of the regression coefficients.
4. If one regression coefficient is greater than unity, then other regression coefficient must be less than unity.
5. Regression coefficients are independent of origin but not of scale.
6. Arithmetic mean of b_1 and b_2 is equal to or greater than the coefficient of correlation. $\frac{b_1 + b_2}{2} \geq r$ Symbolically

7. If r=0, the variables are uncorrelated , the lines of regression become perpendicular to each other.
8. If r= $\pm$1, the two lines of regression either coincide or parallel to each other
9. Angle between the two regression lines is $\theta = \tan^{-1}\left[\frac{m_1 - m_2}{1 + mm}\right]$

   where m_1 and,m_2 are the slopes of the regression lines X on Y and Y on X respectively.
10. The angle between the regression lines indicates the degree of dependence between the variables.

2.8 Difference between Correlation and Regression Analysis:

There are two important points of difference between Correlation and Regression Analysis:

Whereas correlation coefficient is a measure of degree of relationship between X and Y, the objective of regression analysis is to study the 'nature of relationship' between the variables.

The cause and effect relation is clearly indicated through regression analysis than the correlation analysis. Correlation is merely a tool of ascertaining the degree of relationship between two variables and , therefore , we cannot say that one variable is the cause and the other is the effect.

2.9 Uses of Regression Analysis:

1. Regression analysis helps in establishing a functional relationship between two or more variables.
2. Since most of the problems of economic analysis are based on cause and effect relationships, the regression analysis is a highly valuable tool in economic and business research.
3. Regression analysis predicts the values of dependent variables from the values of independent variables.
4. We can calculate coefficient of correlation (r) and coefficient of determination (r^2) with the help of regression coefficients.
5. In statistical analysis of demand curves, supply curves, production function, cost function, consumption function etc., regression analysis is widely used.

REFRENCES

1. FUNDAMENTALS OF MATHEMATICAL STATISTIC
   S.C. GUPTA
   V.K. KAPOOR

2.STATISTICS
   TMT. V.VARALAKSHMI
   TMT. N.SUSEELA THIRU
   G.GNANA SUNDARAM
   TMT.S.EZHILARASI
   TMT. B.INDRANI

3. BUSINESS STATISTICS
   S.P. GUPTA
   M.P. GUPTA

4.AN INTRODUCTION TO BUSINESS STATISTICS
   SURINDER KUNDU

5. FUNDAMENTAL OF STATISTICS
   A.M. GOON
   M.K. GUPTA
   B. DASGUPTA

……