DATA ANALYSIS

Classification of data:
Usually any data which is available as raw data may be written in the form of a table, for better understanding of the data, by adopting the following steps:

(i). Grouping of data using tally mark.
(ii). Formation of classes - width of the class is called as class interval.
(iii). Writing down the Frequencies

The table showing the classes and the corresponding frequencies is called the frequency table.

Thus the set of raw data summarized by distributing it into a No. of classes along with their frequencies is known as a frequency distribution.

Averages or Measures of Central Tendency:

A frequency distribution, in general shows clustering of data around some central value. An average is the central value of the frequency distribution which is the most representative value of the entire distribution.

The following or the measures of central tendency:
i) Mean ii) Median iii) Mode iv) Geometric Mean v) Harmonic Mean

Arithmetic Mean or Mean:

The Arithmetic Mean or Mean of a set of observations is their sum divided by the No. of observations. i.e. The Arithmetic Mean, denoted by $\bar{x}$, of n observations $x_1, x_2, \ldots, x_n$ is

$$\bar{X} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

In the case of a frequency distribution, where f_i is the frequency of x_i ; i=1,2,…,n, we have

$$\bar{X} = \frac{f_1 x_1 + f_2 x_2 + \cdots + f_n x_n}{f_1 + f_2 + \cdots + f_n} = \frac{\sum f_i x_i}{\sum f_i}$$

The same formula will hold for a grouped data, where x_i will then be the mid-points of the clases.

Calculation of Mean:

(i). By taking the deviations of the values of x_i from an arbitrary point A:
Let $d_i = x_i - A$
$\Rightarrow \bar{x} = A + \frac{\sum f_i d_i}{\sum f_i}$

(ii). In case of grouped data:
Let $d_i = \frac{x_i - A}{h}$; where h is the width of the class interval
$\Rightarrow \bar{x} = A + h \frac{\sum f_i d_i}{\sum f_i}$

Measures of Dispersion:

The measures of Central Tendency gives us an idea of the concentration of the observations about a central value but they are inadequate in telling us the complete distribution. i.e. how an individual value differs from this central value. Thus, they must be supported and supplemented by some other measures. One such measure is Dispersion.

The following are the measures of dispersion:
(i) Range (ii) Quantile deviation (iii) Mean deviation (iv) Standard Deviation

The square of the standard deviation is known as the Variance

Standard Deviation (or Variance)::

The standard deviation (σ) is the square root of the mean of the squares of the deviations of the given values from their mean and is given by

$\sigma = \sqrt{\frac{1}{n}\sum(x_i - \bar{x})^2}$; where n is the No. of observations

or, $\sigma = \sqrt{\frac{1}{N}\sum f_i(x_i - \bar{x})^2}$; where $N = \sum f_i$ is the total frequency

The square of standard deviation is the variance.

Calculation of Standard Deviation (or Variance):

$$\sigma = \sqrt{\frac{1}{n}\sum(x_i - \bar{x})^2}$$

$$\sigma = \sqrt{\frac{1}{n}(\sum x_i^2) - \bar{x}^2}$$

and $$\sigma = \sqrt{\frac{1}{N}\sum f_i(x_i - \bar{x})^2}$$

$$= \sqrt{\frac{1}{N}\sum f_i x_i^2 - \left(\frac{1}{N}\sum f_i x_i\right)^2}$$

(i). If $d_i = x_i - A$ so that $\bar{x} = A + \frac{1}{N}\sum f_i d_i$

$$\sigma = \sqrt{\frac{1}{N}\sum f_i d_i^2 - \left(\frac{1}{N}\sum f_i d_i\right)^2} \quad \Rightarrow \sigma^2 \text{ is independent of change of origin.}$$

(ii). If $d_i = \frac{x_i - A}{h}$ so that $\bar{x} = A + \frac{h}{N}\sum f_i d_i$

$$\sigma = h\sqrt{\frac{1}{N}\sum f_i d_i^2 - \left(\frac{1}{N}\sum f_i d_i\right)^2} \quad \Rightarrow \sigma^2 \text{ is not independent of change of scale.}$$

Example: Find the Mean and the Standard Deviation for the following data

Age Group (Class Intervals)	No. of members (frequency) f	Mid value x	$d = \frac{x-55}{10}$	fd	fd^2
20-30	03	25	-3	-9	27
30-40	61	35	-2	-122	244
40-50	132	45	-1	-132	132
50-60	153	55	0	0	0
60-70	140	65	1	140	140
70-80	51	75	2	102	204
80-90	02	85	3	06	18
Total	N=542			-15	765

$$\therefore \text{ Mean} = \bar{x} = A + h\,\frac{\sum fd}{N} = 55 + \frac{10*(-15)}{542} = 55\text{-}0.28 = 54.72$$

$$\text{Variance} = \sigma^2 = h^2\left(\frac{1}{N}\sum fd^2 - \left(\frac{1}{N}\sum fd\right)^2\right) = 100\left(\frac{765}{542} - (0.0285)\right)$$

$$= 100 * 1.4107$$

$$= 141.07$$

$$\Rightarrow \text{Standard Deviation} = \sigma = 11.9$$

Correlation Coefficient:

The Correlation Coefficient is a measure of the degree of linear relationship between two variables.

Let X and Y be the two variables taking values $(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)$. Then the correlation coefficient r_{xy} or 'r' between X and Y, is defined as

$$r = \frac{\frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y} = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

Calculation of r:

$$\text{I). } r = \frac{n\sum xy - [(\sum x)(\sum y)]}{\sqrt{[\{n\sum x^2 - (\sum x)^2\}\{n\sum y^2 - (\sum y)^2\}]}} = \frac{\frac{1}{n}\sum xy - \bar{x}\bar{y}}{\sqrt{\left(\frac{1}{n}\sum x^2 - \bar{x}^2\right)\left(\frac{1}{n}\sum y^2 - \bar{y}^2\right)}}$$

II). If $d_x = x - A$ and $d_y = y - B$, where A and B are arbitrary origins then

$$\bar{x} = A + \frac{\sum d_x}{n} \quad \text{and} \quad \bar{y} = B + \frac{\sum d_y}{n}$$

$$\sigma_x^2 = \frac{\sum d_x^2}{n} - \left(\frac{\sum d_x}{n}\right)^2 \text{ and } \sigma_y^2 = \frac{\sum d_y^2}{n} - \left(\frac{\sum d_y}{n}\right)^2$$

$$r = \frac{n\sum d_x.d_y - (\sum d_x)(\sum d_y)}{\sqrt{[n\sum d_x^2 - (\sum d_x)^2]\,[n\sum d_y^2 - (\sum d_y)^2]}}$$

III). If $U = x - \bar{x}$ and $V = y - \bar{y}$

then $\sigma_x^2 = \frac{\sum U^2}{n}$ and $\sigma_y^2 = \frac{\sum V^2}{n}$

$$r = \frac{\sum UV}{\sqrt{\sum U^2 \sum V^2}};$$

Note: r lies between -1 and +1 i.e. $-1 \leq r \leq +1$

Example: Find the Correlation coefficient for the following data.

Student	Intelligence ratio		Eng. ratio		U^2	V^2	UV
	x	$(x-\bar{x})=U$	y	$(y-\bar{y})=V$			
A	105	6	101	3	36	9	18
B	104	5	103	5	25	25	25
C	102	3	100	2	9	4	6
D	101	2	98	0	4	0	0
E	100	1	95	-3	1	9	-3
F	99	0	96	-2	0	4	0
G	98	-1	104	6	1	36	-6
H	96	-3	92	-6	9	36	18
I	93	-6	97	-1	36	1	6
J	92	-7	94	-4	49	16	28
Total	990	0	980	0	170	140	92

where $\bar{x} = \frac{990}{10} = 99$ and $\bar{y} = \frac{980}{10} = 98$

$\therefore$ Correlation coefficient is given by,

$$r = \frac{\sum UV}{\sqrt{\sum U^2 \sum V^2}} = \frac{92}{\sqrt{170*140}} = \frac{92}{154.3} = 0.596$$

Result:

To show that r_{xy} is $\frac{\sigma_X^2+\sigma_Y^2-\sigma_{X-Y}^2}{2\sigma_X\sigma_Y}$

Proof:

$\sigma_{X-Y}^2 = V(X-Y)$

$\sigma_{X-Y}^2 = V(X) + V(Y) - 2Cov(X,Y)$

$\sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2 - 2r\sigma_X\sigma_Y$

$\Rightarrow r = \frac{\sigma_X^2+\sigma_Y^2-\sigma_{X-Y}^2}{2\sigma_X\sigma_Y}$

Aliter:

Let $Z = X - Y$

$\equiv Z_i = X_i - Y_i$

$\bar{Z} = \bar{X} - \bar{Y}$

$\therefore\ Z_i - \bar{Z} = (X_i - \bar{X}) - (Y_i - \bar{Y})$

$\Rightarrow (Z_i - \bar{Z})^2 = [(X_i - \bar{X}) - (Y_i - \bar{Y})]^2$

Taking summation on both sides we get

$\frac{1}{n}\sum(Z_i - \bar{Z})^2 = \frac{1}{n}\sum[(X_i - \bar{X}) - (Y_i - \bar{Y})]^2$

$\sigma_Z^2 = \frac{1}{n}\sum(X_i - \bar{X})^2 + \frac{1}{n}(Y_i - \bar{Y})^2 - 2\frac{1}{n}\sum(X_i - \bar{X})(Y_i - \bar{Y})$

$\sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2 - 2r\sigma_X\sigma_Y \qquad \left[\because r = \frac{\frac{1}{n}\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sigma_X\sigma_Y}\right]$

On simplification we get

$r = \frac{\sigma_X^2+\sigma_Y^2-\sigma_{X-Y}^2}{2\sigma_X\sigma_Y}$