UNIT-III

STATISTICS

Topic Learning Objectives:

Upon Completion of this unit, students will be able to:

- Expand their knowledge and skills of the Statistical Concepts and a personal development experience towards the needs of statistical data analysis.
- Understand the Central Moments, Skewness and Kurtosis.
- Describe the principle of least squares.
- Fit data using several types of curves.
- Describe & evaluate the concept of correlation and regression coefficients.
- Investigate the strength and direction of a relationship between two variables by collecting measurements and using appropriate statistical analysis.

Introduction:

In many fields of Applied Mathematics and Engineering we face some problems and do the experiments involving two variables. In this chapter, we consider the Mathematical theory of statistics, by presenting an elementary treatment of Central moments, mean, variance, coefficients of skewness and kurtosis in terms of moments, curve fitting, correlation and regression. In mathematics, a moment is a specific quantitative measure of the shape of a function. It is used in both mechanics and statistics. If the function represents physical density, then the zeroth moment is the total mass, the first moment divided by the total mass is the center of mass, and the second moment is the rotational inertia. If the function is a probability distribution, then the zeroth moment is the total probability (i.e. one), the first moment is the mean, the second central moment is the variance, the third standardized moment is the skewness, and the fourth standardized moment is the kurtosis.

Moments:

In mechanics, moment refers to the turning or the rotating effect of a force whereas it is used to describe the peculiarities of a frequency distribution in statistics. We can measure the central tendency of a set of observations by using moments. Moments also help in measuring the scatteredness, asymmetry and peakedness of a curve for a particular distribution. Moments refers to the average of the deviations from mean or some other value raised to a certain power. The arithmetic mean of various powers of these deviations in any distribution is called the moments of the distribution about mean. Moments about mean are generally used in statistics.

Moments for ungrouped data:

Now we first define the moments for ungrouped data. The r[th] moment about origin is denoted by μ'_r and defined by,

$$\mu'_r = \frac{1}{n}\sum_{i=1}^{n} x_i^r \,, \quad r = 1, 2, 3 \ldots \qquad (1)$$

Here the μ'_r is the r[th] moment when we are dealing with the n observations denoted by x_1, $x_2 \ldots x_n$. Thus for r =1, 2, 3 and 4 we get the first four raw moments about the origin.

$$\mu'_1 = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \mu'_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2, \quad \mu'_3 = \frac{1}{n}\sum_{i=1}^{n} x_i^3 \quad \text{and} \quad \mu'_4 = \frac{1}{n}\sum_{i=1}^{n} x_i^4.$$

Similarly we can define the r[th] moment about the arithmetic mean $\overline{x}$ or this is also called the r [th] central moment and it is denoted by the notation μ_r and it is defined as:

$$\mu_r = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^r, \quad r = 1, 2, 3 \ldots \qquad (2)$$

Thus for r =1, we get the first central moment about the mean as $\mu_1 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}) = 0.$

Similarly for r = 2, we get the second central moment about the mean as $\mu_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$ which is equal to variance.

Moments for grouped data:

Suppose we are having observations $x_1, x_2, \ldots, x_n$ which are the mid points of the class-intervals and $f_1, f_2, \ldots, f_n$ are their corresponding frequencies then the r[th] moment about origin is denoted by μ'_r and defined by,

$$\mu'_r = \frac{1}{N}\sum_{i=1}^{n} f_i x_i^r \,, \quad r = 1, 2, 3 \ldots \text{ and } \quad N = \sum_{i=1}^{n} f_i \qquad (3)$$

Similarly the r[th] moment about arithmetic mean is denoted by μ_r and defined by,

$$\mu_r = \frac{1}{N}\sum_{i=1}^{n} f_i\,(x_i - \overline{x})^r, \quad r = 1, 2, 3 \ldots \qquad (4)$$

Also, the r[th] moment about any point A is denoted by μ'_r and defined by,

$$\mu'_r = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - A)^r, \quad r = 1, 2, 3 \ldots \qquad (5)$$

Note: If $d_i = \frac{(x_i - A)}{h}$ or $d_i = \frac{(x_i - \bar{x})}{h}$, Then r^{th} order moments about an arbitrary point A and mean $\bar{x}$ are defined respectively by $\mu'_r = \frac{1}{N}\sum_{i=1}^{n} f_i {d_i}^r h^r$ & $\mu_r = \frac{1}{N}\sum_{i=1}^{n} f_i {d_i}^r h^r$ $r = 1, 2, 3 \ldots$

Relation between raw (Moments about origin or any point) and Central Moments

The central moments can be expressed in terms of raw moments and vice-versa. The general relation between the moments about mean in terms of moments about any point is given by,

$$\mu_r = \mu'_r - {}^rC_1\ \mu'_{r-1}\ \mu'_1 + {}^rC_2\ \mu'_{r-2}\ {\mu'_1}^2 - \ldots + (-1)^r {\mu'_1}^r, \quad r = 1, 2, 3 \ldots \qquad (6)$$

In particular, on putting r = 2, 3 and 4 in equation (6), we get

$$\mu_2 = \mu'_2 - {\mu'_1}^2,\ \mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2{\mu'_1}^3 \text{ and } \mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2{\mu'_1}^2 - 3{\mu'_1}^4.$$

Conversely,

$$\mu'_r = \mu_r + {}^rC_1\ \mu_{r-1}\ \mu'_1 + {}^rC_2\ \mu_{r-2}\ {\mu'_1}^2 - \ldots + {\mu'_1}^r, \qquad r = 1, 2, 3 \ldots \qquad (7)$$

In particular, on putting r = 2, 3 and 4 in equation (7), we get

$$\mu'_2 = \mu_2 - {\mu'_1}^2,\ \mu'_3 = \mu_3 - 3\mu_2\mu'_1 + {\mu'_1}^3 \text{ and } \mu'_4 = \mu_4 - 4\mu_3\mu'_1 + 6\mu_2{\mu'_1}^2 + {\mu'_1}^4.$$

Example 1: The first four moments of a distribution about the value 4 of the variables are -1.5, 17, -30 and 108. Find the moments about the mean.

Solution: Given A = 4, μ'_1=-1.5, μ'_2=17, μ'_3=-30 and μ'_4=108.

<u>Moments about mean:</u>

$$\mu_2 = \mu'_2 - {\mu'_1}^2 = 17 - (-1.5)^2 = 14.75$$

$$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2{\mu'_1}^3 = -30 - 3(17)(-1.5) + 2\,(-1.5)^2 = 39.75$$

$$\mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2{\mu'_1}^2 - 3{\mu'_1}^4 = 108 - 4(-30)(-1.5) + 6(17)(-1.5)^2 - 3(-1.5)^4 = 142.3125.$$

Example 2: Calculate the first four moments of the following distribution about the mean.

0	1	2	3	4	5	6	7	8
1	8	28	56	70	56	28	8	1

Solution:

x	**f**	**d = (x - $\bar{x}$)**	**fd**	**fd^2**	**fd^3**	**fd^4**
0	1	-4	-4	16	-64	256
1	8	-3	-24	72	-216	648
2	28	-2	-56	112	-224	448
3	56	-1	-56	56	-56	56
4	70	0	0	0	0	0
5	56	1	56	56	56	56
6	28	2	56	112	224	448
7	8	3	24	72	216	648
8	1	4	4	16	64	256
			$\sum = 0$	**$\sum = 512$**	**$\sum = 0$**	**$\sum = 2186$**

Moments about the mean $\bar{x} = 4$ are

$$\mu_1 = \frac{\sum fd}{N} = 0, \;\; \mu_2 = \frac{\sum fd^2}{N} = 2, \;\; \mu_3 = \frac{\sum fd^3}{N} = 0, \;\; \mu_4 = \frac{\sum fd^4}{N} = 11$$

Example 3: Wages of workers are given in the following table:

1.5 - 2.5	2.5 - 3.5	3.5 - 4.5	4.5 - 5.5	5.5 - 6.5
1	3	7	3	3

Calculate the first four central moments of the following distribution.

Wages	**f**	**Mid-point x**	**d = (x - $\bar{x}$)**	**fd**	**fd^2**	**fd^3**	**fd^4**
1.5 - 2.5	1	2	-2	-2	4	-8	16
2.5 - 3.5	3	3	-1	-3	3	-3	3
3.5 - 4.5	7	4	0	0	0	0	0
4.5 - 5.5	3	5	1	3	3	3	3
5.5 - 6.5	1	6	2	2	4	8	16
				$\sum = 0$	**$\sum = 14$**	**$\sum = 0$**	**$\sum = 38$**

<u>Moments about mean:</u>

$$\mu_1 = \frac{\sum fd}{N} = 0, \;\; \mu_2' = \frac{\sum fd^2}{N} = 0.933, \;\; \mu_3 = \frac{\sum fd^3}{N} = 0, \; \mu_4 = \frac{\sum fd^4}{N} = 2.533$$

Skewness and Kurtosis:

Averages tell us about the central value of the distribution and measures of dispersion tell us about the concentration of the items around a central value. These measures do not reveal whether the dispersal of value on either side of an average is symmetrical or not. If observations are arranged in a symmetrical manner around a measure of central tendency, we get a symmetrical distribution; otherwise, it may be arranged in an asymmetrical order which gives asymmetrical distribution.

Measures of Skewness and Kurtosis, like measures of central tendency and dispersion, study the characteristics of a frequency distribution. Thus, skewness is a measure that studies the degree and direction of departure from symmetry.

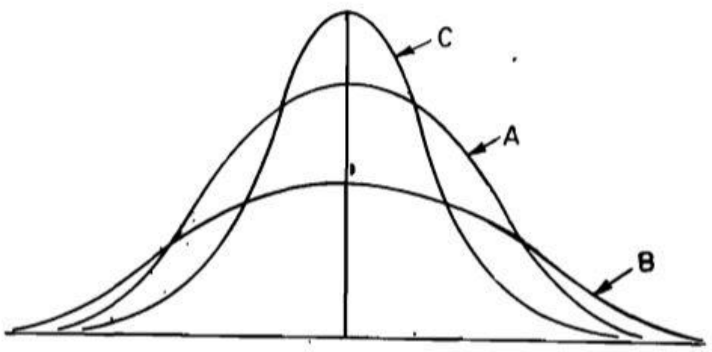
A symmetrical distribution, gives a 'symmetrical curve', where the value of mean, median and mode are exactly equal. On the other hand, in an asymmetrical distribution, the values of mean, median and mode are not equal. When two or more symmetrical distributions are compared, the difference in them is studied with 'Kurtosis'. On the other hand, when two or more symmetrical distributions are compared, they will give different degrees of Skewness. These measures are mutually exclusive i.e. the presence of skewness implies absence of kurtosis and vice-versa.

Measures of Kurtosis:

Kurtosis enables us to have an idea about the flatness or peakedness of the curve. It is measured by the Karl Pearson co-efficient β_2 and given by

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

Kurtosis studies the concentration of the items at the central part of a series. The following figure in which all the three curves A, B and C are symmetrical about the mean.



Curve of the type 'A' which is neither flat nor peaked is called the normal curve or 'MESOKURTIC' curve ($\beta_2 = 3$). If items concentrate too much at the center (more peaked than the normal curve), the curve of the type 'C' becomes 'LEPTOKURTIC' curve ($\beta_2 > 3$).

If the concentration at the center is comparatively less (flatter than the normal curve), the curve of the type 'B' becomes 'PLATYKURTIC' curve ($\beta_2 < 3$).

Measures of Skewness:

Literally, skewness means 'lack of symmetry'. A distribution is said to be skewed if

(i) Mean, Median and Mode fall at different points.

(ii) The curve drawn with the help of the given data is not symmetrical but stretched more to one side than to the other.

Karl Pearson's coefficient of Skewness: The method is most frequently used for measuring skewness. The formula for measuring coefficient of skewness is as follows:

$$S_k = \frac{\text{Mean - Mode}}{\sigma}$$, where σ is the standard deviation of the distribution.

Based upon moments, co-efficient of skewness is defined as follows:

$$S_k = \frac{\sqrt{\beta_1}\,(\beta_2+3)}{2(5\beta_2-6\beta_1-9)},\ \text{where } \beta_1 = \frac{\mu_3^2}{\mu_2^3} \text{ and } \beta_2 = \frac{\mu_4}{\mu_2^2}.$$

Nature of Skewness:

Skewness can be positive or negative or zero. The direction of skewness is determined by observing whether the mean is greater than the mode (positive skewness) or less than the mode (negative skewness).

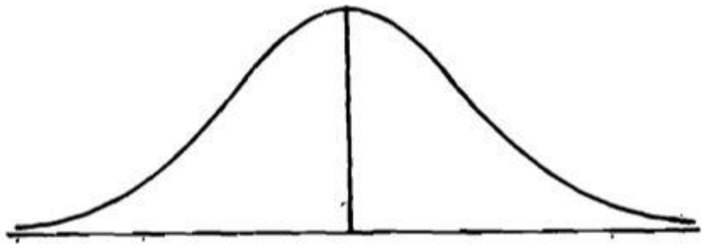
(i) When the values of mean, median and mode are equal, there is no skewness.

(ii) When mean > median > mode, skewness will be positive.

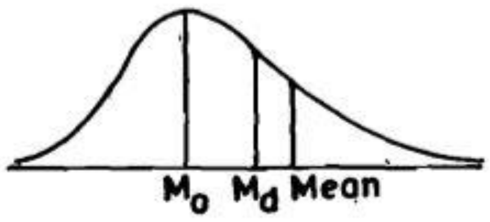
(iii) When mean < median < mode, skewness will be negative.

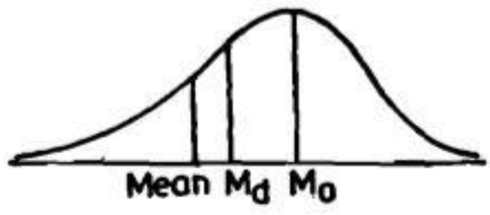
Characteristic of a good measure of skewness:

1. It should be a pure number in the sense that its value should be independent of the unit of the series and also degree of variation in the series.
2. It should have zero-value, when the distribution is symmetrical.
3. It should have a meaningful scale of measurement so that we could easily interpret the measured value.



$\bar{x}$ (*Mean*) = M_0 = M_d
(Symmetrical Distribution)





(Positively Skewed Distribution) (Negatively Skewed Distribution)

Example: Wages of workers are given in the following table:

10-12	12-14	14-16	16-18	18-20	20 - 22	22 - 24
1	3	7	12	12	4	3

Calculate the first four central moments of the following distribution. Also compute β_1 and β_2.

Wages	f	Mid-point x	d = (x -17) / 2	fd	fd^2	fd^3	fd^4
10-12	1	11	-3	-3	9	-27	81
12-14	3	13	-2	-6	12	-24	48
14-16	7	15	-1	-7	7	-7	7
16-18	12	17	0	0	0	0	0
18-20	12	19	1	12	12	12	12
20-22	4	21	2	8	16	32	64
22-24	3	23	3	9	27	81	243
				$\sum = 13$	$\sum = 27$	$\sum = 67$	$\sum = 455$

$$\mu_1' = \frac{\sum fd}{N} \times h = 0.52,\quad \mu_2' = \frac{\sum fd^2}{N} \times h^2 = 2.16,\quad \mu_3' = \frac{\sum fd^3}{N} \times h^3 = 10.72,$$

$$\mu_4' = \frac{\sum fd^4}{N} \times h^4 = 145.6$$

<u>Moments about mean:</u>

$$\mu_1 = 0, \mu_2 = \mu_2' - \mu_1'^2 = 2.16 - 0.2704 = 1.8896$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = 10.72 - 3(2.16)(0.52) + 2\,(0.52)^2 = 7.491$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4 = 145.6 - 4(0.52)(10.72) + 6(2.56)(0.52)^2 - 3 \times 0.07312$$
$$= 126.5874.$$

So, we have $\beta_1 = \dfrac{\mu_3^2}{\mu_2^3} = 8.317,\ \beta_2 = \dfrac{\mu_4}{\mu_2^2} = 35.4527.$

Exercise:

1. The first four raw moments of a distribution are 2, 136, 320 and 40,000. Find the coefficients of skewness and kurtosis.

Ans.

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0.0904, \ \ \beta_2 = \frac{\mu_4}{\mu_2^2} = 2.333.$$

2. Find the second, third and fourth central moments of the frequency distribution given below. Hence, find (i) a measure of skewness and (ii) a measure of kurtosis.

Class limits	Frequency
110.0 – 114.9	5
115.0 – 119.9	15
120.0 – 124.9	20
125.0 – 129.9	35
130.0 – 134.9	10
135.0 – 134.9	10
140.0 – 144.9	5

Ans.

$$\mu_2 = 2.16,\ \mu_3 = 0.804,\ \mu_4 = 12.5232$$

$$\gamma_1 = \sqrt{\beta_1} = 0.25298;\ \gamma_2 = \beta_2 - 3 = -0.317$$

3. Find the second, third and fourth central moments of the frequency distribution given below. Hence, find (i) a measure of skewness and (ii) a measure of kurtosis.

5	10	15	20	25	30	35
4	10	20	36	16	12	2

Ans.

$$\mu_2 = 44.41,\ \mu_3 = -12.504,\ \mu_4 = 5423.5057,\ \beta_1 = 0.001785,$$

$$\beta_2 = 2.7499,\ \gamma_1 = \sqrt{\beta_1} = 0.25298;\ \gamma_2 = \beta_2 - 3 = -0.317.$$

4. Compute the first four moments about mean from the following data. Hence, find (i) a measure of skewness and (ii) a measure of kurtosis.

Class Intervals:	0 -10	10 – 20	20 – 30	30 – 40
Frequency:	1	3	4	2

Ans.

$$\mu_1 = 0,\ \mu_2 = 81,\ \mu_3 = -144,\ \mu_4 = 14817,\ \beta_1 = 0.03902,$$

$$\beta_2 = 0.01909,\ \gamma_1 = \sqrt{\beta_1} = 0.1975;\ \gamma_2 = \beta_2 - 3 = -2.9809.$$

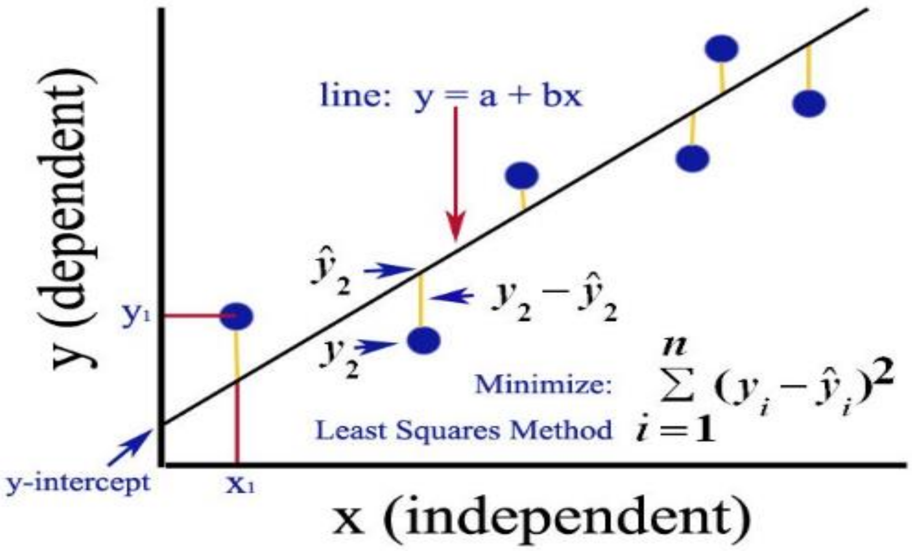
Method of Least squares:

Suppose we are given n values of $x_1, x_2, x_3, \ldots, x_n$ of an independent variable x and the corresponding values $y_1, y_2, y_3, \ldots, y_n$ of a variable y depending on x. Then the pairs (x_1,

y_1), (x_2, y_2),........, (x_n, y_n) give us n- points in the xy-plane. Generally it is not possible to find the actual curve y = f(x) that passes through these points. Hence we try to find a curve that serves as best approximation to the curve y = f(x). Such a curve is referred to as the curve of best fit. The process of determining a curve of best fit is called curve fitting. A method to find curve of best fit is called method of least squares.

The method of least squares tells that the curve should pass as closely as possible to meet all the points. Let y = f(x) be an approximate relation that fits into the data (x_i, y_i), y_i are called observed values and $Y_i = f(x_i)$ are called the expected values. Then $E_i = y_i - Y_i$ are called the estimated error or residuals.

The method of least squares provides a relationship y = f(x) such that sum of the squares of the residues is least. Such a curve is known as least square curve.



Fitting of polynomial:

Approximating a data set using a polynomial equation is useful when conducting engineering calculations as it allows results to be quickly updated when inputs change without the need for manual lookup of the dataset. The most common method to generate a polynomial equation from a given data set is the least squares method. We will discuss the fitting of the following types of the curves.

<u>Fitting of a straight line:</u> $y = a + bx$

Let $y = a + bx$ be the equation of the straight line.

The error estimate is given by $E = y - (a + bx) = y - a - bx$.

By the principle of least squares we have to determine the constants a, b such that

$$E = \sum_{1}^{n} (y - a - bx)^2 \quad \text{is minimum.}$$

For E to be minimum the two necessary conditions are

$$\frac{\partial E}{\partial a} = 0, \qquad \frac{\partial E}{\partial b} = 0,$$

i.e, $\frac{\partial E}{\partial a} = 0 \Rightarrow 2\sum_{1}^{n}(y - a - bx)(-1) = 0,$

$$\Rightarrow 2\sum_{1}^{n}(y - a - bx) = 0,$$

$$\Rightarrow \sum y - \sum a - b\sum x = 0,$$

$$\Rightarrow \sum y = na + b\sum x,$$

$$\frac{\partial E}{\partial b} = 0 \Rightarrow 2\sum_{1}^{n}(y - a - bx)(-x) = 0,$$

$$\Rightarrow \sum xy = a\sum x + b\sum x^2 .$$

The normal equations for estimating the values of a and b are

$$\sum y = na + b\sum x,$$

$$\sum xy = a\sum x + b\sum x^2 .$$

Solving the above normal equations we estimate the values of a & b. With these values of a and b $y = a + bx$ is the line of best fit.

Fitting of a second degree equation (quadratic): $y = a + bx + cx^2$

Let $y = a + bx + cx^2$ be the equation of the curve.

The error estimate is given by $E = y - a - bx - cx^2$.

By the principle of least squares we have to determine the constants a, b and c such that

$$E = \sum_{1}^{n} (y - a - bx - cx^2)^2 \quad \text{is minimum.}$$

For E to be minimum $\frac{\partial E}{\partial a} = 0, \frac{\partial E}{\partial b} = 0, \frac{\partial E}{\partial c} = 0,$

$$\frac{\partial E}{\partial a} = 0 \Rightarrow 2\sum_{1}^{n}(y - a - bx - cx^2)(-1) = 0,$$

$$\Rightarrow \sum y - \sum a - b\sum x - c\sum x^2 = 0,$$

$$\Rightarrow \sum y = na + b\sum x + c\sum x^2,$$

$$\frac{\partial E}{\partial b}=0 \Rightarrow 2\sum_{1}^{n}(y-a-bx-cx^2)(-x)=0,$$

$$\Rightarrow \sum xy = a\sum x + b\sum x^2 + c\sum x^3,$$

$$\frac{\partial E}{\partial c}=0 \Rightarrow 2\sum_{1}^{n}(y-a-bx-cx^2)(-x^2)=0,$$

$$\Rightarrow \sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4.$$

The normal equations for estimating the values of a , b ,c are

$$\sum y = na + b\sum x + c\sum x^2,$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3,$$

$$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4.$$

Solving the above equations we estimate the values of a ,b & c. With these values of a , b & c, $y = a + bx + cx^2$ is the curve of best fit.

Fitting of a curve of the form: $y = ae^{bx}$

Let $y = ae^{bx}$ be the equation of the given curve.

Taking log on both sides we get, logy = loga + $\log e^{bx}$,

$\Rightarrow u = A + bx$, where $A = \log a$ & $u = \log y$.

This is linear in u and x.

Then the normal equations for estimating the values of A and b are

$$\sum u = nA + b\sum x,$$

$$\sum xu = A\sum x + b\sum x^2.$$

By solving these equations, we get the values of A and b .

But $A = \log a \Rightarrow a = \text{antilog} A$.

With these values of a and b, $y = ae^{bx}$ is the curve of best fit.

Fitting of a curve of the form: $y = ax^b$

Let $y = ax^b$.

Taking log on both sides we get

logy = loga + blogx,

$Y = A + bX$ where $Y = \log y, A = \log a, X = \log x$.

The normal equations are

$$\sum Y = nA + b\sum X,$$

$$\sum XY = A\sum X + b\sum X^2 .$$

Solving the above equations we estimate the values of a & b. With these values of a and b, $y = ax^b$ is the curve of best fit.

Examples:

1. Fit a straight line to the following data.

x	1	6	11	16	20	26
y	13	16	17	23	24	31

Let y = a + b x be the straight line.
The normal equations for estimating the values of a and b are

$$\sum y = na + b\sum x,$$

$$\sum xy = a\sum x + b\sum x^2.$$

Given n = 6

x	**y**	**x^2**	**xy**
1	13	1	13
6	16	36	96
11	17	121	187
16	23	256	368
20	24	400	480
26	31	676	806
$\sum x = 80$	$\sum y = 124$	$\sum^{x2} = 1490$	$\sum xy = 1950$

Substituting the above values in the normal equations we get

6a + 80b = 124
80a + 1490 = 1950

Solving, we get a = 11.3227, b = 0.7008.

Therefore the equation of best fit is y = 11.3227 +0.7008x

2. Fit a straight line to the following data.

x	1	2	3	4	5	6
y	6	4	3	5	4	2

Soln:
Let y = a + b x be the straight line.

The normal equations for estimating the values of a and b are

$$\sum y = na + b\sum x, \quad \sum xy = a\sum x + b\sum x^2.$$

Here n = 6 and following the procedure as in example 1 we get

$\sum x = 21,\ \sum y = 24,\ \sum xy = 75,\ \sum x^2 = 91.$

Therefore, we get 24 = 6a + 21b, 75 = 21a + 91b.

Solving, we get a = 5.799, b = – 0.514.

Therefore the equation of best fit is y = 5.799 – 0.514x.

3. Fit a straight line of the form y = ax +b for the following data by the method of least squares.

x	5	10	15	20	25
y	16	19	23	26	30

Soln:

Let y = ax + b be the given straight line

The normal equations are $\sum y = a\sum x + nb,\ \sum xy = a\sum x^2 + b\sum x.$

Here n = 5 and following the procedure as in example 1 we get

$\sum y = 114, \sum x = 75, \sum xy = 1885, \sum x^2 = 1375,$

Substituting in the above equations we get a = 0.7, b = 12.3.

The best fit is y = 0.7 x + 12.3.

4. Fit an exponential curve of the type $y = a e^{bx}$ from the following data by the method of least squares.

x	1	2	4
y	5	10	30

Let $y = a e^{bx}$ ………….(1) be the required curve.
Taking log on both side of (1) and simplifying we get

Y = A + b x, where A = log a, Y = log y

The normal equations for estimating the values of a and b are

$$\sum Y = nA + b\sum x \text{ and } \sum xY = A\sum x + b\sum x^2$$

x	y	Y = log y	xY	x^2
1	5	0.6990	0.6990	1
2	10	1.0000	2.0000	4
4	30	1.4771	5.9085	16

$\sum = 7$		$\sum = 3.1761$	$\sum = 8.6095$	$\sum = 21$

Substituting the above values in the normal equations we get

$$3A + 7b = 3.1761$$
$$7A + 21b = 8.6095$$

Solving, we get A = 0.4604 but a = antilog (0.4604) = 2.8867, b = 0.2564.

Therefore the equation curve of best fit is $y = 2.8867\, e^{0.5624\, x}$.

5. Fit a curve of the form $y = a\, e^{bx}$ to the data by the method of least squares.

x	0	2	4
y	8.12	10	31.82

Soln:

Let $y = a\, e^{bx}$ ....(1) be the required curve.
Taking log on both side of (1) and simplifying we get

Y = A + b x, where A = log a, Y = log y

The normal equations for estimating the values of a and b are

$$\sum Y = nA + b\sum x \text{ and } \sum xY = A\sum x + b\sum x^2$$

Here n = 3 and following the procedure as in example 4 we get

$\sum x = 6$, $\sum Y = 7.85$, $\sum xY = 18.44$, $\sum x^2 = 20$.

Substituting the above values in the normal equations we get

$$3A + 6b = 7.85$$
$$6A + 20b = 18.44$$

By solving these equations, we get A = 1.932 but a = antilog (A) = 6.903, b = 0.3425.

Therefore $y = 6.903\, e^{0.3425x}$ is the curve of best fit.

6. Fit a curve of the form $y = a\, b^x$ ....(1) to the data by the method of least squares.

x	2	4	6	8	10
y	1	3	6	12	24

Soln:
Let $y = a\, b^x$ ....(1) be the required curve.
Taking log on both side of (1) and simplifying we get

Y = A + B x, where A = log a, B = log b and Y = log y

The normal equations for estimating the values of a and b are

$$\sum Y = nA + B\sum x \text{ and } \sum xY = A\sum x + B\sum x^2$$

Here n = 5 and following the procedure as in example 4 we get

$\sum x = 30$, $\sum Y = 3.7147$, $\sum xY = 29.0130$, $\sum x^2 = 220$.

Substituting the above values in the normal equations we get

$$5A + 30B = 3.7147, \quad 30A + 220B = 29.0130$$

By solving these equations, we get A = -0.26566 but a = antilog (A) = 1.8436,

B = 0.1681 but b = antilog (B) = 1.4727.

Therefore $y = (1.8436)(1.4727)^x$ is the curve of best fit.

7. At constant temperature the pressure P and the volume V of a gas are connected by the relation $PV^{\gamma} = K$ (constant). Find the best fitting equation of this form to the following data and estimate V when P = 4.

P	0.5	1.0	1.5	2.0	2.5	3.0
V	1620	1000	750	620	520	460

Soln: Let $PV^{\gamma} = K$. …….. (1) be the given relation. Taking log on both side of (1) and simplifying we get

$$\sum \log V = 39.73, \sum \log P = 2.42,$$

$$\sum \log V \log P = 14.4786, \ \sum (\log V)^2 = 264.1689.$$

Here n = 3 and following the procedure as in example 4 we get

γ = 1.42 and K = 18144

Therefore $PV^{1.42} = 18144$ is the curve of best fit.

At P = 4, V = 375.9428 ≈ 376.

8. Fit a second degree parabola for the following data.

x	0	1	2	3	4
y	1	3	4	5	6

Soln: Let $y = a + bx + cx^2$ be the second degree polynomial and we have to determine a, b and c.

Normal equations for the second degree parabola are

$$\sum y = na + b\sum x + c\sum x^2,$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3,$$

$$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4.$$

x	**y**	**xy**	**x^2**	**x^2y**	**x^3**	**x^4**
0	1	0	0	0	0	0
1	3	3	1	3	1	1
2	4	8	4	16	8	16
3	5	15	9	45	27	81
4	6	24	16	96	64	256
$\sum x = 10$	$\sum y = 19$	$\sum xy = 50$	$\sum x^2 = 30$	$\sum x^2 y = 160$	$\sum x^3 = 100$	$\sum x^4 = 354$

Substituting the above values in the normal equations and solving we get a = 1.114, b = 1.7717, c = 0.1429.

Therefore the second degree of parabola of best fit is $y = 1.114 + 1.7717\,x - 0.1429x^2$

9. Fit a curve of the form $y = a + bx + cx^2$ to the data by the method of least squares.

x	0	1	2	3	4
y	1	1.8	1.3	2.5	6.3

Soln: Let $y = a + bx + cx^2$ be the second degree parabola and we have to determine a, b and c.

Normal equations for the second degree parabola are

$$\sum y = na + b\sum x + c\sum x^2,$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3,$$

$$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4.$$

Here n = 5 and following the procedure as in example 7 we get

$$\sum x = 10, \sum y = 12.9, \sum xy = 38.1, \sum x^2 = 30, \sum x^3 = 100, \sum x^4 = 354, \sum x^2 y = 131.3.$$

Substitute these values in normal equations

$$\sum y = na + b\sum x + c\sum x^2,$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3,$$

$$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4.$$

Solving we get a = 0.7914, b = - 0.1128, c = 0.3357.

Then the curve of best fit is $y = 0.7914 - 0.1128\,x + 0.3357\,x^2$.

10. The following table gives the production (in thousand units) of a certain commodity in different years:

Year(x)	1968	1978	1988	1998	2008
Production(y)	8	10	12	10	16

Fit a straight line to the data and estimate the production in the year 2015.

Soln:

For convenience in computations, let us set X = x -1967 and Let y = a + b X be the straight line.
The normal equations for estimating the values of a and b are

$$\sum y = na + b\sum X, \quad \sum Xy = a\sum X + b\sum X^2.$$

Here n = 6 and following the procedure as in example 1 we get

$$\sum y_i = 56, \sum X_i = 105, \sum X_i y_i = 1336, \sum x_i^{\,2} = 3025,$$

Substitute these values in normal equations we get

56 = 5a + 10b,

1336 = 105a + 3205b.

Solving these equations, we get a = 7.84 and b = 0.16. Therefore the line of best fit is given by y = a +b X = 7.84 + 0.16X = 7.84 + 0.16 (x -1967) = 0.16x - 306.88.

For x = 2015, this gives y = 15.52.

Thus, for the year 2015, the estimated production is 15.52(thousand units).

Exercise:

1. An experiment gave the following data:

x	1	3	4	6	8	9	11	14
y	1	2	4	4	5	7	8	9

It is known that x and y are connected by the relation y = $a_0 + a_1 x$. Find the best values of a and b using least square method.

Ans. a_0 = 1, a_1= 0.5420 and y = 1 + 0.5420 x

2. The number y of bacteria per unit volume present in a culture after x hours is given by the following table :

x	0	1	2	3	4	5	6

y	32	47	65	92	132	190	275

Fit a curve of the form $y = a\, b^{x}$ to the data. Estimate the value of y when x = 7.

Ans. a = 32.14, b= 1.4270 and $y = 32.14\,(1.4270)^{x}$, $y_7 = 387$.

3. The following table gives the production (in thousands units) of a certain commodity in different years:

Year (x)	1941	1951	1961	1971	1981	1991	2001
Production (y)	3.9	5.3	7.3	9.6	12.9	17.1	23.2

Fit a curve of the form $y = a\, b^{x}$ to this data and estimate the production in the year 2006.

Ans. a = 9.5735, b = 1.3433 and $y = 9.5735\,(1.3433)^{x}$, $y_{2006} = 27.5$ x1000 quintals

4. The latent heat of vaporization of steam r is given in the following table at different temperatures t: For this range of temperature fit a relation of the form r = a + b t using the method of least squares.

t	40	50	60	70	80	90	100	110
r	1069.1	1063.6	1058.2	1052.7	1049.3	1041.8	1036.3	1030.8

Ans. a = 1090.26, b = -0.534 and r = 1090.26 – 0.534 t.

5. The following table gives the results of the measurements of train resistances; V is the velocity in mile per hour and R is the resistance in pound per ton.

V	20	30	40	50	60	70
R	54	90	138	206	292	396

If R is related to V by the relation $R = a + bV + cV^2$. Find a, b and c by the method of least squares and estimate R when V = 45 miles / hour.

Ans. a = 41.77, b = -1.096 and c = 0.08786 $R = 41.77 + (-1.096)\,V + 0.08786\,V^2$,

R = 170 Pound when V = 45 miles / hour.

<u>Correlation and Regression</u>:

The word correlation is used in everyday life to denote some form of association. In statistical terms we use correlation to denote association between two quantitative variables. We also assume that the association is linear, that one variable increases or decreases a fixed amount for a unit increase or decrease in the other. The other technique that is often used in these circumstances is regression, which involves estimating the best straight line to summarize the association.

Correlation:

Correlation means simply a relation between two or more variables.

Two variables are said to be correlated if the change in one variable results in a corresponding change in the other.

Ex: 1. x: supply   y: price

2. x: demand   y: Price.

Positive correlation:

If **an** increase or decrease in one variable corresponds to an increase or decrease in the other then the correlation is said to be positive correlation or direct correlation.

Ex: 1. Demand and price of commodity.   2. Income and expenditure.

Negative correlation:

If an increase or decrease in one variable corresponds to an decrease or increase in the other then the correlation is said to be negative correlation or inversely correlated.

Ex: 1.Supply and Price of a commodity.

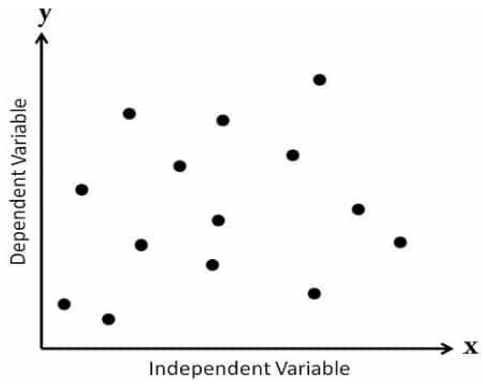
2. Correlation between Volume and pressure of a perfect gas.

No correlation

If there exist no relationship between two variables then they are said to be non correlated.

Scatter diagram

To obtain a measure of relationship between two variables x and y we plot their corresponding values in the xy - plane. The resulting diagram showing the collection of the dots is called the dot diagram or scatter diagram.



Correlation Coefficient (Karl Pearson correlation coefficient)

The degree of association is measured by a correlation coefficient, denoted by r. It is sometimes called Karl Pearson's correlation coefficient and is a measure of linear association.

If a curved line is needed to express the relationship, other and more complicated measures of the correlation must be used.

Let $x_1, x_2, x_3, \ldots\ldots, x_n$ be n values of x and $y_1, y_2, y_3, \ldots\ldots y_n$ be the corresponding n values of y, then the coefficient of correlation between x and y is

$r = \dfrac{\Sigma(x-\bar{x})(y-\bar{y})}{n\sigma_x\sigma_y}$, where $\sigma_x^{\,2}$ - variance of the x series, $\sigma_y^{\,2}$ - variance of the y series,

$\bar{x} = \dfrac{\Sigma x}{n} \rightarrow$ Mean of the x series $\quad \bar{y} = \dfrac{\Sigma y}{n} \rightarrow$ mean of the y series.

For computation purpose we can use the formula

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

Limits for correlation coefficient

The coefficient of correlation numerically does not exceed unity ($-1 \le r \le 1$).

Proof:

We have $r = \dfrac{\frac{1}{n}\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\Sigma(x_i-\bar{x})^2}\ \sqrt{\frac{1}{n}\Sigma(y_i-\bar{y})^2}}$, i=1,2,………n,

$$r = \frac{\frac{1}{n}\Sigma a_i \Sigma b_i}{\sqrt{\frac{1}{n}\Sigma a_i^{\,2}}\ \sqrt{\frac{1}{n}\Sigma b_i^{\,2}}},\quad r^2 = \frac{(\Sigma a_i\ \Sigma b_i)^2}{\Sigma a_i^{\,2}\ \ \Sigma b_i^{\,2}}. \qquad (1)$$

By Schwartz inequality, which states that if a_i, b_i i=1, 2… n are real quantities then

$(\Sigma a_i\ \Sigma b_i)^2 \le \Sigma a_i^{\,2}\ \ \Sigma b_i^{\,2}$ and the sign of equality holding if and only if

$$\frac{a_1}{b_1} = \frac{a_2}{b_2} = \frac{a_3}{b_3} = \ldots\ldots\ldots\ldots = \frac{a_n}{b_n}.$$

Using this equation (1) becomes $r^2 \le 1$,

$$\Rightarrow |r| \le 1,$$
$$\Rightarrow -1 \le |r| \le 1.$$

Hence correlation coefficient cannot exceed unity numerically.

Note:

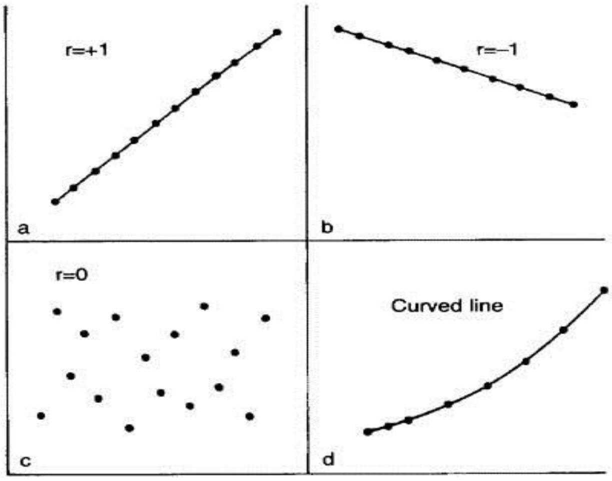


Figure 1.1 Correlation illustrated.

1. If r =-1 there is a perfect negative correlation.
2. If r =1 there is a perfect positive correlation.
3. If r =0 then the variables are non-correlated.
4. When r = 0, $\theta = \frac{\pi}{2}$. i.e, when the variables are independent the two lines of regression are perpendicular to each other.
5. When $r = \pm 1, \theta = 0$ or π. i,e the lines of regression coincide.

Examples:

1. If r is the correlation coefficient between x and y and z= ax+by. Show that

$$r = \frac{\sigma_z^2 - (a^2\sigma_x^2 + b^2\sigma_y^2)}{2ab\sigma_x\sigma_y}.$$

Soln:

Let z = ax + by $\Rightarrow \frac{1}{n}\sum z = \frac{a}{n}\sum x + \frac{b}{n}\sum y \Rightarrow \bar{z} = a\bar{x} + b\bar{y}$,

$$\frac{1}{n}\sum (z-\bar{z})^2 = a^2\frac{1}{n}\sum (x-\bar{x})^2 + b^2\frac{1}{n}\sum (y-\bar{y})^2 + 2ab\frac{1}{n}\sum (x-\bar{x})(y-\bar{y}),$$

$$\Rightarrow \quad \sigma_z^2 = a^2\sigma_x^2 + b^2\sigma_y^2 + 2abr\sigma_x\sigma_y,$$

$$\Rightarrow r = \frac{\sigma_z^2 - (a^2\sigma_x^2 + b^2\sigma_y^2)}{2ab\sigma_x\sigma_y}.$$

2. While calculating the correlation coefficient between x and y from 25 pairs of observations a person obtained the following values. $\Sigma x_i = 125, \Sigma x_i^2 = 650$, $\Sigma y_i = 100, \Sigma y_i^2 = 460, \Sigma x_i y_i = 508$. It was later discovered that he had copied

down the pairs (8,12) and (6,8) as (6,12) and (8,6) respectively. Obtain the correct value of the correlation coefficient.

Soln:

Correct $\sum x_i = 125, \sum x_i^2 = 650, \sum y_i = 102, \sum y_i^2 = 488, \sum x_i y_i = 532$,

n = 25,

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{\{n\sum x^2 - (\sum x)^2\}\{n\sum y^2 - (\sum y)^2\}}} = 0.51912.$$

3. The following Table gives the age (in years) of 10 married couples. Calculate the coefficient of correlation between these ages.

Age of Husband(x)	23	27	28	29	30	31	33	35	36	39
Age of wife(y)	18	22	23	24	25	26	28	29	30	32

Soln:

Here n=10

We find $\bar{x} = \frac{1}{n}\sum x_i = \frac{311}{10} = 31.1$ $\bar{y} = \frac{1}{n}\sum y_i = \frac{257}{10} = 25.7$.

x_i	$X_i = x_i - \bar{x}$	X_i^2	$Y_i = y_i - \bar{y}$	Y_i^2	$X_i Y_i$
23	-8.1	65.61	-7.7	59.29	62.37
27	-4.1	16.81	-3.7	13.69	15.17
28	-3.1	9.61	-2.7	7.29	8.37
29	-2.1	4.41	-1.7	2.89	3.57
30	-1.1	1.21	-0.7	0.49	0.77
31	-0.1	0.01	0.3	0.09	-0.03
33	1.9	3.61	2.3	5.29	4.37
35	3.9	15.21	3.3	10.89	12.87
36	4.9	24.01	4.3	18.49	21.07
39	7.9	62.41	6.3	39.69	49.77
		$\sum X_i^2 = 202.9$	$\sum Y_i^2 = 158.10$		$\sum X_i Y_i = 178.$

$$r = \frac{\sum X_i Y_i}{\sqrt{\sum X_i^2 \sum Y_i^2}} = 0.9955 \approx 1.$$

i.e, the ages of husbands and wives are almost perfectly correlated.

Regression :

Correlation describes the strength of an association between two variables, and is completely symmetrical, the correlation between A and B is the same as the correlation between B and A. However, if the two variables are related it means that when one changes by a certain amount the other changes on an average by a certain amount. The relationship can be represented by a simple equation called the regression equation. In this context "regression" (the term is a historical anomaly) simply means that the average value of y is a "function" of x, that is, it changes with x.

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

Line of regression:

Line of regression is the line which gives the best estimate to the value of one variable for any specific value of the other variable. So the line of regression is the line of best fit.

Regression line of y on x:

Let regression line of y on x be $y = a + bx$.

The normal equations by the method of least squares is

$\sum y = na + b\sum x$,

$\sum xy = a\sum x + b\sum x^2$,

$\frac{1}{n}\sum y = a + \frac{b}{n}\sum x$.

$\bar{y} = a + b\bar{x}$ is the regression line passing through ($(\bar{x}, \bar{y})$

$$b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{\sum(XY)}{\sum X^2} = \frac{\sum(XY)}{n\sigma_x^{\ 2}} = r\frac{\sigma_y}{\sigma_x},$$

$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \Rightarrow Y = b_{yx}X$ is the regression line of y on x.

Note:

1. Regression coefficient of y on x

$$b_{yx} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{n\sum xy - \sum x\sum y}{n\sum x^2 - (\sum x)^2} = r\frac{\sigma_y}{\sigma_x}.$$

2. Regression coefficient of x on y

$$b_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{n\sum xy - \sum x \sum y}{n\sum y^2 - (\sum y)^2} = r\frac{\sigma_x}{\sigma_y}.$$

Examples:

1. If two regression equations of the variables x and y are x = 19.13 - .87y, y = 11.6 – 0.5x, find

   (a) mean of x

   (b) mean of y

   (c)The correlation coefficient between x and y.

Soln:

Since $\bar{x}$ and $\bar{y}$ lie on two regression lines,

$\bar{x} = 19.13 - 0.87\bar{y},\ \bar{y} = 11.64 - 0.5\bar{x},$

Solving we get $\bar{x} = 15.79,\ \bar{y} = 3.74.$

$$b_{yx} = -0.5, b_{xy} = -0.87, r = \sqrt{-0.5 \times -0.87} = -0.66.$$

2. In the following table data is showing the test scores made by sales man on an intelligent test and their weekly sales.

Test scores(x)	1	2	3	4	5	6	7	8	9	10
sales(y)	2.5	6	4.5	5	4.5	2	5.5	3	4.5	3

Calculate the regression line of sales on test scores and estimate the most possible weekly volume if a sales man scores 70.

Soln:

$\bar{x} = 60, \bar{y} = 4.05$, Regression line of y on x is $y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$,

y = 0.06x + 0.45.

When x = 70, y = 4.65.

3. In a partially destroyed laboratory, record of an analysis of correlation data, the following results only are legible.

   Variance of x=9, Regression equations 8x -10y + 66 = 0, 40x - 18y = 214

   what are (i) the mean values of x and y

   (ii) the correlation coefficient between x and y

(iii) the standard deviation of y.

Soln:

(i) Since both the lines of regression pass through the point $(\bar{x}, \bar{y})$

$8\bar{x} - 10\bar{y} + 66 = 0,$

$40\bar{x} - 18\bar{y} - 214 = 0.$

Solving these equations we get $\bar{x} = 13$, $\bar{y} = 17$

(ii) $\sigma_x{}^2 = 9$

$\sigma_x = 3$

Let $8x - 10y + 66 = 0$ and $40x - 18y = 214$ be the lines of regression of y on x and x on y respectively

$$b_{yx} = \frac{4}{5}, b_{xy} = \frac{18}{40} = \frac{9}{20}, \text{ Hence } r^2 = b_{yx}\, b_{xy} = \frac{9}{25}, \; r = \pm\frac{3}{5} = \pm 0.6.$$

Since both the regression coefficients positive we take $r = 0.6$.

Standard deviation of y = 4.

4. The following table gives the stopping distance y in meters of a motor bike Moving at a speed of x Kms/hour when the breaks are applied

x	16	24	32	40	48	56
y	0.39	0.75	1.23	1.91	2.77	3.81

Find the correlation coefficient between the speed and the stopping distance, and the equations of regression lines. Hence estimate the maximum speed at which the motor bike could be driven if the stopping distance is not to exceed 5 meters.

Soln:

$\bar{x} = 36, \; \bar{y} = 1.81, \; , \sigma_x = 13.663, \sigma_y = 1.1831,$

$b_{yx} = 0.0851, b_{xy} = 11.352,$

$$r = r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{\{n\sum x^2 - (\sum x)^2\}\{n\sum y^2 - (\sum y)^2\}}} = 0.983.$$

The equation of the line of regression of y on x is $y = 0.0851x - 1.2536$ (i)

and the equation of the line of regression of x on y is $x = 11.352y + 15.453$. (ii)

For $y = 5$, equation (ii) gives $x = 72.213$.

Accordingly, for the stopping distance not to exceed 5 meters, the speed must not exceed 72 Kms/hour.

Exercise:

1. If the coefficient of correlation between the variables x and y is 0.5 and the acute angle between their lines of regression is $\tan^{-1}\left(\frac{3}{5}\right)$. Find the ratio of the standard deviation of x and y.

   Ans. $\frac{\sigma_x}{\sigma_y} = \frac{1}{2}$ or $\frac{\sigma_x}{\sigma_y} = \frac{2}{1}$.

2. Prove the following formulas for the coefficient of correlation r (in the usual notation)

   a) $r = 1 - \frac{1}{2n}\Sigma\left(\frac{X_i}{\sigma_x} - \frac{Y_i}{\sigma_y}\right)^2$, $r = -1 + \frac{1}{2n}\Sigma\left(\frac{X_i}{\sigma_x} + \frac{Y_i}{\sigma_y}\right)^2$.

3. The following table shows the ages x and the systolic pressures of 12 persons.

Age (x)	56	42	72	36	63	47	55	49	38	42	68	60
Blood Pressure **(y)**	147	125	160	118	149	128	150	145	115	140	152	155

   Calculate the coefficient of correlation between x and y. Estimate the blood pressure of a person whose age is 45 years.

   Ans. r = 0.8961, y = 80.78 + 1.138 x , when x = 45, y = 132.

4. The height (inches) and weight (pounds) of baseball players are given below:
   (76, 212), (76, 224), (72, 180), (74, 210), (75, 215), (71, 200), (77, 235), (78, 235), (77, 194), (76, 185).

   (i) Estimate the coefficient of correlation between weight and height of baseball players.

   (ii) Find the regression line between weight and height. Use the regression equation to find the weight of a baseball player that is 68 inches tall.

   Ans. r = 0.5529, y = 4.737 x – 147.227, x = 0.064 y + 61.712, when x = 68, y = 97.37.

5. The equations of regression lines of two variables x and y are 4 x – 5y + 33 = 0 and 20x - 9y = 107, Find the correlation coefficient and the means of x and y.

   Ans. r = 0.6, Mean of x = 13 and Mean of y = 17.

6. If the tangent of the angle between the lines of regression of y on x and x on y is 0.6 and the standard deviation of y is twice the standard deviation of x. find the coefficient

of correlation between x and y.

Ans. r = 0.5.

Resources:

1. https://nptel.ac.in/courses/111105042/
2. http://www.nptelvideos.in/2012/12/regression-analysis.html
3. https://nptel.ac.in/courses/111104074/