3.2 Characteristics of Good Measure of Central Tendency

An average with following characteristics can be called as an ideal average :

(1) It should be well defined and rigid.

(2) It should be easy to understand and calculate.

(3) It should be based on all the observations of the data.

(4) It should be suitable for further algebraic operations.

(5) It should be a stable measure. It means that values of averages found for different samples of same size from the same population should be almost same.

(6) It should not be unduly affected by a few very large or very small observations.

We will discuss the following measures of central tendency which are widely used in data analysis.

(1) Mean (2) Median and other positional averages (3) Mode.

3.3 Arithmetic Mean or Mean

This is one of the most commonly used averages.

3.3.1 Meaning

Arithmetic mean is defined as the value obtained by sum of all observations divided by the total number of observations.

The arithmetic mean of variable x is denoted by $\bar{x}$.

Calculation of mean :

For raw or ungrouped data :

Suppose $x_1, x_2, ..., x_n$ are the n observations in the data, then Arithmetic mean is $\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$

$$= \frac{\Sigma x_i}{n}$$

Where $\Sigma x_i = x_1 + x_2 + ... + x_n$ = Sum of observations $x_1, x_2, ..., x_n$

and n = number of observations

Note : For the sake of simplicity while solving examples, we will not write the suffix i. Thus we will take x instead of x_i, d instead of d_i and f instead of f_i.

Illustration 1 : The following data [illegible]

Thus, the correct mean weight is 55.45 kg.

r grouped data :

r discrete frequency distribution :

Suppose $x_1, x_2, \ldots x_k$ are the observations in the given data with frequencies $f_1, f_2, \ldots f_k$ respectively.

Here, n = total number of observations

$$= f_1 + f_2 + \ldots + f_k = \Sigma f_i$$

The frequency of x_1 is f_1 means observation x_1 is repeated f_1 times. The sum of all x_1 observations ll be $f_1 \times x_1$ or $f_1 x_1$. Similarly, sum of all x_2 observations will be $f_2 x_2$ and so on.

$$\text{Mean } \bar{x} = \frac{\text{sum of all observations}}{\text{total number of observations}}$$

$$= \frac{f_1x_1 + f_2x_2 + \ldots + f_kx_k}{n}$$

$$= \frac{\Sigma f_i x_i}{n}$$

Where $\Sigma f_i x_i = f_1 x_1 + f_2 x_2 + \ldots + f_k x_k$

Short cut method :

As done earlier for the raw data, an assumed mean A can be suitably chosen and the deviations of values $x_1, x_2, ..., x_k$ can be taken from A. Further, if all these deviations have a common factor c, we can further simplify the calculations by dividing all the deviations by c.

Thus we will have the values $d_1 = \frac{x_1 - A}{c}$, $d_2 = \frac{x_2 - A}{c}$, ..., $d_k = \frac{x_k - A}{c}$

Now, the formula for mean is written as follows :

$$\text{Mean } \bar{x} = A + \frac{\Sigma f_i d_i}{n} \times c$$

Where $\Sigma f_i d_i = f_1 d_1 + f_2 d_2 + ... + f_k d_k$

and n = total number of observations

$= f_1 + f_2 + ... + f_k = \Sigma f_i$

Note : The choice of values of A and c does not change the value of mean.

Illustration 5 : The time (in minutes) taken by a bus to travel between two towns on different days

3.3.2 Combined Mean and Weighted Mean

Combined Mean :

If we know the means of two or more groups of observations, we can find mean of the combined group. Such a value is called as combined mean. It is denoted by $\bar{x}_c$.

Suppose $\bar{x}_1, \bar{x}_2, ..., \bar{x}_k$ are means of k groups having $n_1, n_2, ...n_k$ observations respectively.

The formula for combined mean is as follows :

$$\bar{x}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + ... + n_k\bar{x}_k}{n_1 + n_2 + ... + n_k}$$

Illustration 10 : A factory owner knows that the mean monthly production from January to March is 350 items, from April to August it is 254 items and from September to December it is 315 items. Find the mean monthly production for that year.

Here $n_1 = 3$ months, $n_2 = 5$ months, $n_3 = 4$ months,

$\bar{x}_1 = 350$, $\bar{x}_2 = 254$, $\bar{x}_3 = 315$

Combined mean $\bar{x}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + n_3\bar{x}_3}{n_1 + n_2 + n_3}$

$$= \frac{3(350) + 5(254) + 4(315)}{3 + 5 + 4}$$

$$= \frac{1050 + 1270 + 1260}{12}$$

$$= \frac{3580}{12}$$

$$= 298.3333$$

$$\approx 298.33$$

Thus, mean monthly production of the factory over the year is 298.33 items.

Weighted Mean :

We said that using arithmetic mean is not appropriate if the importance of all observations is not same. special mean called as weighted mean can be found in such cases. Weighted mean is denoted by $\bar{x}_w$. Each ervation is assigned a numerical value called weight according to its importance. The most important value iven maximum weight.

Suppose $w_1, w_2, \ldots, w_n$ are the weights assigned to observations $x_1, x_2, \ldots, x_n$ respectively.

The formula for weighted mean is given as follows :

$$\text{Weighted mean } \bar{x}_w = \frac{w_1x_1 + w_2x_2 + \ldots + w_nx_n}{w_1 + w_2 + \ldots + w_n}$$

$$= \frac{\Sigma w_i x_i}{\Sigma w_i}$$

Here $\Sigma w_i x_i = w_1 x_1 + w_2 x_2 + \ldots + w_n x_n$

And $\Sigma w_i = w_1 + w_2 + \ldots + w_n$

= sum of weights

ion 13: A student gets 35 marks in theory paper, 15 marks in practical examination and 5 marks in oral examination of a subject. The school ...

3.4 Measures of Positional Average

Median, Quartiles, Deciles, Percentiles :

We studied that mean is an appropriate average if we have data which are evenly distributed around the average and the data which do not have too large or too small values. It is said that mean does not become a good representative of data if these conditions are not satisfied. Another average is more suitable in such situations which is called as median. It is a positional average. In addition to median, quartiles, deciles and percentiles are also other positional averages.

3.4.1 Meaning

Median, quartiles, deciles and percentiles are called positional averages because their values are found using the value of an observation at a specific position among the values of variable in the ordered data.

Median :

Median is defined as the value of middlemost observation when the data are arranged in either ascending or descending order. It is denoted by M. In other words, 50% values of observations in the data are above the median and 50% observations have value less than the median.

Calculation of median :

For raw data :

As we have to find the value at the centre, the observations have to be arranged in ascending or descending order.

For n observations $x_1, x_2, \ldots, x_n$, median is found as follows :

$$\text{Median } M = \text{value of the } \left(\frac{n+1}{2}\right) \text{ th observation}$$

For example, if we have 15 observations, the value of the $\left(\frac{15+1}{2}\right)$th that is the 8th observation will be exactly the central value, which is called as median.

Suppose the given data consists of 20 observations. Then, as $\frac{n+1}{2} = \frac{20+1}{2} = 10.5$, we say that the 10th and the 11th observations are both in the centre. In this case, median will be taken as the mean of these two central values.

Thus, the [illegible] required for typing is 12.5 min.

For continuous frequency distribution :

A continuous frequency gives the values of the variable in the form of class intervals and they are generally arranged in ascending order. In such cases, we will use the cumulative frequencies to find the median. These cumulative frequencies will show us the class containing median. For this, we take Median class = class containing the $\left(\frac{n}{2}\right)$th observation

Where $n = f_1 + f_2 + \ldots + f_k = \Sigma f_i$ = total number of observations

The following formula is used to find the median :

$$\text{Median } M = L + \frac{\frac{n}{2} - cf}{f} \times c$$

Where L = lower boundary point of the median class

cf = cumulative frequency of the class prior to median class

f = frequency of the median class

c = length of median class

percentile P_j.

Type of Data	*j*th Quartile $j = 1, 2, 3$	*j*th Decile $j = 1, 2, \ldots 9$	*j*th Percentile $j = 1, 2, \ldots, 99$
Raw data and Discrete frequency distribution	Q_j = value of the $j\left(\frac{n+1}{4}\right)$ th observation	D_j = value of the $j\left(\frac{n+1}{10}\right)$ th observation	P_j = value of the $j\left(\frac{n+1}{100}\right)$ th observation
Continuous frequency distribution	Class of Q_j = class of the $j\left(\frac{n}{4}\right)$ th observation $Q_j = L + \frac{j\left(\frac{n}{4}\right) - cf}{f} \times c$ Where L = lower boundary point of class of Q_j cf = cumulative frequency of the class prior to class of Q_j f = frequency of the class of Q_j c = length of class of Q_j	Class of D_j = class of the $j\left(\frac{n}{10}\right)$ th observation $D_j = L + \frac{j\left(\frac{n}{10}\right) - cf}{f} \times c$ Where L = lower boundary point of class of D_j cf = cumulative frequency of the class prior to class of D_j f = frequency of the class of D_j c = length of class of D_j	Class of P_j = class of the $j\left(\frac{n}{100}\right)$ th observation $P_j = L + \frac{j\left(\frac{n}{100}\right) - cf}{f} \times c$ Where L = lower boundary point of class of P_j cf = cumulative frequency of the class prior to class of P_j f = frequency of the class of P_j c = length of class of P_j

1, 10, 2, 6, 3, 4, 15, 1, 5, 9, 2, 4, 3, 1, 10,
7, 3, 5, 4, 2, 4, 8, 5, 3, 1, 9, 6, 2, 3, 7

-pital stays (in days) of 30 patients admitted to a hospital.

Find the median stay. Further convert this information in a continuous frequency distribution (inclusive type) by taking classes of equal length starting from 1 –3. Find the median from the frequency distribution and compare it with your earlier answer.

*

3.5 Mode

We have earlier studied the mean and the median as the measures of central tendency. We shall now study 'mode' as another measure which is extensively used in business and commercial fields.

3.5.1 Meaning :

The value which gets repeated maximum number of times or the value occurring with maximum frequency in the given data is called as **mode.** It is denoted by M_o.

It is very often used in business to give a representative value for a set of data. For example, see the following statements :

(1) On an average 3 languages are known to the students of this school.
(2) The average height of the men in our country is 1.7 m.
(3) The average daily production of our company is 50 items.
(4) The average daily overtime put in by the workers of a factory is 3 hours.

The value which is repeated most number of times is considered in the calculation of the average in these situations. As per the first statement, it is implied that most of the students know three languages. Thus we can say that mode is used as an average here.

Calculation of mode

The class having maximum frequency is called as modal class of the frequency distribution. The mode is further obtained using the following formula.

$$\text{Mode } M_o = L + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times c$$

Where L = lower boundary point of the modal class

f_m = frequency of the modal class

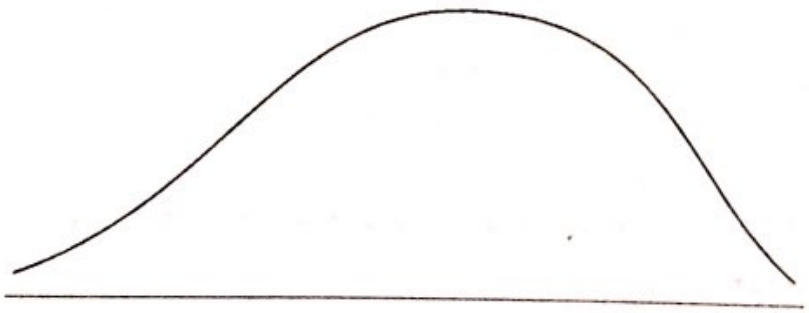
f_1 = frequency of the class prior to modal class

f_2 = frequency of the class succeeding to modal class

c = class length of the modal class

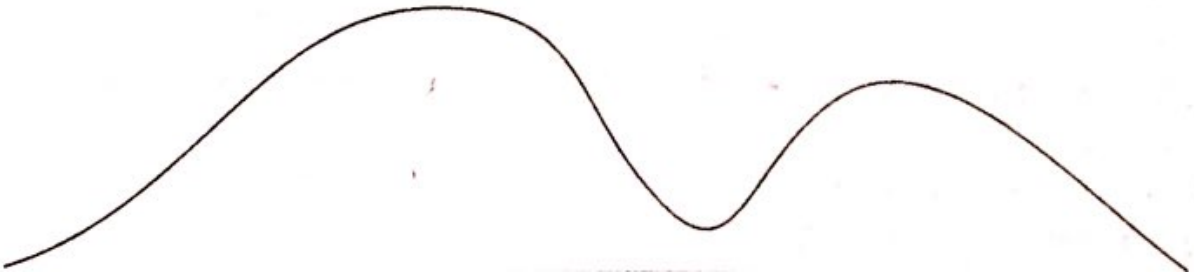
Note : The above formula can be used only if the distribution has classes of equal class length. Moreover, the formula can be used only in those cases where the maximum frequency is only for one class.

The frequency distribution in which the frequencies increase initially and then start decreasing after attaining the maximum frequency is called as a **regular frequency distribution**. Such distributions are also called as **unimodal distributions** as the distribution has only one mode. The frequency curve of such distributions is as follows :



Frequency curve of regular distribution

For bimodal distribution, the frequencies increase and then decrease but then again increase and decrease. Such a frequency distribution is called as an **irregular frequency distribution** whose frequency curve is as follows.



4.3.1 Range

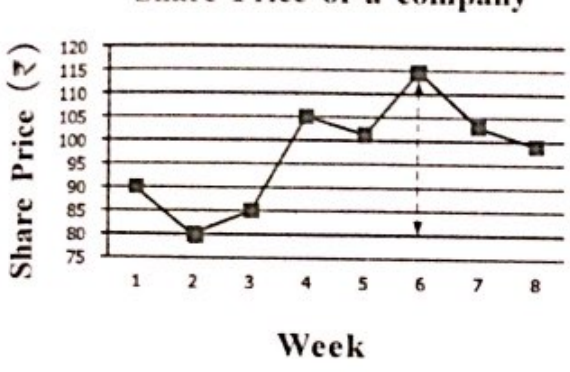
The difference between the highest and lowest observation of the data is called the Range and it is denoted by the symbol R.

$\therefore$ Range $R = x_H - x_L$

where x_H = the highest observation

x_L = the lowest observation

Range R is an absolute measure of dispersion having same unit of measurement as that of the observations.



The data of weekly closing share prices of a company for 8 weeks are given.

From the graph, it is clear that the maximum price is ₹ 115 and its minimum price is ₹ 80. So the range will be 115 – 80 = ₹ 35.

It is obvious from the definition of range that frequency does not play any role in determining the range, even for grouped data. For any grouped frequency distribution, the range can be obtained as the difference between the upper limit of the highest class interval and the lower limit of the lowest class interval.

If we divide the range R of the data by the sum $x_H + x_L$, we get the relative range.

$$\therefore \quad \text{Relative Range} = \frac{R}{x_H + x_L} = \frac{x_H - x_L}{x_H + x_L}$$

The relative range is also known as **coefficient of range.** It is free from the unit of measurement.

If the coefficient of range for a population is small, then it can be said that variability is less in the observations of the population i.e. the values of the observations are not far from each other. But if the coefficient of range is high then it can be said that variability is more in the observations of the population, i.e. the values of the observations are very far from each other.

Illustration 1 The

4.3.4 Standard Deviation

We have seen that the definition of mean deviation is based on absolute values of the deviations of observations of the data from the mean. Since the algebraic signs of the deviations are ignored, mean deviation is less used in advanced study of statistics. This limitation of mean deviation is overcome by an important measure of dispersion known as Standard Deviation. Instead of taking the absolute value of deviation of each observation from the mean, the square of the deviation is taken. If the sum of squares of these deviations is divided by the total number of observations, we get an important measure of dispersion known as Variance. It is denoted by s^2. The positive square root of the variance is called the Standard Deviation. It is denoted by s.

Well known statistician Karl Pearson defined the Standard Deviation as, "Standard Deviation is the positive square root of the mean of the squares of the deviations measured from the mean."

After mean, standard deviation is another very useful measure which gives information about values of the observations of a population.

Note that the standard deviation is an absolute measure of dispersion. If the standard deviation is divided by the mean of the data, we get its relative measure of dispersion. It is called the **coefficient of standard deviation.**

$\therefore$ Coefficient of standard deviation $= \frac{s}{\bar{x}}$

Note : Among all the measures of dispersion, standard deviation is the most important and widely

List of Formulae :

	Measure of Dispersion	Absolute Measure	Relative Measure
1.	**Range**	$R = x_H - x_L$	Coefficient of Range $= \frac{x_H - x_L}{x_H + x_L}$
2.	**Quartile Deviation**	$Q_d = \frac{Q_3 - Q_1}{2}$	Coefficient of Quartile Deviation $= \frac{Q_3 - Q_1}{Q_3 + Q_1}$
3.	**Mean Deviation**	$MD = \frac{\Sigma\lvert x_i - \bar{x}\rvert}{n}$ (For Ungroped Data) $MD = \frac{\Sigma f\lvert x - \bar{x}\rvert}{n}$ (For Groped Data)	Coefficient of Mean Deviation $= \frac{MD}{\bar{x}}$
4.	**Standard Deviation**	**For Ungrouped Data :** $s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}$ OR $\sqrt{\frac{\Sigma x^2}{n} - \bar{x}^2}$ **Short-cut Method :** $s = \sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2}$ **For Grouped Data :** $s = \sqrt{\frac{\Sigma f(x-\bar{x})^2}{n}}$ OR $\sqrt{\frac{\Sigma f x^2}{n} - \bar{x}^2}$ **Short-cut Method :** When, $d = x - A$ $s = \sqrt{\frac{\Sigma f d^2}{n} - \left(\frac{\Sigma f d}{n}\right)^2}$ When, $d = \frac{x - A}{c}$ $s = \sqrt{\frac{\Sigma f d^2}{n} - \left(\frac{\Sigma f d}{n}\right)^2} \times c$	Coefficient of standard Deviation $= \frac{s}{\bar{x}}$ Coefficient of Variation $= \frac{s}{\bar{x}} \times 100$
5.	**Combined Standard Deviation** $s_c = \sqrt{\frac{n_1(s_1^2 + d_1^2) + n_2(s_2^2 + d_2^2)}{n_1 + n_2}}$		

2.2 Meaning and Definition of Linear Correlation

Let us first understand the meaning of correlation. We know that in many situations, simultaneous changes are seen in the values of two variables. The simultaneous changes in the values of two variables are mainly due to the following reasons.

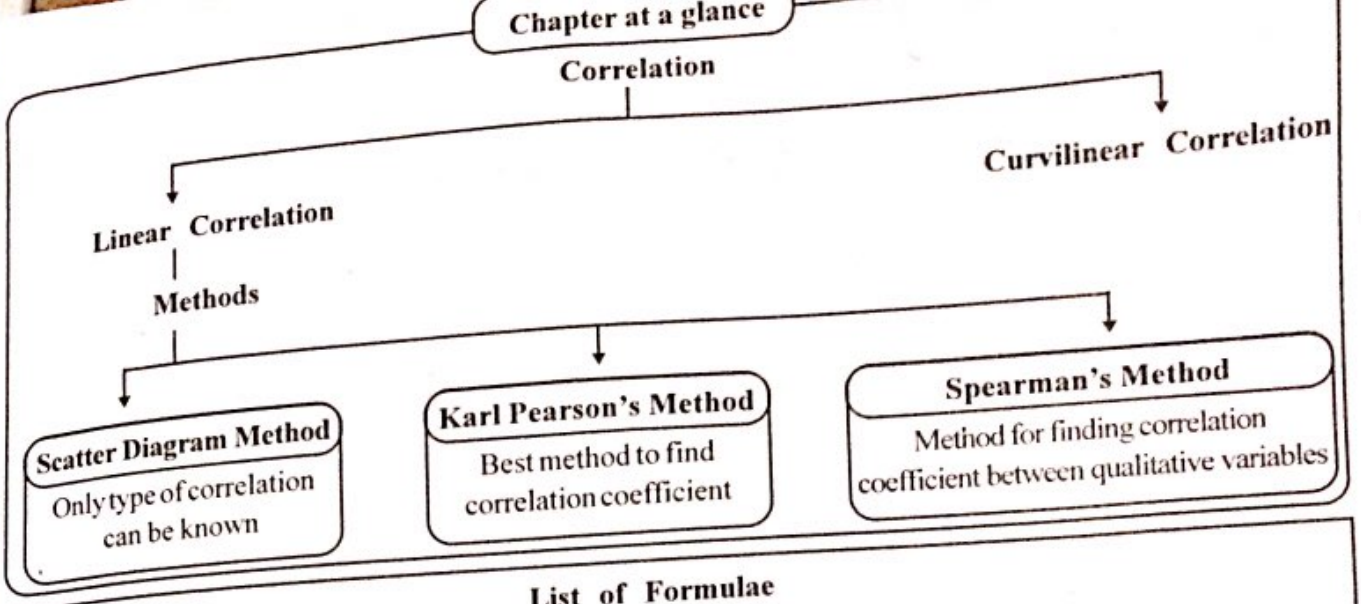
(1) There is a cause-effect relation between two variables.

(2) The values of two variables change due to the effect of some other factor.

In case of yearly rainfall and yield of rice of a region, usually if rainfall increases (up to some extent), yield of rice also increases and if rainfall decreases, yield of rice also decreases. So, 'rainfall' is a 'cause' and 'yield of rice' is an 'effect'. Similarly, when the income of a person remains more or less same, if his expenditure increases, saving decreases and if his expenditure decreases, saving increases. 'Expenditure' is a 'cause' here and 'savings' is the 'effect'. The changes in two variables in the above two examples indicate cause-effect relationship. Sometimes both the variables may be mutually dependent and therefore neither can be specifically said as the 'cause' and the other the 'effect'. Generally, it happens in case of economic variables. e.g. demand and supply. If demand increases, it is necessary to increase supply (which is not always possible instantly) and when supply increases, price tend to decrease and because of that demand goes up. Thus, demand and supply are interdependent. The ages of husband and age of wife is also an example of such situation.

In case of sale of raincoats and sale of rain shoes, the values of both the variables increase in monsoon. There is no direct cause-effect relation here between two variables but the changes in the sale of raincoats and rain shoes are observed due to the presence of the third variable, namely the monsoon. This is an example of indirect cause-effect relationship.

Chapter at a glance



List of Formulae

Karl Pearson's Method :

Correlation coefficient $= r$

(1) $r = \dfrac{\text{Covariance}}{(\text{S.D of } X)(\text{S.D of } Y)} = \dfrac{Cov(X,Y)}{s_x \cdot s_y}$

Where, $Cov(X,Y) = \dfrac{\Sigma(x-\bar{x})(y-\bar{y})}{n} = \dfrac{\Sigma xy - n\bar{x}\,\bar{y}}{n}$

$s_x = \sqrt{\dfrac{\Sigma(x-\bar{x})^2}{n}}$ and $s_y = \sqrt{\dfrac{\Sigma(y-\bar{y})^2}{n}}$

(2) $r = \dfrac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2} \cdot \sqrt{\Sigma(y-\bar{y})^2}}$

(3) $r = \dfrac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2} \cdot \sqrt{n\Sigma y^2 - (\Sigma y)^2}}$

(4) $r = \dfrac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{n\Sigma u^2 - (\Sigma u)^2} \cdot \sqrt{n\Sigma v^2 - (\Sigma v)^2}}$ Where, $u = x - A$ or $\dfrac{x-A}{c_x}$, $v = y - B$ or $\dfrac{y-B}{c_y}$

(5) $r = \dfrac{\Sigma(x-\bar{x})(y-\bar{y})}{n \cdot s_x \cdot s_y}$

(6) $r = \dfrac{\Sigma xy - n\bar{x}\,\bar{y}}{n \cdot s_x \cdot s_y}$

Specially for short sums

Spearman's Rank Correlation Method

(7) $r = 1 - \dfrac{6\Sigma d^2}{n(n^2-1)}$ When the observations are not repeated

(8) $r = 1 - \dfrac{6\left[\Sigma d^2 + CF\right]}{n(n^2-1)}$ When some of the observations are repeated

Where, d = Rank of x – Rank of y = $R_x - R_y$

CF = Correction Factor $= \Sigma\left(\dfrac{m^3-m}{12}\right)$

m = Number of times an observation is repeated

of regression under the assumption that ...

3.2 Linear Regression Model

A set of one or more equations representing a relation or a problem is called a model. A statistical model which describes the cause and effect relationship between two variables is called a regression model. Generally, out of two variables having cause-effect relationship, the causal variable is denoted by X. We shall call this variable as independent or explanatory variable and effect variable is denoted by Y. We shall call this variable as dependent or explained variable. Let us understand the meaning of independent variable and dependent variable from the following illustrations :

(i) In case of 'advertisement cost' and 'sales', generally, because of increase (decrease) in the 'advertisement cost', corresponding 'sales' also increases (decreases), so we shall take 'advertisement cost' as independent variable X and 'sales' as dependent variable Y.

(ii) In case of 'rainfall' and 'yield of rice' in some region, it is very clear that 'yield of rice' depends on 'rainfall'. So, we shall take 'rainfall' as independent variable X and 'yield of rice' as dependent variable Y.

In a regression model, the dependent variable Y is expressed in the form of an appropriate mathematical function of the independent variable X.

Now, we shall define a linear regression model as follows.

$$Y = \alpha + \beta X + u$$

Where, Y = Dependent Variable

X = Independent Variable

α = Constant

β = Constant

u = Disturbance Variable of the Model

List of Formulae :

Equation of Regression Line

$\hat{y} = a + bx$

Where, $b = b_{yx}$ = Regression Coefficient

(1) $b = \dfrac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}$

(2) $b = \dfrac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$

(3) $b = \dfrac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma u^2 - (\Sigma u)^2}$ Here, $u = x - A$ and $v = y - B$

(4) $b = \dfrac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma u^2 - (\Sigma u)^2} \times \dfrac{c_y}{c_x}$ Here, $u = \dfrac{x-A}{c_x}$ and $v = \dfrac{y-B}{c_y}$

(5) $b = r \cdot \dfrac{s_y}{s_x}$

(6) $b = \dfrac{Cov(x, y)}{s_x^2}$

(7) $a = \bar{y} - b\bar{x}$

(8) Coefficient of Determination $R^2 = \left[r(y, \hat{y})\right]^2 = \left[r(x, y)\right]^2 = r^2$