MEASURES OF CENTRAL TENDENCY

Average are the values around which other items of the distribution congregate. They are the values which lie between the two extreme observations (i.e. smallest and the largest observations), of the distribution and give us an idea about the concentration of the values in the central part of the distribution According they are also sometimes referred to as a Measures of Central Tendency.

Various Measures of Central Tendency: The following are the five measures of Central Tendency:-

1. Mathematical Averages
   a. Arithmetic Averages or Mean
   b. Geometric Mean
   c. Harmonic Mean
2. Averages of Position
   a. Median
   b. Mode

Arithmetic Mean: Arithmetic Average of Mean of a given set of observations is their sum divided by the number of observations.

If x_1, x_2, x_3…… x_n are the given n observations, then their Arithmetic Mean (A.M) usually denoted by $\bar{x}$, is given by

$$\bar{x} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n} = \frac{\sum x}{n} \quad \text{[where } \sum x \text{ = Sum of n observations]}$$

Calculation of arithmetic mean in a discrete series.

Direct method: In a discrete series the values of the variable are multiplied by their respective frequencies and the product so obtained are totaled. The total is divided by the number of items which in a discrete series is equal to the total of the frequencies. Th resulting quotient is a simple arithmetic average of the series.

Algebraically if f_1, f_2, f_3, ……… etc, stand respectively for the frequencies of the values x_1, x_2, x_3 …….. etc.

$$\bar{x} = \frac{1}{n}(x_1 f_1 + x_2,f_2 , + x_3f_3 + …….. + x_nf_n) \Rightarrow \bar{x} = \frac{\sum fx}{N} \text{ or } \frac{\sum fx}{\sum f}$$

Short Cut Method (Step Deviation Method): In this method the deviations of the items from an assumed mean are first found out and they are multiplied by their respective frequencies. The total of these products is divided by the total frequencies and added to the assumed average.

Algebraically $\bar{x} = A + \frac{\sum fdx}{\sum N}$, Where, $\sum fdx$ = the total of the product of the deviations from the assumed mean and their respective frequencies of the items

A = assumed average,

N = Number of items

Calculation of the arithmetic average in a continuous series: The process of the calculation of arithmetic average in a continuous series is the same as in case of discrete series In a continuos series the midpoints of the various class intervals are written down to replace the class intervals. Once it is done there is no difference between a continuous series and a discrete series.

EXAMPLE

Ex.1 What is the mean wage of the data given below:

Wage (in Rs.)	800	820	860	900	920	980	1000
No. of workers	7	14	19	25	20	10	5

Sol. Let the assumed mean be A = 900. The given data can be written as under

Wage (in Rs.) x_i	No. of workers f_1	$d_i = x_1 – A$	$f_i\ d_i$
800	7	–100	–700
820	14	–80	–1120
860	19	–40	–760
900	25	0	0
920	20	20	400
980	10	80	800
1000	5	100	500
	$\sum f_i = 100$		$\sum f_id_i = 880$

Here A = 900. ∴ Mean = $\bar{x} = A + \frac{1}{N}\sum_{i=1}^{n} f_id_i = 900 + \left(-\frac{800}{100}\right) = 891.2.$

Hence, mean wage = Rs. 891.2.

Step Deviation Method:

Lets assumed mean A = 900

Wage (in Rs.) x_i	No. of workers f_1	$d_i = x_1 – A = x_i – 900$	$f_i\ d_i / h$
800	7	–100	–35
820	14	–80	–56
860	19	–40	–38
900	25	0	0
920	20	20	20
980	10	80	40
1000	5	100	25
		$\sum f_i = 100$	$\sum f_id_i / h = – 44$

Here A = 900, h = 20. ∴ Mean $\bar{x}$ = A + h $\left(\frac{1}{N}\sum_{i=1}^{n} f_i u_i\right)$ = 900 + 20$\left(-\frac{44}{100}\right)$ = 891.2

Hence, mean wage = Rs. 891.2.

Weighted Arithmetic Average: In the calculation of simple each then item of the series is considered equally important but there may be cases where all items may not have equal importance and some of them may be comparatively more important than others. In such cases proper weightage is to be given to various items – the weights attached to each item being proportional to the importance of the item in the distribution.

Let $w_1, w_2, w_3, \ldots\ldots. w_n$ be the weights attached to variable values $x_1, x_2, x_3, \ldots\ldots.. x_n$ respectively.

Then the weighted arithmetic mean, usually denoted by $\bar{x}_W = \frac{w_1x_1 + w_2x_2 + w_3x_3 + \ldots.. + w_nx_n}{w_1 + w_2 + w_3 + \ldots.. + w_n}$

= $\frac{\sum wx}{\sum w}$ where, $w_1, w_2, w_3, \ldots\ldots. w_n$ are the respective weights of $x_1, x_2, x_3, \ldots\ldots.. x_n$

In case of frequency distribution, $f_1, f_2, f_3, \ldots.. f_n$ are the frequencies of the variable values $x_1, x_2, x_3, \ldots\ldots.. x_n$ respectively, then the weighted arithmetic is given by

$$\bar{x}_W = \frac{w_1(fx_1) + w_2(fx_2) + w_3(fx_3) + \ldots.. + w_n(fx_n)}{w_1 + w_2 + \ldots.. + w_n}$$

Where, $w_1, w_2, w_3, \ldots\ldots. w_n$ are the respective weights of $x_1, x_2, x_3, \ldots\ldots.. x_n$

CONTINUOUS DISTRIBUTION

Ex. Calculate the arithmetic mean for the following frequency distribution

Class	0 – 8	8 – 16	16 – 24	24 –32	32 – 40	40 – 48
Frequency	8	7	16	24	15	7

Sol. Let the assumed mean be A = 28 and h = 8. The calculations are:

Class	Mid-value	frequency	$u_i = (x_i – A)/h$	$f_i u_i$
	x_i	f_i	$= (x_i – 28)/8$	
0 – 8	4	8	–3	–24
8 –16	12	7	–2	–14
16–24	20	16	–1	–16
24 – 32	28	24	0	0
32 –40	36	15	1	15
40 – 48	44	7	2	14
		N = $\sum f_i$ = 77		$\sum f_iu_i$ = – 25

Here A = 28, h = 8. $\therefore$ Mean $\bar{x} = A + h\left(\frac{1}{N}\sum_{i=1}^{n} f_i u_i\right) = 28 + 8\left(-\frac{25}{77}\right) = 25.404$

Weighted Average

Ex. Find out the weighted arithmetic average wage rate of 31 building trade workers from the following table:

Kind of worker	***Daily wages in Rupees***	***Number employed***
Masons	15	4
Laborers	8	20
Carpenters	12	5
Painters	10	2

Sol.

Kind of workers	Daily wages in Rupees(w)	Number employed(x)	w X x
Masons	15	4	60
Laborers	8	20	160
Carpenters	12	5	60
Painters	10	2	20
		$\sum w = 31$	$\sum wx = 300$

$\therefore$ The required weighted average = $\sum wx / \sum w = 300/31 = 9.68$ rupees

Geometric Mean *The geometric mean is the nth root of the product of n items of a series. Thus if x_1, x_2, x_3, …….. x_n are the given n observations, then their G.M is given by*

$Gm = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 ..... \cdot x_n} = (x_1, x_2, x_3, …….. x_n)^{1/n}$,

Taking logarithm on both sides, we get

$\log(GM) = \frac{1}{n}\log(x_1, x_2, x_3, …….. x_n) = \frac{1}{n}(\log x_1 + \log x_2 + …….+ \log x_n) = \frac{1}{n}\sum \log(x)$

Thus, we see that the logarithm of the G.M. of a set of observations is the arithmetic mean of their logarithms.

Taking Antilog on both sides, we get GM = Antilog $\left[\frac{1}{n}\sum \log(X)\right]$

<u>Calculation of GM in Discrete Series</u> In discrete series the geometric mean or GM = $\left[\frac{1}{N}\sum f \log(X)\right]$,

where, f is the frequency, X is the value of the items and N is the total number of items.

<u>Calculation of GM in Continuous Series:</u> In the calculation of geometric mean in a Continuous Series the procedure is the same as in case of discrete series with the only difference that the mid values of the class intervals are taken as the values of the variable.

Weighted Geometric Mean: Weighted geometric mean is the n^{th} root of the product of various values raised to the power of their respective weights. Thus if x_1, x_2, x_3, …….. x_n stand for the values of a variable, w_1, w_2, w_3, …….. w_n etc. for their respective weights n for the sum of weights.

Weighted Geometric mean of GM(w) is $\quad GM\ (w) = \sqrt[n]{x^{w_1} \cdot x^{w_2} \cdot x^{w_3} ..... \cdot x^{w_n}}$ → GM(w)

$$= \text{Anti log}\left[\frac{1}{N}\sum w \log(X)\right] = \text{Antilog}\ \frac{\sum w \log X}{\sum w}$$

Compound Interest Formula: Let us suppose that P_0 is the initial value of the variable and P_n be the value at the end of the period n and let r be the rate of growth per unit per period. Since r, is the rate of growth per unit per period, growth for period 1 is rP_o and thus the value of the variable ate the end of period is $rP_o + P_o = P_0\ (1 + r)$. For the second period the initial value of the variable becomes $P_0(\ 1 + r)$.

The growth for the second period is $P_o\ (1 + r)r$ and consequently the value of the variate at the end of 2 nd period is $P_0\ (1 + r) + P_0\ (1 + r)r = P_0\ (1 + r)\ (1 + r) = P_0\ (1 + r)^2$.

Similarly, the value of the variate at the end of period n is $p_0\ (1 + r)^n$

HARMONIC MEAN

Harmonic mean is the reciprocal of the arithmetic average of the reciprocals of the values of its various items. Symbolically the Harmonic Mean or HM of a series is

$$\text{H.M.} = \frac{1}{\frac{1}{n}\left[\frac{1}{X_1} + \frac{1}{X_2} + ... + \frac{1}{X_n}\right]} = \frac{1}{\frac{1}{n}\sum\left(\frac{1}{X}\right)} = \frac{n}{\sum \frac{1}{x}}$$

Indiscrete series we first find out the reciprocal of each value and multiply it by the concerned frequency Then total the products and divide the total frequency and then find out the reciprocal of the values, which is the harmonic mean of the series. Thus in a discrete series.

$$\frac{1}{H} = \frac{1}{N}\sum\frac{f}{X},\ \ H = \frac{N}{\sum \frac{f}{X}} = \frac{\sum f}{\sum \frac{f}{X}}\ (N = \textstyle\sum f;\ \text{the total frequency})$$

In Continuous Series: The value of the variable is the mid point of the class interval and the formula for calculation is same as in discrete series.

Weighted Harmonic Mean: Weighted HM = $\dfrac{\sum W}{\sum \frac{W}{X}}$ where W is the weight and X is the value of the variable

MEDIAN

The median is that value of the variable which divides the group in two equal parts. One part comprising all the values greater and the other, all the values less than median. Then median of the distribution may defined as that value of the variable which exceeds and is exceeded by same numbers of observations, i.e., it is the value such that number of observations above it is equal to the number of observations below it. Hence, the median is only positional average its value depends on position occupied by a value in the frequency distribution

Median of individual observations: In case of individual observations x_1, x_2, ……. x_n to find the median we use the following algorithm.

Step 1: Arrange the observations x_1 , x_2, ……. x_n in ascending or descending order of magnitude.

Step 2: Determine the total number of observations, say, n

Step 3: If n is odd, then median is the value of $\left(\frac{n+1}{2}\right)^{th}$ observation. If n is even, then median is the AM of the values of $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n+1}{2}\right)^{th}$ observation

Ex. (i) The following are the makes of 9 students in a class. Find the median

34, 32, 48, 38, 24, 30, 27, 21, 35

(ii) Find the median of the daily wages of ten workers.

Rs. 20, 25, 17, 18, 8, 15, 22, 11, 9, 14

Sol. (i) Arranging the data in ascending order of magnitude, we have 21, 24, 27, 30, 32, 34,35, 38, 48. Since, there are 9, and odd number of items, therefore median is the value of $\left(\frac{9+1}{2}\right)^{th}$ observation, i.e., 32.

(ii) Arranging the wages in ascending order of magnitude, we have 8, 9, 11, 14, 15, 17, 18, 20, 22, 25, Since, there are 10 observations, therefore median is the arithmetic mean of $\left(\frac{10}{2}\right)^{th}$ and $\left(\frac{10}{2}+1\right)^{th}$ observations. So median = (15 + 17)/2 = 16.

Median of discrete frequency distribution In case of a discrete frequency distribution x_i/f ; i = 1, 2, …., n; we calculate the median by using the following algorithm

Step 1: Find the cumulative frequencies (c.f.)

Step 2: Find N/2, where N = $\sum_{i-1}^{n} f_i$

Step 3: See the cumulative frequency (c.f.) just greater than N/2 and determine the co-responding value of the variable.

Step 4: The value obtained in step III is the median

Ex. Obtain the median for the following frequency distribution:

x:	1	2	3	4	5	6	7	8	9
f:	8	10	11	16	20	25	15	9	6

Sol. Calculation of Median

x	f	cf
1	8	8
2	10	18
3	11	29
4	16	45
5	20	65
6	25	90
7	15	105
8	9	114
9	6	120
		N = 120

Here N = 120 $\Rightarrow$ N/2 = 60. We find that the cumulative frequency just greater than N/2, is 65 and the value of x corresponding to 65 is 5. Therefore, median is 5.

Median of a grouped or continuous frequency distribution To calculate the median of a grouped or continuous frequency distribution we use the following algorithm

Step 1: Obtain the frequency distribution

Step 2: Prepare the cumulative frequency column and obtain N = $\sum f_i$ Find N/2

Step 3: See the cumulative frequency just grater than N/2 and determine the corresponding class .This class is known as the median class.

Step 4 Use the formula: Median = $1 + \left(\dfrac{\frac{N}{2} - F}{f}\right) \text{x } h$

where l = lower limit of the median class,

f =frequency of the median class and N = $\sum f$

h = width (size) of the median class,

F = cumulative frequency of the class preceding the median class,

Ex. Calculate the median from the following distribution:

Class :	5 – 10	10 – 15	15 – 20	20 – 25	25 – 30	30 – 35	35 – 40	40 – 45
Freq:	5	6	15	10	5	4	2	2

Sol.

Class	Frequency	Cumulative Frequency
5 – 10	5	5
10 – 15	6	11
15 – 20	15	26
20 – 25	10	36
25 – 30	5	41
30 – 35	4	45
35 – 40	2	47
40 – 45	2	49

Here N = 49 ⇒ N/2 = 49/2 = 24.5.

The cumulative frequency just greater than N/2, is 26 and the corresponding class is 15 – 20.

Thus 15 – 20 is the median class such that l = 15, f = 15, F = 11, h = 5

$$\therefore \text{ Median} = l + \frac{\frac{N}{2} - F}{f} \text{ x } h = 15 + \frac{24.5}{15} \text{ x } 5 = 15 + \frac{13.5}{15} = 19.5$$

Ex. Compute the median from the following data:

Mid-value:	115	125	135	145	155	165	175	185	195
Frequency:	6	25	48	72	116	60	38	22	3

Sol. Here we are given the mid-values. So, we should first find the upper and lower limits of the various classes. The difference between two consecutive values is h = 125 – 115 = 10. So, the

lower limit of a class = mid – value – h/2, and upper limit = mid-value + h/2

Mid-value	Class Groups	Frequency	Cumulative Frequency
115	110 – 120	6	6
125	120 – 130	25	31
135	130 – 140	48	79
145	140 – 150	72	151
155	150 – 160	116	267
165	160 – 170	60	327
175	170 – 180	38	365
185	180 – 190	22	287
195	190 – 200	3	390

$N = \sum f_i = 390$

We have N = 390. ∴ N/2 = 390/2 = 195.

The cumulative frequency just greater than N/2 = 195 is 267 and the corresponding class is 150 – 160. So, 150 – 160 is the median class. ∴l = 150, f = 116, h = 10, F = 151.

$\therefore$ Median = $1 + \dfrac{\frac{N}{2} - F}{f}$ x h → Median = 150 $\dfrac{195 - 151}{116}$ x 10 = 153.8

Ex. If the median of the following frequency distribution is 46, find the missing frequency:

Variable:	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70	70 – 80	Total
Frequency:	12	30	?	65	?	25	18	229

Sol. Let the frequency of the class 30 – 40 be f_1 and that of 50 – 60 be f_2. The total frequency is 229.

$\therefore$ 12 + 30 + f_1 + 65 + f_2 + 25 + 18 = 229 → $f_1 + f_2 = 79$.

Median = 46. Clearly, 46 lies in the class 40 – 50. So, 40 – 50 is the median class.

∴l = 40, h = 10, f = 65 and F = 12 + 30 + f_1 = 42 + f_1, N = 229.

Now, ∴Median = $1 + \dfrac{\frac{N}{2} - F}{f}$ x h → 46 = 40 + $\dfrac{\frac{229}{2} - (42 + f_1)}{65}$ x 10

$\Rightarrow f_1 = 34$ (approx.). Since $f_1 + f_2 = 79$. Therefore $f_2 = 45$. Hence, $f_1 = 34$ and $f_2 = 45$.

<u>Partition Values:</u> The values which divide the series into number of equal parts are called the partition values.

<u>Quartiles:</u> The values which divide the given data into four equal parts are known as Quartiles

Obviously there will be three such point Q_1, Q_2 and Q_3 such that $Q_1, \le Q_2, \le Q_3$ termed as three quartiles. Mathematically,

$$Q_1 = l + \frac{h}{f}\left(\frac{N}{4} - C\right), \qquad Q_2 = \text{Median}, \qquad Q_3 = l + \frac{h}{f}\left(\frac{3N}{4} - C\right)$$

where l = lower limit of the class containing Q_1 or Q_3

f = frequency of the class containing Q_1 or Q_3

h = magnitude of the class containing Q_1 or Q_3

c = c–f of class preceeding the class containing Q_1 or Q_3

<u>Deciles:</u> Deciles are the values which divide the series into ten equal parts Obviously there are nine deciles D_1, S_2, D_3,……, D_9 such that $D_1 < D_2 < D_3 < \ldots\ldots < D_9$. D_5 coincides with the median.

Mathematically

$$D_i = l + \frac{h}{f}\left(\frac{ixN}{10} - C\right)(i = 1, 2, 3, \ldots., 9)$$

Percentiles: Percentiles are the values divide the series into 100 equal parts. Obviously there are 99 percentiles P_1, P_2, P_3,…. P_{99} such that $P_1 \le P_2 \le P_3 \le …. \le P_{99}$. Mathematically

$$P_i = l + \frac{h}{f}\left(\frac{ixN}{100} - C\right) \quad (i = 1, 2, 3, …… , 99)$$

Thus $P_{25} = Q_1$, $P_{50} = D_5 = Q_2$, $P_{75} = Q_3$, $D_1 = P_{10}$, $D_2 = P_{20}$, $D_3 = P_{30}$ …….. $D_9 = P_{90}$

Ex. The following table gives the distribution of male population according to age groups. Find out the first quartile, third quartile and 40th percentile

Age Group:	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70
No. of males:	150	350	400	200	75	50	25

Sol. The cumulative frequency distribution is given by

Age group (years) (x)	No. of Males (f)	Cumulative Frequency
0 – 10	150	150
10 – 20	350	500
20 – 30	400	900
30 – 40	200	1100
40 – 50	75	1175
50 – 60	50	1225
60 – 70	25	1250
		$N = \sum f_i = 1250$

First Quartile: N = 1250 → N/4 = 312.5. The cumulative frequency just greater than N/4, i.e., 312.50 is 500. The corresponding class, i.e., 10 – 20 is the class such that l = 30, h = 20, f = 350, F = 150.

$$\therefore Q_1 = l + \frac{\frac{3N}{4} - F}{f} x\ h = 30 + \frac{937.7 - 900}{200} x\ 10 = 30 + \frac{37.5x10}{200} = 31.875$$

Hence, Q_3 = Third Quartile = 31.875 years.

MODE

Mode is the value which occurs most frequently in a set of observations and around which the other items of the set cluster densely OR the mode of a distribution is the value at the point around which the item tends to be most heavily concentrated.

Computation of Mode In case of frequency distribution mode is the value of the variable corresponding to the maximum frequency. This method can be applied if the distribution is unimodal. In case of continuous frequency distribution the class corresponding to the maximum frequency is called the modal class the value of mode is obtained by the interpolation formula

$$\text{Mode} = l + \frac{h(f_1 - f_0)}{(f_1 - f_0) - (f_2 - f_1)},$$

where l = lower limit of modal class, h = magnitude of the modal class, f_1 = frequency of the modal class,

f_0 = frequency of class preceding the modal class, f_2 = frequency of class succeeding the modal class.

Ex. Compute the mode of the following distribution:

Class – intervals:	0 – 7	7 – 14	14 – 21	21 – 28	28 – 35	35 – 42	42 – 49
Frequency:	19	25	36	72	51	43	28

Sol. Here maximum frequency 72 lies in the class interval 21 – 28. Therefore 21 – 28 is the modal class. $\therefore t = 21, f = 72, f_{-1} = 36, f_1 = 51, i = 7$

$$\text{Mode } (M_0) = l + \frac{f_1}{f_{-1} + f_1} \text{ x } i = 2l + \frac{21}{36 + 51} \text{ x } 7 = 2l + \frac{357}{87} = 2l + 4.103 = 25.103$$

Again using the other formula (which is more accurate) for the mode, we have

$$\text{Mode } (M_0) = l + \frac{f - f_1}{2f - f_{-1} + f_1} \text{ x } i = 2l + \frac{72 - 36}{144 - 36 - 51} \text{ x } 7 = 2l + \frac{252}{57} = 2l + 4.42 = 25.42$$

DISPERSION & STANDARD DEVIATION

Dispersion means "Scatteredness" Dispersion gives us an idea of the homogeneity (compactness) or heterogeneity (scatter) of the distribution. Let us consider the following three series A, B and C have the same size and same mean viz . 15.

	Series	Total	Mean
A	15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15	135	15
B	11, 12, 13, 14, 15, 16, 17, 18, 19	135	15
C	3, 6, 9, 12, 15, 18, 21, 24, 27	135	15

Series A is stationary, i.e., it is constant and shows no variability. Series B is dispersed and Series C is relatively more dispersed or we can say series B is more homogenous (or uniform) as compared with series C or the series C is more heterogeneous than Series B. (Hence, we see that despite the fact that the series are very different from each other, the Average is the same.)

Absolute and Relative Measures of Dispersion: The measures of dispersion which are expressed in terms of the original units of a series are termed as Absolute Measures. On the other hand, Relative measures of dispersion are obtained as percentages and thus are pure numbers independent of the units of measurement

Measures of Dispersion

The various measures of dispersion are as follows

1. **Range:** It is defined as the difference between the two extreme observations of the distributions.

$$\text{Range} = X_{max} - X_{min}$$

Where X_{max} is the greatest observation and X_{min} is the smallest observation of the variable value. In case of a grouped frequency distribution (for discrete) or the continuos frequency distribution, range is defined as the difference between the upper limit of the highest class and the lower limit of the smallest class.

$$\text{Coefficient of Range} = \frac{X_{max} - X_{min}}{X_{max} + X_{min}} = \frac{L - S}{L + S}$$

2. **Quartile Deviation of Semi Inter-Quartile Range**

Inter Quartile Range = $Q_3 - Q_1$

Quartile Deviation (Q. D) = $\frac{Q_3 - Q_1}{2}$ (Also called Semi-Inter Quartile Range)

Quartile Deviation is defined as absolute measure of dispersion. Relative measure is coefficient of Quartile deviation is given by;

$$\text{Coefficient of Q.D.} = \frac{(Q_3 - Q_1)/2}{(Q_3 + Q_1)/2} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

For a symmetrical distribution $Q_3 - Md = Md - Q_1 \Rightarrow Md = \frac{Q_3 + Q_1}{2}$ median lies half way on the scale from Q_1 to Q_3. Thus for symmetrical distribution.

$$QD + Q_1 = \frac{Q_3 + Q_1}{2} + Q_1 = \frac{Q_3 + Q_1}{2} = Md$$

$$Q_3 - QD = Q_1 - \frac{Q_3 - Q_1}{2} = \frac{Q_3 + Q_1}{2} = Md \text{ or } Q_1 = Md - QD \text{ and } Q_3 = Md + QD$$

MEAN DEVIATION

Mean deviation of a series is the arithmetic average of the deviations of various items from a measure of central tendency (either mean, median or mode).

If X_1, X_2, X_3…….. X_n are n given observations then the mean deviation (M.D.) about an average A, is given by

$$\text{M.D. (about an average A)} \ \frac{1}{n}\sum|X-A| = \frac{1}{n}\sum|d|$$

Where $|d|$ = $|X-A|$ i.e. modulus value or absolute value of the deviation (After ignoring the negative sign)

In case of Frequency distribution, Mean Deviation about an average A is given by

$$\text{M.D. (about an average A)} \ \frac{1}{N}\sum|X-A| = \frac{1}{N}\sum|d|$$

Where X is the value of the variable or it is the mid value of the class interval(in case of grouped or continuos frequency distribution)Usually, we obtain the M.D. about any one of the three averages (mean, median, and mode)

$$\text{M.D. (about mean)} = \frac{1}{N}\sum f|X-M|$$

$$\text{M.D. (about median)} = \frac{1}{N}\sum f|X-Md|$$

$$\text{M.D. (about mode)} = \frac{1}{N}\sum f|X-Mo|$$

Series of Individual observations (Direct Method) In direct method the mean deviation would be calculated by totaling the deviations from the mean, median or mode (ignoring +, – sign) and dividing the total by number of items.

$$\text{Thus, M.D.} = \frac{\sum|d|}{N}$$

Coefficient of M.D. = $\dfrac{\text{Mean Deviation}}{\text{Average about which it is calculated}}$

Coefficient of M.D. about Mean = $\dfrac{\text{M.D.}}{\text{Mean}}$

Coefficient of M.D. about Median = $\dfrac{\text{M.D.}}{\text{Median}}$

STANDARD DEVIATION

It is defined as positive square root of the A.M. of the square of the deviations of the given observations from their A.M.

If X_1, X_2,……X_n is a set of N observations then its standard deviation is given by Standard Deviation

$$\sigma = \sqrt{\frac{\sum d^2}{N}} \text{ or } \sigma = \sqrt{\frac{1}{N}\sum(X-\overline{X})^2} \text{ where } \overline{X} = \frac{\Sigma X}{N} \text{ or } \sigma = \sqrt{\frac{1}{N}\sum f(X-\overline{X})^2}$$

Calculation of Standard deviation Series of individual observation

1. Direct method No. 1 S.D. = $\sigma = \sqrt{\frac{\sum d^2}{N}}$ $(d = X - \overline{X})$

2. Direct method No. 2 $\sigma = \sqrt{\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2}$ or $\sigma = \sqrt{\frac{1}{N}\Sigma fx^2 - \left(\frac{1}{N}\Sigma fx\right)^2}$

3. Short cut method $\sigma = \sqrt{\frac{\Sigma dX^2}{N} - \left(\frac{\Sigma dX}{N}\right)^2}$ where dX = deviation of each item from assumed mean.

   Standard deviation is an absolute measure of dispersion for purpose of comparison a relative measure of dispersion is calculated by dividing Standard deviation by A.M. and it is called as **coefficient of dispersion** i.e.,

   Coefficient of dispersion = $\frac{\sigma}{\overline{X}}$. This is also called Coefficient of Standard Deviation.

   Indiscrete Series

   1. Direct method S.D. = $\sigma = \sqrt{\frac{\sum d^2}{N}}$ $(d = X - \overline{X})$

   2. Short cut method $\sigma = \sqrt{\frac{\Sigma fdX^2}{N} - (\overline{X} - A)^2}$ or $\sqrt{\frac{\Sigma fdx^2}{N} - \left(\frac{\Sigma fdx}{N}\right)^2}$

   where $\overline{X} - A = \frac{\sum fdx}{N}$, A is the assumed mean.

Step deviation method First find step deviations and result is multiplied by magnitude of common factor or i.

$$\sigma = \sqrt{\frac{\Sigma fdx^2}{N} - \left(\frac{\Sigma fdx}{N}\right)^2}$$

Calculation of S.D. in continuos series: In continuous series procedure is same as in discrete series the class intervals are represented by their mid points and thus converting the continuos series in a discrete series.

Variance It is the square of the standard deviation and is denoted by σ^2 . Thus for a F.D. variance is

$$\sigma^2 = \frac{1}{N}\sum f(X-\bar{X})^2$$

Mean Square Deviation: (Usually denoted by S^2) $S^2 = = \frac{1}{N}\sum f(X-A)^2$ where A is any arbitrary number, thus root mean square deviation (S) = $\sqrt{\frac{1}{N}\sum f(X-A)^2}$

Coefficient of Variation: Standard deviation is absolute measure of dispersion and is expressed in terms of units of the variable. The corresponding relative measure is known as coefficient of variation. The series for which C.V. is more will be less consistent, less homogenous and less stable. On the other hand the series for which C.V. is less is considered to be more uniform, more consistent and homogeneous

$$\text{C. V} = \frac{\sigma}{\bar{X}} \text{ x } 100$$

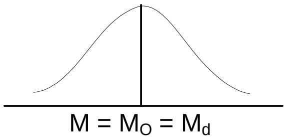
Note: C.V. is a percentage and coefficient of standard deviation is a ratio of standard deviation to mean.

Skewness, Moments & Kurtosis

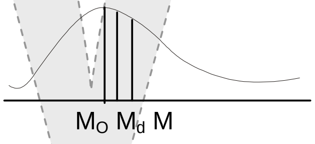
Skewness mean lack of symmetry A distribution is said to be skewed when the mean and median fall at different points in the distribution and balance is shifted to one side or the other – to left or right or a distribution is said to be skewed if:

1. The frequency curve of the distribution is not a symmetric bell shaped curve but it is stretched more to one side than to the other. In other words it has a longer tail to one side (left or right) than to the other. A frequency curve which has longer tail towards right is said to be positively skewed and if longer tail lies towards the left it is said to be negatively skewed.

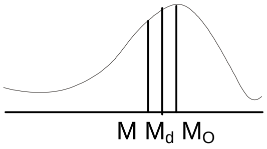
Symmetrical Distribution



Positively Skewed Distribution



Negatively Skewed Distribution



2. The value of mean (M), Median (M_d) and mode (M_o) fall at different point i.e. they do not coincide.

3. $Q_3 - M_0 \neq M_0 - Q_1$

Measures of Skewness

- $Sk = \text{Mean} - \text{Median} = M - M_d$
- $Sk = \text{Mean} - \text{Mode} = M - M_o$
- $Sk = (Q_3 - M_o) - (M_o - Q_1) = Q_3 + Q_1 - 2m_0$

These are absolute measures of skewness. The relative measures of skewness (Coefficients of Skewness) are as follows.

- ♦ Karl Pearson's Coefficient of Skewness
- ♦ Bowley's Coefficient of Skewness
- ♦ Kelly's Coefficient of Skewness

Karl Pearson's Coefficient of skewness

$$Sk = \frac{\text{Mean} - \text{Mode}}{\text{S.D.}} = \frac{M - M_o}{\sigma}$$

Because M_o is maximum times ill defined, we use the empirical relationship between mean, median and mode for moderately asymmetrical distribution $M_o = 3M_o - 2M$

$$Sk = \frac{M - 3M_d + 2M}{\sigma}$$

$$Sk = \frac{3(M - M_o)}{\sigma}$$

Thus,

1. Sk lies between the limit – 1 to + 1.
2. Also skewness is zero if $M = M_d = M_o$.
3. $Sk > 0$ if $M > M_d > M_o$ + vely Skewed distribution
4. $Sk < 0$ if $M < M_o < M_o$ – vely Skewed distribution
5. Skewness studies the direction of variation while dispersion studies the degree or extent of variation.

Bowely's coefficient of Skewness

$$Sk = \frac{(Q_3 - M_d) - (M_d - Q_1)}{(Q_3 - M_d) + (M_d - Q_1)}, \qquad Sk = \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1}$$

It is also known as Quartile Coefficient of Skewness. Thus

1. $Sk = 0$ If $Q_3 - M_d = M_d - Q_1$
2. $Q_3 - M_d > M_d - Q_1$ or $Q_3 + Q_1 > 2M_d$ + vely Skewed distribution
3. $Q_3 - M_d < M_d - Q_1$ or $Q_3 + Q_1 < 2M_d$ + vely Skewed distribution
4. Bowely's coefficient of skewness ranges from – 1 to + 1.
5. Used when open end classes are there or when mode is ill-defined

Limitations: It is based only on central 50 % of the data and ignores the remaining 50% of the data towards the extremes

Kelly's Measure of Skewness

Coefficient of Skewness = $S_k = \frac{(P_{90} - P_{50}) - (P_{50} - P_{10})}{(P_{90} - P_{50}) + (P_{50} - P_{10})} = \frac{D_9 - 2D_5 + D_1}{D_9 - D_1} = \frac{P_{10} + P_{90} - 2 \text{ Median}}{P_{90} - P_{10}}$

This method is primarily of theoretical importance.

MOMENTS

In statistics, the term moment is used with reference to frequencies and the respective values against them. Moments are calculated to study the nature of distribution μ_n denotes the n^{th} order moment about mean (central moment) and is given by

$$\mu_n = \frac{1}{N}\sum f(X - \overline{X})^n$$

μ'_n denotes the n^{th} order moment about A(an arbitrary origin) and is given by

$$\mu_n = \frac{1}{N}\sum f(X - A)^n$$

Relation between central moments and moments about any value A(called an arbitrary origin)

$\mu_1 = 0$, $\mu_2 = \mu'_2 – (\mu'_1)^2$, $\mu_3 = \mu_3 – 3\mu'_1 \ \mu'_2 + (2\mu'_1)^3$, $\mu_4 = \mu'_4 – 4\mu'_3 + 6\mu'_3 \ \mu'_1 + 6\mu'_2(\mu'_1)^2 – 3(\mu'_1)^4$

Beta Coefficients $\beta_1 = (\mu_3^2/\mu_2^3)$, $\beta_2 = (\mu_4/\mu_2^2)$. β_2 is called **kurtosis**

1^{ST} coefficient of skewness = $(\mu_3)/ \sqrt{(\mu_2^3)} = \sqrt{\beta_1} = \alpha_3 = \gamma_1$.

2^{nd} coefficient of skewness = $\frac{\sqrt{\beta_1}(\beta_1 + 3)}{2(5\beta_2 - 6\beta_1 - 9)}$

CORRELATION

The distributions in which each unit of the series assumes two values is called as bi-variate distribution and if we measure more than two variables in each unit of distribution, it is called as multivariate distribution.

Correlation is a statistical tool which studies the relationship between two variables and correlation analysis involves various methods and techniques used for studying and measuring the extent of the relationship between the two variables. Two variables are said to be correlated, if the change in one variable results in a corresponding change in the other variable.

Utility:

- It reduces the range of uncertainty associated with decision making.
- It is very helpful in understanding the economic behaviour.
- It helps in identifying such factors which can stabilize a distribution.
- It helps to estimate the likely change in a variable with a particular amount of change in related variable.
- Interrelationship studies between different variables are very helpful tools in promoting the research and opening new frontiers of knowledge.

Types of Correlation.

1. Positive or Negative
2. Simple Multiple and Partial
3. Linear and Non-Linear

Positive and Negative Correlation. If the values of two variables deviate in same direction so that increase in value of one variable is associated with an increase in value of the other variable and decrease in value of one variable is associated with decrease in value of the other. The correlation is said to be positive. e.g.,

Heights and Weights

Family income and expenditure on luxury items

Amount of rainfall and yield of crop (up to a point)

On the other hand, if the values of two variables deviate in different directions so that with increase in one value of one variable the value of other variable decreases and with a decrease in the value of one variable the value of the other variable increases, correlation is said to be negative e.g.,

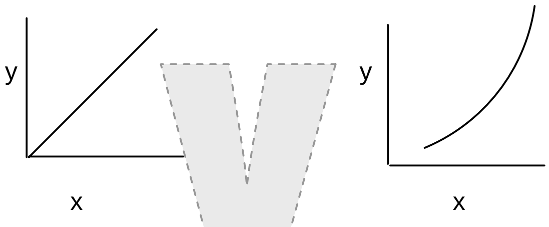
Price and demand of a commodity

Volume and pressure of a perfect gas

Linear and Non Linear Correlation: When variations in the values of two variables are in a constant ratio, correlation is said to be linear, there is a linear relationship between two variables. Their relationship is of type y = a + bx.

On the other hand if the ratio of change in the two variables is not constant e.g., in the corresponding figure of two variables would not give a straight line.

The following two diagrams illustrate the difference between linear and non-linear correlation.



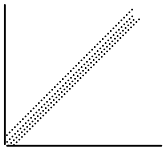
Covariance: By covariance we measure how the pairs of observations $(x_1\ y_1)$, $(x_2\ y_2)$ $(x_3\ y_3)$ are varying or changing. Covariance means the measurement of effect of change of one variable over the other in pair of observations.

The Covariance cov (x, y) between two variables x and y is given by

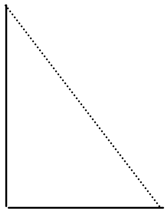
$$\text{cov}\,(x, y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{n} \qquad \text{or } \text{cov}(x, y) = \frac{n\sum xy - (\sum x)(\sum y)}{n^2}$$

1. If + ve value of covariance is very small then there is no trend.
2. If – ve value of covariance is large in magnitude, this means, except for a few pair of values, and increase in value of one variable is associated with decrease in value of other variable and vice versa.
3. Large + ve value means value of two variables mostly increase or decrease together though it is not true for all the pairs.

Scatter diagram. For each pair of x and y values we put a dot. Thus we obtain as many points as the number of observations By looking at the scatter we get an idea about the relationship between the two variables. The greater the scatter, the less is the relationship between the variables.



+ve co-relation



–ve co-relation

Merits: It is simple and non mathematical method of studying correlation between the variables. It can easily be understood.

Demerits: It gives only direction of correlation and also whether it is high or low. We do not get the exact degree of correlation.

Karl Person's Coefficient of Correlation.

Coefficient of correlation

$$r(x,\ y) = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)}\sqrt{\text{var}(y)}} = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} = \frac{\sum(x - \overline{x})(y - \overline{y})}{\sqrt{\sum(x - \overline{x})^2}\sqrt{\sum(y - \overline{y})^2}} = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}$$

Rank Method: Some times we come across statistical series in which variables under consideration are not capable of quantitative measurement but can be arranged in serial order i.e. when we are dealing with qualitative characteristics. e.g., A and B are two qualitative characteristics. We have to arrange a group of n individuals in order of their (merits) ranks. w.r.t. proficiency in two characteristics. When the variables x and y represent distinct ranks, then it can be proved that

$r = 1 - \frac{6\sum(x - y)^2}{n(n^2 - 1)}$. This formula is popularly known as Spearman's rank correlation coefficient.

EXAMPLES

Ex. Find cov (x, y) between x and y from the following data:

$\sum x = 15,\ \sum y = 36,\ \sum xy = 110,\ n = 5.$

Sol. We have $\sum x = 15,\ \sum y = 36,\ \sum xy = 110,\ n = 5$

$$\text{cov}(x, y) = \frac{n\sum xy - (\sum x)(\sum y)}{n^2} = \frac{5(110) - (15)(36)}{(5)^2} = \frac{10}{25} = 0.4$$

Ex. Find cov (x, y) for the following data:

x:	1	2	3	4	5	6	7	8	9	10
y:	6	9	6	7	8	5	12	3	17	1

Sol. $$\text{cov}(x, y) = \frac{n\sum xy - (\sum x)(\sum y)}{n^2} = \frac{10(411) - (55)(74)}{(10)^2} = \frac{40}{100} = 0.4$$

Ex. Find the coefficient of correlation for the following data

x:	5	7	1	3	4
y:	2	2	4	5	6

Sol. Calculation of r

S. No	x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$	$(y - \bar{y})^2$
1	5	2	1	–1.8	–1.8	1	3.24
2	7	2	3	–1.8	–5.4	9	3.24
3	1	4	– 3	0.2	–0.6	9	0.04
4	3	5	– 1	1.2	–1.2	1	1.44
n = 5	4	6	0	2.2	0	0	4.84
Total	20	19	0	0	– 9	20	12.80

$\bar{x} = 4$ $\bar{y} = 3.8$

Ex. Find the coefficient of correlation for the following data:

x:	5	7	1	3	4
y:	2	2	4	5	6

Sol. $n = 5,\ \sum x = 20,\ \sum y = 19,\ \sum xy = 67,\ \sum x^2 = 100,\ \sum y^2 = 85.$

$$r(x, y) = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} = –0.5625$$

Ex. Calculate the coefficient of correlation between the ranks obtained by 10 students in English and Hindi in a class test as given below:

Rank in English :	1	2	3	4	5	6	7	8	9	10
Rank in Hindi:	3	10	5	1	2	9	4	8	7	6

Sol. Let the variables "Rank in English" and "Rank in Hindi" be denoted by x and y respectively.

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} = 0.2242$$

We find r by using Spearman's formula:

S. No	x	y	x – y	(x – y)^2
1	1	3	–3	4
2	2	10	–8	64
3	3	5	–2	4
4	4	1	3	9
5	5	2	3	9
6	6	9	–3	9
7	7	4	3	9
8	8	8	0	0
9	9	7	2	4
10	10	6	4	16
Total				128

$$r = 1 - \frac{6\sum(x-y)^2}{n(n^2-1)} = 0\ .2242$$

REGRESSION

Definition: Regression is the measure of average relationship between two or more variables in terms of original units of the data. Regression analysis is done for estimating or predicting the unknown values of one variable from the known values of other variable.

Methods of studying Regression.
Regression can be studied either Graphically or Algebraically.

Graphical study of Regression: Free hand curve method & the method of least squares

In free hand curve method, first plot the pairs of the values of X and Y in the form of a scatter diagram (one point for one pair of values). After this draw two free hand St. lines. One of the lines is drawn in such a way that-ve deviations of y series from its mean are cancelled by the negative deviations. The sum of deviations on one side is equal to the sum of deviations on the other side. This will be the regression line of Y on X. Similarly draw regression line of X and Y i.e. for X series. The two regression lines cut at mean of two series. If there is perfect + ve or – ve correlation between the two variables there will be only one regression line.

Method of Least Squares: If the value of Y are plotted on Y axis, then the regression line of Y on X will be such that it minimizes the total of the squares of the vertical deviations. Similarly if the value of X are plotted on X axis, then the regression line of X on Y will be such that it minimizes the total of the squares of the horizontal deviations.

Line of best fit is obtained by the equation of straight line Y = a + bX

$$\sum y = na + b\sum X \qquad (1)$$

$$\sum XY = a\sum X + b\sum X^2 \qquad (2)$$

These equations will give us the values of a and b which we fit in the equation of the straight line Y = a + bX. This will give us regression line of Y and X.

Comparison of Correlation and Regression Studies:

1. Correlation analysis is done for studying the co-variation of two variables i.e. whether the variables move in same direction or in reverse direction and also degree of their co-variation, but regression analysis studies the nature of relationship i.e. relative movement in the variable.
2. Correlation between two series is not necessarily a cause and effect relationship while regression presumes one variable as a cause and other as its effect.
3. The coefficient of Correlation varies between ±1. The regression coefficient have the same signs as the correlation coefficient. If r is + ve regression coefficients would also be + ve and if r is – ve regression coefficients would also be -ve.
4. Correlation coefficient cannot exceed unity. One of the regression coefficients can have a higher value than unity but the product of the two regression coefficients can never exceed unity because r is the square root of product of the two regression coefficients.

Comparison between correlation and regression: There is no denying the fact that there are some basic differences in studying correlation and regression between variable. These are as follows:

Sr. No	Correlation	Regression
1.	It studies the degree of relationship between variables	It studies the nature of relationship between the variables
2.	It need not imply cause and effect relationship between variables.	It implies cause and effect relationship between variables.
3.	There may be non-sense correlation between the variables.	In regression analysis, there is nothing like non-sense regression.
4.	The correlation coefficient is independent of change of origin and scale.	The regression coefficients are independent of only change of origin but not of scale.
5.	The correlation coefficient cannot be used for prediction.	The regression lines can be used for prediction.

Regression Lines: Let the variables under consideration be denoted by 'x' and 'y'. The line used to estimate the value of y for a given value of x is called the regression line of y on x. Similarly the line used to estimate the value of x for a given value of y is called the regression line of x on y. In regression line of y on x (x on y) the variable y is considered as the dependent (independent), regression lines depends upon the given pair of values of the variables Regression lines are also known as estimating lines. We shall see that in case of perfect correlation between the variables, the regression lines will be coincident. The angle between the regression lines will increase from 0^0 to 90^0 as the correlation coefficient numerically decreases from 1 to 0. If for a particular pair of variables r = 0, then the regression lines will be perpendicular to each other. The regression lines are determined by using the principle of least squares.

Regression Equations: We have already noted that for two variables x and y, there can be two regression lines If the intention is to depict the change in y for a given change in x, then the regression line of y on x is to the used. Similar argument also works for regression line of x on y.

Regression Equations of y on x: The regression equation of y on x is estimated by using the 'principle of least squares'. This principle will ensure that the sum of the squares of the vertical deviations of actual values of y form estimated values for all possible values of x is minimum.

Mathematically, $\sum (y - y_0)^2$ is least, where y and y_0 are the corresponding actual and computed values of y for a particular value of x.

By using derivatives, it can be proved that the regression equation of y on x is given by

$y - \bar{y} = b_{yx}(x - \bar{x})$ where $\bar{x} = \frac{\sum x}{n}, \bar{y} = \frac{\sum y}{n}, b_{yx} = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$

EXAMPLES

Ex. Find b_{yx} from the data {(x, y)} = {(5, 2), (7, 4), (8, 3), (4, 2), (6, 4)}

Sol. Here n = 5, $\sum x = 30$, $\sum y = 15$, $\sum xy = 94$, $\sum x^2 = 190$

Hence $b_{yx} = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = 0.4$

Ex. x and y are correlated variable. Ten observations of values of (x, y) have the following results: $\sum = 55$, $\sum y = 55$, $\sum xy = 350$, $\sum x^2 = 385$. Predict the value of y when the value of x is 6.

Sol. To predict the value of y for a given value of x, we shall require the equation of regression line of y on x. The equation of regression line of y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$ …. (1)

We have $\sum x = 55$, $\sum y = 55$, $\sum xy = 350$, $\sum x^2 = 385$, n = 10.

Now, $\bar{x} = \frac{\sum x}{n} = \frac{55}{10} = 5.5$, $\bar{y} = \frac{\sum x}{n} = \frac{55}{10} = 5.5$, $b_{yx} = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \frac{475}{825} = 0.5758$

Therefore, from (1) → $y - 5.5 = 0.5758\,(x - 5.5)$ → $y = 0.5758x + 2.3331$.

This is the equation of regression line of y on x. When x = 6, the predicted value of y is $0.5758(6) + 2.3331 = 5.7879$.

Ex. For the following data, find the regression line of y on x.

x:	1	2	3	4	5	8	10
y:	9	8	10	12	14	16	15

Sol. Here we have $n = 7$, $\sum x = 33$, $\sum y = 84$, $\sum xy = 451$, $\sum x^2 = 219$.

The regression line of y on x is $y - \overline{y} = b_1 (x - \overline{x})$ ….. (1)

Now, $\overline{x} = \frac{\sum n}{n} = \frac{33}{7} = 4.714$, $\overline{y} = \frac{\sum y}{n} = \frac{84}{7} = 12$, $b_{yx} = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \frac{385}{444} = 0.867$

There equation of regression line of y on x is

$y - 12 = 0.867\ (x - 4.714)$ or $y = 0.867x + 12 - (0.867)\ (4.174)$ or $y = 0.867 + 7.913$

Ex. Find the regression coefficient b_{xy} between x and y for the following data:

$\sum x = 30$, $\sum y = 42$, $\sum xy = 199$, $\sum x^2 = 184$, $\sum y^2 = 318$, $n = 6$.

Sol. $b_{xy} = \beta\beta = -0.4583$

Ex. For the observation of pairs (x, y) of the variables x and y, the following results are obtained:

$\sum = 110$, $\sum y = 70$, $\sum x^2 = 2500$, $\sum y^2 = 2000$, $\sum xy = 100$, $n = 20$.

Sol. We have $\sum x = 110$, $\sum y = 70$, $\sum x^2 = 2500$, $\sum y^2 = 2000$, $\sum xt = 100$, $n = 20$. The equation of regression line of x on y is $x - \overline{x} = b_{xy}\ (y - \overline{y})$.

Now, $\overline{x} = \frac{\sum x}{n} = \frac{110}{20} = 5.5$, $\overline{y} = \frac{\sum x}{n} = \frac{70}{20} = 3.5$, $b_{xy} = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \frac{-5700}{35100} = -0.1624$

Therefore, from (1) → $x - 5.5 = -0.1624\ (y - 3.5)$ → $x = -0.1624 + (0.1624)\ (3.5) + 5.5$

When y = 4, the estimated value of x is

$-0.1624(4) + 6.0684 = 5.4188$.

Ex. For the following data, find the regression line of x on y.

x:	1	2	3	4	5	6	7
y:	9	8	10	12	11	13	14

Sol. Here we have $\sum x = 28$, $\sum y = 77$, $\sum xy = 83$, $\sum y^2 = 875$, $n = 7$

The equation of regression line of y on x is $x - \overline{x} = b_{xy}\ (y - \overline{y})$

$\overline{x} = \frac{\sum x}{n} = \frac{28}{7} = 4$, $\overline{y} = \frac{\sum y}{n} = \frac{77}{7} = 11$, $b_{xy} = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \frac{182}{196} = 0.929$.

$\therefore$ The equation is $x - 4 = 0.929\ (y - 11)$ → $x = 0.929y - (0.929)\ (11) + 4$ → $x = 0.929y - 6.219$.

Method of Least Squares & Curve Fitting

Let us consider m independent linear equations

$a_{11} X_1 + a_{12} X_2 + a_{12} X_3 + \ldots + a_{1n} X_n = b_1$

$a_{21} X_1 + a_{22} X_2 + \quad a_{23} X_3 + \ldots + a_{2n} X_n = b_2$

…. $\qquad$ …. $\qquad$ ……

$a_{m1} X_1 + a_{m2} C_2 + \quad a_{m3} X_3 + \ldots + a_{mn} X_n = b_n$

Where a's and b's constants and X_1, X_2 X_3 … X_n are the n variables.

If m = n, we can find in general a unique set of values of satisfying the given system of equations.

If m > n i.e if the number of equations is greater than the number of variables, then no solution may exists so our aim is to find these values of the variables X_1, X_2 … X_n which satisfy as nearly as possible the given equations.

These values are called the most possible values or the best values in the least square sense.

METHOD OF LEAST SQUARES

Let X_1, X_2 … X_n be most possible values, then $a_{11} X_1 + a_{12} X_2 + a_{13} X_3 + \ldots.. a_m X_n - b$, is defined as the deviation or the residual or error of this i^{th} equation out of the given m equation and is denoted by E_i i.e E_i

$E_i = a_{i1} X_1 + a_{i2} X_2 + \ldots. a_{in} X_n - b_i$

where i = 1, 2, 3…m. Also S denote the sum of squares of these errors then we have;

$$S = \sum_{i=1}^{m} (a_{i1} X_1 + a_{i2} X_2 + \ldots + a_{in} X_n - b_i)^2 = \sum_{i=1}^{m} E_i^2 \ldots\text{(ii)}$$

Also from our knowledge of Differential calculus (Maxima and Minima), we know the theorem that the extreme values of the function

F = F(X_1, X_2 X_3 … X_n) are given by $\frac{\partial F}{\partial x_1} = 0 = \frac{\partial F}{\partial x_2} = \frac{\partial F}{\partial x_3} = \ldots.\frac{\partial F}{\partial x_n}$

provided the partial derivatives exist. With the help of this theorem S will have a maximum or minimum value for those values of the variables X_1, $X_{2,}$ … X_n which satisfy the following n equations.

$$\frac{\partial S}{\partial x_1} = 0 = \frac{\partial S}{\partial x_2} = 0 \text{ ie } \sum_{i=1}^{m} a_{i1} E_i = 0 \sum_{i=1}^{m} a_{i2} E_i = 0 \ldots \sum_{i=1}^{m} a_n E_i = 0 \ldots.\text{(iii)}$$

These equations are known as the normal equations and can solved like simultaneous equations for n variables X_1, X_2 … X_n and values obtained are most plausible or best values of the unknowns.

EXAMPLES

Ex. Fit a straight line to the following data regarding x as the independent variable.

x:	0	1	2	3	4
y:	1	1.8	3.3	4.5	6.3

Sol. Suppose a straight line to be fitted to the given data is as follows

y = a + bx ….. (1), then the normal equations are as:

$\sum y = ma + b\sum x$ …. (2) and $\sum xy = a\sum x + b\sum x^2$ ….. (2)

Now, from the given data, we have

x	y	xy	x^2
0	1	0	0
1	1.8	1.8	1
2	3.3	6.6	4
3	4.5	13.5	9
4	6.3	25.2	16

Total $\sum x = 10$ $\sum y = 16.9$ $\sum xy = 47.1$ $\sum x^2 = 30$

Here m = 5. Now substituting these values in the normal equations, we get 16.9 = 5a + 10b ….(4)

47.1 = 10a + 30b …..(5)

Solving (4) and (5) we have a = 0.72 and b = 1.33.

Thus, the required equation of the straight line is y = 0.72 + 1.33 x [Putting a and b in (1)]

Ex. Fit a second degree parabola to the following data regarding x as an independent variable:

x:	0	1	2	3	4
y:	1	5	10	22	38

Sol. Let the equation of second degree parabola to be fitted to the given data be $y = a + bx + cx^2$ …(1)

Then its normal equations are $\sum y = ma + b\sum x + c\sum x^2$ ….. (2)

$\sum xy = a\sum x + b\sum x^2 + c\sum x^3$ …..(3)

$\sum x^2y = a\sum x^2 + b\sum x^3 + c\sum x^4$ …… (4)

x	**y**	**x^2**	**x^3**	**x^4**	**xy**	**x^2y**
0	1	0	0	0	0	0
1	5	1	1	1	5	5
2	10	4	8	16	20	40
3	22	9	27	81	66	198
4	38	16	64	256	152	608
$\sum x$ =10	$\sum y$ =76	$\sum x^2$=16.9	$\sum x^3$=47.1	$\sum x^4$=30	$\sum xy$ =243	$\sum x^2y$ = 851

Substituting these values in the normal equations, we get 76 = 5a + 10b + 30c …. (i)
243 = 10a + 30b + 100c …..(ii)

851 = 30a + 100b + 354c …. (iii)

Solving these equations, we have

a = 1.43, b = 0.24, c = 2.21.

∴The required equation of second degree parabola is $y = 1.43 + 2.24x + 2.21\ x^2$

Ex. From normal equations and hence find the most plausible values of x and y from the following $x + y = 3.01$, $2x - y = 0.03$, $x + 3y = 7.03$, $3x + y = 4.97$

Sol. Suppose $U = (x + y - 3.01)^2 + (2x - y - 0.03)^2 + (x + 3y - 7.03)^2 + (3x + y - 4.97)^2$ ….. (1)

These values of x and y which make U minimum are called most possible values. They will be given by following equations.

$$\frac{\partial U}{\partial x} = 0 \frac{\partial U}{\partial y}$$

Differentiating (1) partially w.r.t. x, we have

$(x + y - 3.01) + (2x - y - 0.03)2 + (x + 3y - 7.03) + (3x + y - 4.97)3 = 0$ i.e.,$15x + 5y = 25$ .. (2)

Differentiating (1) partially w.r.t. y, we have

$(x + y - 3.01) + (2x - y - 0.03) +3(x + 3y - 7.03) +(3x + y - 4.97) = 0$ i.e,$5x +12y =29.04$ .. (3)

Solving (2) and (3), we have $x = 0.999$ and $y = 2.004$.

These values are the most plausible values of x and y and so integral values of x and y are 1 and 2 respectively.