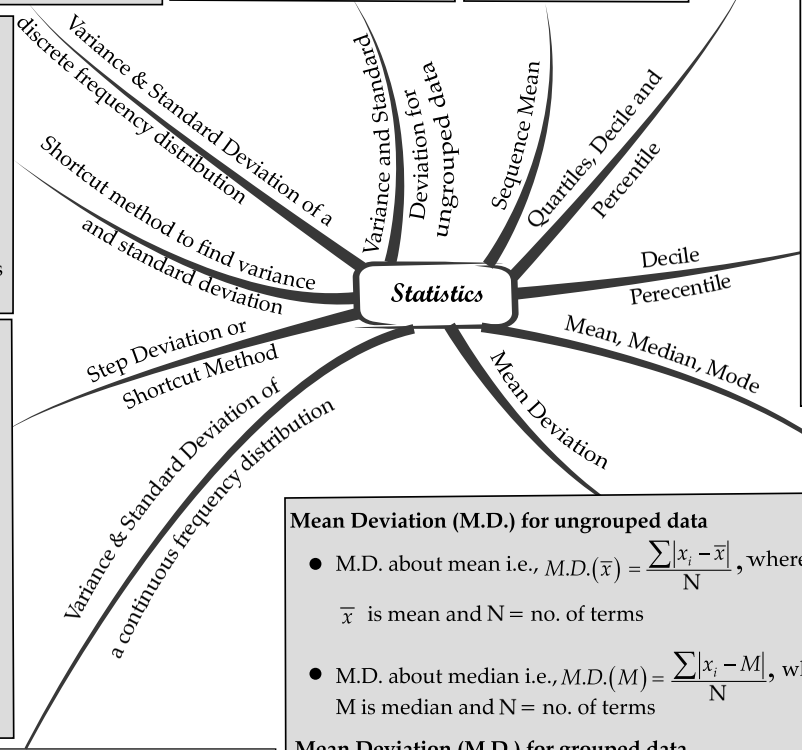
CHAPTER - 22 STATISTICS



Variance of a discrete frequency distribution

$$\text{Var}\left(\sigma^2\right)=\frac{1}{N}\sum f_i\left(x_i-\bar{x}\right)^2$$

Standard Deviation (S.D.) of a discrete frequency distribution

$$\text{S.D}\left(\sigma\right)=\sqrt{\frac{1}{N}\sum f_i\left(x_i-\bar{x}\right)^2}$$

where $N=\sum_{i=1}^{n} f_i$ and $\bar{x}$ = mean

Variance of ungrouped data

$$\text{Var}\left(\sigma^2\right)=\frac{1}{n}\sum\left(x_i-\bar{x}\right)^2$$

Standard Deviation (S.D.) for ungrouped data:

$$\text{S.D.}\left(\sigma\right)=\sqrt{\frac{1}{n}\sum\left(x_i-\bar{x}\right)^2}$$

- Geometric Mean (G) = antilog $\dfrac{\sum_{i=1}^{n} f_i \log x_i}{\sum_{i=1}^{n} f_i}$
- Harmonic Mean (H) = $\dfrac{\sum_{i=1}^{n} f_i}{\sum_{i=1}^{n} \frac{f_i}{x_i}}$

Quartiles : $Q_1=l+\dfrac{\frac{N}{4}-c}{f}\times h$, $Q_3=l+\dfrac{\frac{3N}{4}-c}{f}\times h$

Inter quartile range = Q_3-Q_1

Coeff. of quartile deviation = $\dfrac{Q_3-Q_1}{Q_3+Q_1}$

Decile: $D_i=\dfrac{i\,(n+1)}{10}$ team

Percentile = $P_i=100\,\dfrac{(i-0.5)}{n}$

Variance

$$\left(\sigma^2\right)=\frac{h^2}{N^2}\left[N\sum f_i y_i^{\,2}-\left(\sum f_i y_i\right)^2\right]$$

Standard Deviation

$$\sigma=\frac{h}{N}\sqrt{N\sum f_i y_i^{\,2}-\left(\sum f_i y_i\right)^2}$$

where $y_i=\dfrac{x_i-A}{h}$, $N=\sum_{i=1}^{n} f_i$

A = assumed mean, h = width of class interval

- $M.D.(\bar{x})=\dfrac{1}{N}\sum f_i\left|x_i-\bar{x}\right|$

where $\bar{x}=a+\dfrac{\sum_{i=1}^{n} f_i d_i}{N}\times h$

and $d_i=\dfrac{x_i-a}{h}$

Here, a = assumed mean

h = height

N = sum of frequencies

- $M.D.(M)=\dfrac{1}{N}\sum f_i\left|x_i-M\right|$

where, M(median) $=l+\dfrac{\frac{N}{2}-C}{f}\times h$

Karl Pearson's Coeff. of correlation

$$r=\frac{\text{Cov}\,(x,\,y)}{\sqrt{\text{Var}\,(x)\ \text{Var}\,(y)}}=\frac{n\sum x_i y_i-\sum x_i\sum y_i}{\sqrt{n\sum x_i^2-\left(\sum x_i\right)^2}\sqrt{n\sum y_i^2-\left(\sum y_i\right)^2}}$$

Here, $\text{Cov}(x,\,y)=\frac{1}{n}\sum x_i y_i-\left(\frac{1}{n}\sum x_i\right)\left(\frac{1}{n}\sum y_i\right)$

Regression Equations

Regression equation of x on y

$x-\bar{x}=r\frac{\sigma_x}{\sigma_y}(y-\bar{y})=b_{xy}\,(y-\bar{y})$

Regression equation of y on x.

$(y-\bar{y})=r\frac{\sigma_y}{\sigma_x}(x-\bar{x})=b_{yx}\,(x-\bar{x})$

Angle between two regression lines

$\theta=\tan^{-1}\left(\frac{1-r^2}{b_{xy}+b_{yx}}\right)=\tan^{-1}\left(\frac{1-r^2}{r}\right)\left(\frac{\sigma_x\times\sigma_y}{\sigma_x^2+\sigma_y^2}\right)$

Mode (for grouped data) $M=L+\dfrac{f_1-f_0}{2f_1-f_0-f_2}\times h$

Relation among Mean, Median and Mode

3 Median = 2 Mean + Mode

Mean Deviation (M.D.) for ungrouped data

- M.D. about mean i.e., $M.D.(\bar{x})=\dfrac{\sum\left|x_i-\bar{x}\right|}{N}$, where $\bar{x}$ is mean and N = no. of terms
- M.D. about median i.e., $M.D.(M)=\dfrac{\sum\left|x_i-M\right|}{N}$, where M is median and N = no. of terms

Mean Deviation (M.D.) for grouped data

M.D. about mean i.e., $M.D.(\bar{x})=\dfrac{\sum f_i\left|x_i-\bar{x}\right|}{N}$

M.D. about median i.e., $M.D.(M)=\frac{1}{N}\sum f_i\left|x_i-M\right|$

Variance of a continuous frequency distribution

$\text{Var}\,(\sigma^2)=\frac{1}{N}\sum f_i\,(x_i-\bar{x})^2$

Standard deviation of a continuous frequency distribution

$\text{S.D}\,(\sigma)=\frac{1}{N}\sqrt{N\sum f_i x_i^2-(\sum f_i x_i)^2}$

CHAPTER

22 STATISTICS

Chapter Objectives

Measures of Dispersion : Range & Mean deviation, Variance and standard deviation of ungrouped / grouped data, The Median, Quartiles, Decile, Percentile, Mode of grouped and ungrouped data, analysis of frequency distributions with equal means but different variances. Correlation, types of correlation, Covariance, Karl Pearson Coefficient of correlation, Regression and its analysis.

STUDY MATERIAL

I. Concept Clarified

1. Some important definitions

1. Limit

The starting value of each class is called lower limit and end value of each class is called upper limit.

2. Class interval

The difference between upper and lower limit of the class is called class interval or size of the class.

3. Frequency

Frequency is the maximum number of occurrence of the data. In other words the frequency (or absolute frequency) of an event i is the number n_i of times the event occurred in an experiment.

4. Cumulative frequency

The cumulative frequency is the total of the absolute frequencies of all events at or below a certain point in an ordered list of events

5. Variable or Variate

Characteristics that varies in magnitude from observation to observation. For example- height, weight, age, income e.t.c.

6. Primary and Secondary Data

Primary data refers to the first hand data gathered by the researcher himself whereas secondary data means data collected by someone else earlier.

7. Discrete frequency distribution

Discrete data is calculated by counting, exactly each and every observation.

When an observation is repeated, it is counted. The number for which the observation is repeated is called the frequency of that observation.

8. Continuous frequency distribution

A frequency distribution in which the data is arranged in group which are not exactly measurable is called continuous frequency distribution.

9. Cumulative frequency distribution

Cumulative frequency is defined as the running total of frequencies. It is the sum of all the previous frequencies up to the current point.

$$N = \sum_{i=1}^{n} f_i$$

2. Measure of Dispersion

While measures of central tendency are used to estimate "normal" values of a dataset, measures of dispersion are important for describing the spread of the data, or its variation around a central value. Two distinct samples may have the same mean or median, but completely different levels of

variability, or vice versa. A proper description of a set of data should include both of these characteristics. There are various methods that can be used to measure the dispersion of a dataset, each with its own set of advantages and disadvantages.

- **Range**

It is defined as the difference between the largest and smallest sample values. This is one of the simplest measures of variability to calculate and depends only on extreme values and provides no information about how the remaining data is distributed.

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

3. Mean

The mean, or expected value, are a measure of the central tendency either of a probability distribution or of the random variable characterized by that distribution. It is of three types- Arithmetic mean, Geometric mean and harmonic mean.

- **Arithmetic mean(ungrouped data)**

The arithmetic mean (or simply mean) of an ungrouped sample $x_1, x_2, x_3,..., x_n$ is the sum of the sampled values divided by the number of items in the sample. It is usually denoted by $\bar{x}$.

$$\bar{x} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n} = \frac{1}{n}\sum_{i=1}^{i=n} x_i$$

- **Arithmetic mean for discrete frequency distribution**

The arithmetic mean (or simply mean) of a grouped sample $x_1, x_2, x_3, ... x_n$ with frequencies $f_1, f_2, f_3,.....f_n$ respectively is obtained by

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + f_3x_3 + ... + f_nx_n}{f_1 + f_2 + f_3 + ... + f_n} = \frac{\sum_{i=1}^{n} f_ix_i}{\sum_{i=1}^{n} f_i}$$

- **Arithmetic mean(grouped data)**

By direct method long cut $\bar{x} = \frac{\sum_{i=1}^{n} f_ix_i}{\sum_{i=1}^{n} f_i}$

By shortcut method $\bar{x} = x + \frac{\sum_{i=1}^{n} f_id_i}{\sum_{i=1}^{n} f_i}$, here, $d_i = x_i - x$ and x is assumed mean

By step Deviation Method $\bar{x} = x + \left(\frac{\sum_{i=1}^{n} f_iu_i}{\sum_{i=1}^{n} f_i}\right)h$

Here, x – assumed mean u_i – deviation $= \frac{x_i - x}{h}$ and h is the class width

By Weighted arithmetic mean $\bar{x}_w = \left(\frac{\sum_{i=1}^{n} w_ix_i}{\sum_{i=1}^{n} w_i}\right)$

Where w_1, w_2, w_3, w_n are weights assigned to the values $x_1, x_2, x_3, ... x_n$ respectively.

- **Geometric Mean**

The geometric mean is an average that is useful for sets of positive numbers that are interpreted according to their product and not their sum (as is the case with the arithmetic mean). It is usually denoted by G.

$$G = (x_1, x_2, x_3 \ldots\ldots x_n)^{\frac{1}{n}}$$

If $f_1, f_2, f_3, \ldots\ldots f_n$ be the corresponding frequencies of non-zero observations $x_1, x_2, x_3, \ldots\ldots x_n$ then geometric mean is obtained as :

$$\text{G=antilog of} \left(\frac{\sum_{i=1}^{n} f_i \log x_i}{\sum_{i=1}^{n} f_i} \right)$$

➢ **Harmonic Mean**

The harmonic mean is an average which is useful for sets of numbers which are defined in relation to some unit, for example speed (distance per unit of time).

$$H = n \sum_{i=1}^{n} \left(\frac{1}{x_i} \right)^{-1}$$

If f_1, f_2, f_3, be the corresponding frequencies of non-zero observations $x_1, x_2, x_3, \ldots\ldots x_n$ then their harmonic mean is obtained as :

$$H = \frac{f_1 + f_2 + f_3 + \ldots\ldots\ldots + f_n}{\left(\frac{f_1}{x_1} + \frac{f_2}{x_2} + \frac{f_3}{x_3} \ldots\ldots\ldots + \frac{f_n}{x_n} \right)} = \frac{\sum_{i=1}^{n} f_i}{\sum_{i=1}^{n} \frac{f_i}{x_i}}$$

4. Mean deviation

It is defined as how far a mean is from the middle in the given set of data. It is obtained using following steps :

1. Find mean

2. Calculate the distance from the mean of each data from middle. Distance will be obtained by subtracting mean from each value ignoring minus sign.

3. Find the mean of the distances

Example : Find the mean deviation of given data - 2, 3, 4, 4, 4, 5, 6, 6, 8, 8, 9, 9, 10.

Mean $\bar{x} = \frac{2+3+4+4+4+5+6+6+8+8+9+9+10}{13} = 6$

Value	Distance from 6
2	4
3	3
4	2
4	2
4	2
5	1
6	0
6	0
8	2
8	2
9	3
9	3
10	4

Mean of distance = mean deviation

$$= \frac{4+3+2+2+2+1+0+0+2+2+3+3+4}{13} = 2.15$$

So the mean is 6 and mean deviation is 2.15

➢ **Variance**

The variance of a random variable x is the expected value of the squared deviation from the mean of x. In other words it is the average of squared differences from the mean.

$$\text{Variance} = \frac{1}{n} \sum_{i=1}^{i=n} (x_i - \bar{x})^2$$

Example : Find the variance of given data : 3,4,5,2,8,6,7.

Mean of given data $\bar{x} = \frac{3+4+5+2+8+6+7}{7} = 5$

Value	Distance from 5 $(x_i - \bar{x})$
3	–2
4	–1
5	0
2	–3
8	+3
6	+1
7	+2

$$\text{Variance} = \frac{(-2)^2 + (-1)^2 + (-3)^2 + 3^2 + 1^2 + 2^2}{7} = 4$$

Here mean is 5 and variance is 4.

Important :

Variance is square times of standard deviation.

➢ Standard Deviation of ungrouped data

Standard deviation of ungrouped data is defined as the square root of variance when all the elements are different. It is denoted by σ.

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

Example : Find the standard deviation of given data : 11, 13, 12, 14, 12, 15, 16, 19.

Mean $\bar{x} = \frac{11+13+12+14+12+15+16+19}{8} = 14$

Value	Distance from 14 $(x_i - \bar{x})$	$(x \quad \bar{x})$
11	–3	9
13	–1	1
12	–2	4
14	0	0
12	–2	4
15	+1	1
16	+2	4
19	+5	25

$$\text{Variance} = \frac{(-3)^2 + (-1)^2 + (-2)^2 + 0^2 + (-2)^2 + (+1)^2 + (+2)^2 + (+5)^2}{8}$$

$$= \frac{9+1+4+0+4+1+4+25}{8}$$

$$\text{Variance } \sigma^2 = 6$$

So standard deviation $\sigma = 2.44$

➢ Standard deviation of grouped data

Standard deviation of grouped data is defined as the square root of variance when few of the elements are same and few are different. It is also denoted by σ.

$$\text{Standard Deviation } \sigma = \sqrt{\frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})^2}{\sum_{i=1}^{n} f_i}}$$

Example : Find the standard deviation of tabulated data :

Data x_i	Frequency f
4	5
6	2
5	8
9	4
7	8
8	6

Solution : Mean of grouped data $\overline{x} = \dfrac{\sum_{i=1}^{i=n} f_i x_i}{\sum_{i=1}^{n} f_i}$

Data x_i	Frequency f_i	$f_i x_i$
4	5	20
6	2	12
5	8	40
9	4	36
7	8	56
8	6	48
	$\sum_{i=1}^{n} f_i = 33$	$\sum_{i=1}^{i=n} f_i x_i = 212$

$$\overline{x} = \frac{212}{33} = 6.42$$

Value	Distance from 6.42 $(x_i - \overline{x})$	$(x_i - \overline{x})^2$	Frequency f_i	$f_i (x_i - \overline{x})^2$
4	–2.42	5.85	5	29.25
6	–0.42	0.17	2	0.34
5	–1.42	2.01	8	16.08
9	2.58	6.65	4	26.6
7	0.58	0.34	8	2.72
8	1.58	2.49	6	14.94
			$\sum f_i = 33$	$\sum f_i (x_i - \overline{x})^2 = 89.93$

Standard Deviation $\sigma = \sqrt{\dfrac{\sum_{i=1}^{n} f_i (x_i - \overline{x})^2}{\sum_{i=l}^{n} f_i}}$

Standard Deviation $\sigma = \sqrt{\dfrac{89.93}{33}} = 1.64$

5. The Median

The median is the value separating the higher half from the lower half of a data sample or in other words the median of a set of data is the middle number in the set. To find the median the data must be arranged in ascending order first.

Cases :

1. **I**f the number of sampling data is odd then median will be $\left(\frac{n+1}{2}\right)^{th}$ data, where n is total no. of sampling data.

2. If the number of sampling data is even then median will be mean of $\left(\frac{n}{2}\right)^{th}$, $\left(\frac{n}{2}+1\right)^{th}$ data, where n is total number of sampling data.

➢ Median for grouped frequency

For a continuous distribution, median is

$$\text{Median} = l + \frac{\frac{N}{2} - C}{f} \times h$$

Where l – lower limit of median class

f – frequency of the median class

N – total number frequency = $\sum_{i=1}^{n} f_i$

C – Cumulative frequency of the class just before the median class

h – length of the median class

6. Quartiles :

Quartiles in statistics are values that divide the data into quarters. The quartile divides data into four segments according to where the numbers fall on the number line. The four quarters that divide a data set into quartiles are :

1 – The lowest 25% of numbers.

2 – The next lowest 25% of numbers (up to the median).

3 – The second highest 25% of numbers (above the median).

4– The highest 25% of numbers.

Quartiles for a continuous distribution is given by :

$$Q_1 = l + \frac{\frac{N}{4} - C}{f} \times h$$

$$Q_2 = l + \frac{\frac{N}{2} - C}{f} \times h$$

$$Q_3 = l + \frac{\frac{3N}{4} - C}{f} \times h$$

Inter quartile range = $Q_3 - Q_1$

Coefficient of quartile deviation = $\frac{Q_3 - Q_1}{Q_3 + Q_1}$

f – frequency of the first quartile class

l – lower limit of the first quartile class

C – Cumulative frequency corresponding to the class just before the first quartile of class

h – length of the first quartile

➢ Method to find Quartile

Use the following rules to find quartile :

1. Use the median to divide the ordered data set into two halves.

2. Do not include the median (the central value in the ordered list) in either half if there are an odd number of data points in the original ordered data set.

3. Divide the data set exactly in half if there is an even number of data points in the original ordered data set,
4. The lower quartile value is the median of the lower half of the data. The upper quartile value is the median of the upper half of the data.

7. Decile

A decile is any of the nine values that divide the sorted data into ten equal parts, so that each part represents 1/10 of the sample or population. In other words a decile is a quantitative method of splitting up a set of ranked data into 10 equally large subsections.

Decile is obtained by the following method $D_i = \frac{i(n+1)}{10^{th}\text{ data}}$

Where D_i – decile, n – total number of data

D_2 = value of $\frac{2(n+1)^{th}}{10}$ term, D_9 = value of $\frac{9(n+1)^{th}}{10}$ term

8. Percentile

A percentile is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations falls.

The meaning of percentile can be captured by stating that the p^{th} percentile of a distribution is a number such that approximately p percent ($p\%$) of the values in the distribution are equal to or less than that number. So, if '20' is the 90th percentile of a larger batch of numbers, 90% of those numbers are less than or equal to 20.

To calculate percentiles, sort the data so that x_1 is the smallest value, and xn is the largest, with n = total number of observation.

$$p_i = 100\frac{(i-0.5)}{n}$$

Where p_i is percentile of i^{th} data.

Example : Calculate percentile of 5, 1, 6, 4, 10, 20.

Arrange in ascending order 1,4, 5, 6, 10, 20.

$$p_1 = 100\frac{(1-0.5)}{6} = 8.33$$

$$p_2 = 100\frac{(2-0.5)}{6} = 25$$

$$p_3 = 100\frac{(3-0.5)}{6} = 41.66$$

$$p_4 = 100\frac{(4-0.5)}{6} = 58.33$$

$$p_5 = 100\frac{(5-0.5)}{6} = 75$$

$$p_6 = 100\frac{(6-0.5)}{6} = 91.66$$

x_i	1	4	5	6	10	20
i	1	2	3	4	5	6
p_i	8.33	25	41.33	58.33	75	91.33

9. Mode of grouped and ungrouped data

The mode of a set of data is simply the value that appears most frequently in the set. In other words it is the value about which the observations are tend to be most heavily concentrated.

1. Mode of simple ungrouped data is the value which is repeated maximum number of times.
2. Mode of grouped data is given by :

$$M = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$$

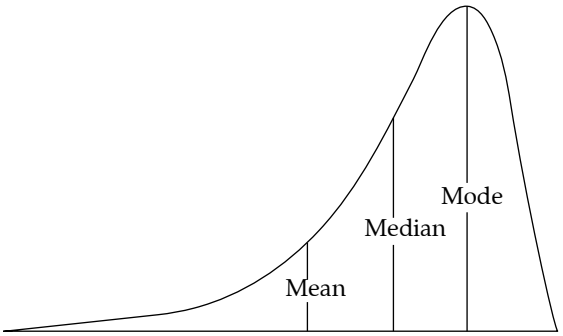
l = lower limit of the modal class
f_1 - frequency of the modal class
f_0 - frequency of the pre-modal class
f_2 - frequency of the post - modal class
h - length of the class interval

Note :

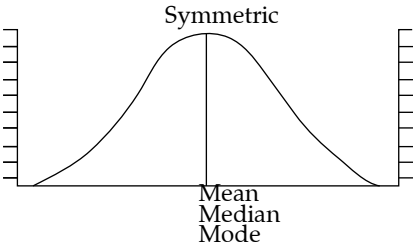
1. The class having maximum frequency is called the modal class.
2. Middle point of the model class is called the crude mode .
3. The class just before the model class is call pre-modal class.
4. The class just after the model class is called post modal class.

➢ **Comparison of Mean, Median and Mode graphically**

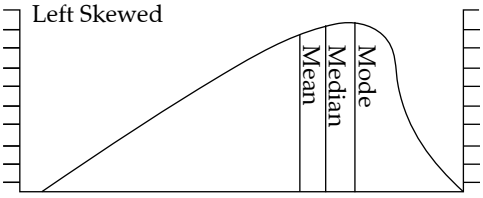
1. Mode is the maximum no. of occurrence and it is greatest among the mean, median and mode while value of median lies between mean and mode.



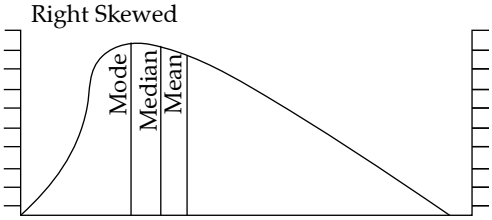
2. A distribution is said to be symmetric if the same number of frequency is found to be distributed at same linear distance on the either side of the mode.



3. If the frequency increase slowly till modal value and decreases rapidly after model value, then skewness is said to be negative and is left skewed.



4. If the frequency increase sharply till modal value and decreases slowly after modal value, then skewness is said to be positive and is right skewed.



➢ **Relation between Mean, Median and Mode**

(i) Mean - Mode = 3(Mean – Median)
(ii) Mode = 3 Median – 2 Mean
(iii) 3 Median = 2 Mean + Mode

10. Correlation

Correlation is a statistic that measures the degree to which two variables move in relation to each other. For example, height and weight are related; taller people tend to be heavier than shorter people. The relationship isn't perfect. People of the same height vary in weight.

Like all statistical techniques, correlation is only appropriate for certain kinds of data. Correlation works for quantifiable data in which numbers are meaningful, usually quantities of some sort. It cannot be used for purely categorical data, such as gender, brands purchased, or favourite colour.

➢ **Types of correlation**

1. Perfect Correlation

When the variables vary in such a way that their ratio is always constant, then this correlation is said to be perfect correlation.

2. Positive or Direct Correlation

If an increase or decrease in one variable results in increase or decrease of other *i.e.* variable are related in direct relation, then the correlation is called positive or direct correlation.

3. Negative or Indirect Correlation

If an increase or decrease in one variable results in decrease or increase of other *i.e.* variable are related in indirect relation, then the correlation is called negative or indirect correlation.

➢ **Properties of Correlation**

1. Coefficient of Correlation lies between -1 and +1, *i.e.* $-1 \le r \le +1$
The coefficient of correlation cannot take value less than -1 or more than +1.
2. If value of coefficient of correlation is 1 then it is perfectly positive.
3. If value of coefficient of correlation is -1 then it is perfectly negative.
4. If x and y are independent variable, then $r = 0$.
5. Coefficients of Correlation are independent of Change of Origin
This property reveals that if we subtract any constant from all the values of x and y, it will not affect the coefficient of correlation.
6. Coefficients of Correlation possess the property of symmetry.
7. Coefficient of Correlation is independent of Change of Scale.
This property reveals that if we divide or multiply all the values of x and y, it will not affect the coefficient of correlation.
8. Coefficient of correlation measures only linear correlation between x and y.
9. If r is the correlation coefficient in the sample of n pair of observations then,

$$\text{Standard error} = \frac{1-r^2}{\sqrt{n}}$$

$$\text{Probable error} = 0.6745\ \frac{1-r^2}{\sqrt{n}}$$

➢ **Karl Pearson's Coefficient of Correlation**

Karl Pearson's Coefficient of Correlation is widely used mathematical method wherein the numerical expression is used to calculate the degree and direction of the relationship between linear related variables. The coefficient of correlation is denoted by "r".

(i) The correlation coefficient r(x , y) between the variables (having small number) x and y is obtained by

$$r = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x).\text{var(y)}}}$$

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

(ii) If $(x-\overline{x})$, $(y-\overline{y})$ are small non- fractional number, we use

$$r = \frac{\Sigma(x-\overline{x})(y-\overline{y})}{\sqrt{\Sigma(x-\overline{x})^2}\sqrt{\Sigma(y-\overline{y})^2}}$$

(iii) If we consider assumed means A and B, where $u = x$ – A and $v = y$ – A them,

$$r = \frac{n\sum_{i=1}^{n} u_i v_i - \sum_{i=1}^{n} u_i \sum_{i=1}^{n} v_i}{\sqrt{n\sum_{i=1}^{n} u_i^2 - \left(\sum_{i=1}^{n} u_i\right)^2}\sqrt{n\sum_{i=1}^{n} v_i^2 - \left(\sum_{i=1}^{n} v_i\right)^2}}$$

➢ Covariance

Covariance provides a measure of the strength of the correlation between two or more sets of random variates. The covariance for two random variates x and y, each with sample size n, is defined by the expectation value.

If (x_i, y_i) be a bivariate distribution such that x_1, x_2, x_3, x_n is sampled values of variable x and y_1, y_2 y_3, y_n is the sampled values of variable y then their covariance of (x, y) is given by :

1. Covariance of $(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$, where $\overline{x}$ and $\overline{y}$ are the mean of variable x and y.

2. Covariance of $(x, y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)$

11. Regression

Regression is a technique for determining the statistical relationship between two or more variables where a change in a dependent variable is associated with, and depends on, a change in one or more independent variables. In other words Regression is a technique used to model and analyse the relationships between variables and often times how they contribute and are related to producing a particular outcome together.

➢ Types of regression

1. Linear Regression
2. Logistic Regression
3. Polynomial Regression
4. Stepwise Regression
5. Ridge Regression
6. Lasso Regression
7. Elastic Net Regression

➢ Lines of regression

A linear regression line has an equation of the form $y = a + bx$, where x is the independent variable and y is the dependent variable. The slope of the line is 'b' and 'a' is the intercept (the value of y when $x = 0$).

➢ Regression coefficients

Regression coefficient shows that a small change made in value of y or x variables, what will be the resultant average change in the value of x or y variables. It is denoted by b_{xy} or b_{yx}

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} \quad \text{and} \quad b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

➢ **Regression Equations**

1. $y-\overline{y}=r\frac{\sigma_y}{\sigma_x}(x-\overline{x})$, this is a line of regression of y on x

2. $x-\overline{x}=r\frac{\sigma_x}{\sigma_y}(y-\overline{y})$, this is a line of regression of x on y

3. $\theta=\tan^{-1}\left(\frac{1-r^2}{b_{xy}+b_{yx}}\right)=\tan^{-1}\left[\left(\frac{1-r^2}{r}\right)\left(\frac{\sigma_x\sigma_y}{\sigma_x^{\,2}+\sigma_y^{\,2}}\right)\right]$, where θ is angle between two regression lines.

Note :

1. When $r = 0$, then two regression lines are mutually perpendicular and $\theta=\frac{\pi}{2}$.

2. When r = 1 or –1 then two regression lines coincides and $\theta=0$

➢ **Properties of regression Coefficients**

1. The correlation coefficient is the geometric mean of two regression coefficients.

2. The value of the coefficient of correlation cannot exceed unity *i.e.* 1. Therefore, if one of the regression coefficients is greater than unity, the other must be less than unity.

3. The sign of both the regression coefficients will be same, *i.e.* they will be either positive or negative. Thus, it is not possible that one regression coefficient is negative while the other is positive.

4. The coefficient of correlation will have the same sign as that of the regression coefficients, such as if the regression coefficients have a positive sign, then "*r*" will be positive and vice-versa.

5. The arithmetic mean of the two regression coefficients will be greater than or equal to the value of the correlation. Symbolically it can be represented as

$$\frac{b_{yx}+b_{xy}}{2}\geq r$$

6. The regression coefficients are independent of the change of origin, but not of the scale.

II. Important Formulae

1. Range = $x_{max}-x_{min}$

2. Mean = $\overline{x}=\frac{\sum_{i=1}^{n}x_i}{n}$

3. Mean of discrete frequency distribution = $\overline{x}=\frac{\sum_{i=1}^{n}f_ix_i}{\sum_{i=1}^{n}f_i}$

4. By weighted arithmetic mean = $\overline{x}_w=\frac{\sum_{i=1}^{n}w_ix_i}{\sum_{i=1}^{n}w_i}$

5. Median of grouped frequency median = $l+\frac{\frac{N}{2}-C}{f}\times h$

6. Mode of grouped data

$$L+\frac{f_1-f_0}{2f_1-f_0-f_2}\times h$$

7. Geometric mean

$$\mathrm{G}=\left(x_1,x_2,x_3,.....x_{\mathrm{n}}\right)^{\frac{1}{n}}$$

8. Harmonic mean

$$\mathrm{H}=n\sum_{i=1}^{n}\left(\frac{1}{x_i}\right)^{-1}$$

9. Variance $=\frac{1}{n}\sum_{i=1}^{i=n}(x_i-\overline{x})^2$

10. Standard deviation of ungrouped data

$$\sigma=\sqrt{\frac{1}{n}\sum_{i=1}^{i=n}(x_i-\overline{x})^2}$$

11. Standard deviation of grouped data

$$\sigma=\sqrt{\frac{\sum_{i=1}^{n}f(x_i-\overline{x})^2}{\sum_{i=1}^{n}f}}$$

12. Quartile

$$Q_1=l+\frac{\frac{N}{4}-C}{f}\times h,\ Q_2=l+\frac{\frac{N}{2}-C}{f}\times h,\ Q_3=l+\frac{\frac{3N}{4}-C}{f}\times h$$

13. Decile $D_i=\frac{i(n+1)}{10^{th}\text{ data}}$

14. Percentile $p_i=100\frac{(i-0.5)}{n}$

15. Covariance

Covariance of $(x,y)=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})$, where $\overline{x}$ and $\overline{y}$ are the mean of variable x and y.

Covariance of $(x,y)=\frac{1}{n}\sum_{i=1}^{n}x_iy_i-\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)\left(\frac{1}{n}\sum_{i=1}^{n}y_i\right)$

16. Equation of line of regression of y on x $\quad y-\overline{y}=r\frac{\sigma_y}{\sigma_x}(x-\overline{x})$

17. Equation of line of regression of x on y $\quad x-\overline{x}=r\frac{\sigma_x}{\sigma_y}(y-\overline{y})$

18. Angle between two regression of lines $\theta=\tan^{-1}\left(\frac{1-r^2}{b_{xy}+b_{yx}}\right)=\tan^{-1}\left[\left(\frac{1-r^2}{r}\right)\left(\frac{\sigma_x\sigma_y}{\sigma_x^{\ 2}+\sigma_y^{\ 2}}\right)\right]$

19. Regression coefficients $\mathrm{b}_{xy}=r\frac{\sigma_x}{\sigma_y}$ and $\mathrm{b}_{yx}=r\frac{\sigma_y}{\sigma_x}$

20. Karl pearson coefficient $r=\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{\sqrt{n\sum_{i=1}^{n}x_i^{\ 2}-\left(\sum_{i=1}^{n}x_i\right)^2}\sqrt{n\sum_{i=1}^{n}y_i^{\ 2}-\left(\sum_{i=1}^{n}y_i\right)^2}}$

●●