UNIT- IV

MEASURES OF CENTRAL TENDENCY

DEFINITION

Average is a value which is typical or representation of a set of data given by MURRY.R. SPEIGAL.

Average is an attempt to find one single figure to describe the whole of figures. - CLARK & SEKKODS

Average is sometime described as number which is typical of the whole group - LEABO.

FUNCTIONS OF AVERAGES

(1) To facilitate quick understanding of complex Data.

(2) To facilitate comparison

(3) To known about the universe from a sample.

(4) To help in Decision making.

(5) To estimate or Establish mathematical Relationship.

CHARACTERISTICS OF A GOOD AVERAGES

* It should rigidly defined (well defined)
* It should be easy to understand.
* It should be simple to complete.
* It's definition should be in the form of a mathematical formula.
* It should be based on all items in the form of data.
* It should not be influenced by any single item (or)

Group of item.

* It should be capable of further algebraic treatment.
* It should be capable of being used in further statistical computation.
* It should have sampling stability.

TYPES OF AVERAGES

The following are fourtypes of averages:

(i) Arithmetic mean
   a) Simple
   b) weighted
(ii) Median
(iii) Mode
(iv) Geometric mean
(v) Hormonic mean

(i) ARITHMETIC MEAN

Average average is also called as 'Mean'. Arithmetic average of a series is the figure obtained dividing the total value of the various items by their number. There are two types:

a) SIMPLE : * INDIVIDUAL SERIES

- Add up all the values of the variables (x) and find Σx.
- Divide Σx by their no. of observation (n)

$$\therefore \bar{x} = \frac{\Sigma x}{n}$$

b) WEIGHTED : INDIVIDUAL SERIES

The weighted is a number standing for the relative importance of the items weighted average can be defined as an average are components items are multiplied by their respective weights and the aggregate of the product are divided by total of arithmetic average.

$$\therefore \overline{x}_w = \frac{\Sigma wx}{\Sigma w}$$

ARITHMETIC DISCRETE AVEARAGE:

- Multiply each size of item (variable by its frequency) that is Fx.
- Add all the Fx to Σfx divided Σfx by total frequency

$$\therefore \overline{x} = \frac{\Sigma fx}{N}$$

ARITHMETIC CONTINUOUS AVERAGE:

- The Midpoint value must be find.
- Assume any one of midpoint as a assumed mean.
- Find out the deviation of mid value of each class.
- Deviation are divided by a common pattern to obtained 'd'
- Multiply d of each class by its frequency to get fd.
- Add up the product value to ged summation fd.

$$\therefore \overline{x} = A + \left(\frac{\Sigma fd}{N}\right) \times C$$

A - ASSUMED MEAN
N = Σf
d - (M-A)/C
C - WIDTH OF THE CLASS-INTERVAL OF A COMMON FACTOR.

MEDIAN

The median is that value of the variable which divides the group.

INDIVIDUAL SERIES:

Median refers to the middle value in the distribution.

If the total no. of items is an odd figure, the median is the middle value. If the total is even figure, the average of items in the center of the distribution.

DISCRETE SERIES:

- Arrange the data in Ascending order.
- Find the cumulative frequency
- Apply the formula

$(N = \Sigma f)$

$$\text{Median} = \text{Size of } \left(\frac{N+1}{2}\right)^{th} \text{ item}$$

CONTINUOUS SERIES:

- Find out the Median by using $\left(\frac{N}{2}\right)$.
- Find out the class in which Median lies.
- The formula,

$$\text{Median} = L + \frac{\left(\frac{N}{2}\right) - cf}{f} \times C$$

L - lower limit

f - frequency

c.f - cumulative frequency

C - class interval

PARTITIONED VALUES:

Median divides the distribution into two equal parts. There are other values also which the divide the series into equal parts and they are called the partition values.

ONE POINT divides the series (HALVES)
THREE POINT divides the series (QUARTILES)
NINE POINT divides the series (DECILES)
99 POINT divides the series (PERCENTILES)

CALCULATION OF QUARTILES (INDIVIDUAL & DISCRETE SERIES)

- Find out the cummulative frequency.
- Then apply the formula.

FIRST QUARTILE (Q1) = size of $\left(\frac{N+1}{4}\right)^{th}$ item

THIRD QUARTILE (Q3) = size of $\left(\frac{3(N+1)}{4}\right)^{th}$ item

FOR CONTINUOUS SERIES

$$Q_1 = L + \left(\frac{N/4 - c.f}{f}\right) \times c$$

$$Q_3 = L + \left(\frac{\frac{3N}{4} - c.f}{f}\right) \times c$$

DECILES:

$$D_2 = 4 + \left\{\frac{\frac{2N}{10} - f}{f}\right] \times c$$

PERCENTILES:

$$P_{25} = L + \left(\frac{\frac{25N}{100} - c.f}{f}\right) \times c$$

MODE

The Mode of the distribution is the value at the point around which the item tend to most nearly concentrated.

INDIVIDUAL :

Mode can often be found out by which occur more no. of times than the other.

DISCRETE :

Mode is the variable value which corresponds to the maximum frequency.

CONTINUOUS :

First we have to find Modal class and then value of mode by formula,

$$\text{Mode} = L + \frac{(f_1 - f_0)}{(2f_1 - f_0 - f_2)} \times C$$

ASYMPTOTIC MODE :

$$\text{MODE} = 3\,\text{MEDIAN} - 2\,\text{MEAN}$$

Symmetrical = $\bar{x}$ = Median = Mode

Asymmetrical = $\bar{x} \neq$ Median = Mode

GEOMETRIC MEAN

It is defined as the n^{th} root of the product of n times.

$$\therefore G.M. = \sqrt[n]{x_1 x_2 \dots x_n}$$

$$GM = \text{antilog}\left[\frac{\Sigma \log x}{n}\right]$$

DISCRETE :

$$G.M. = \text{Antilog}\left[\frac{\Sigma f \log x}{N}\right]$$

CONTINUOUS :

$$G.M. = \text{Antilog}\left[\frac{\Sigma f \log m}{N}\right]$$

HARMONIC MEAN

Harmonic mean is the reciprocal of the arithmetic average of values of various items in the variable.

INDIVIDUAL :

$$H.M. = \frac{n}{\Sigma \frac{1}{x}}$$

DISCRETE :

$$H.M. = \frac{N}{\Sigma f\left(\frac{1}{x}\right)}$$

CONTINUOUS :

$$H.M. = \frac{N}{\Sigma f \frac{1}{M}}$$

MEASURES OF DISPERSION

DEFINITION

Disporsion is the measure of variation of the items.
— Bowley

Disporsion is a measure of extent to which the individual items - CONNER.

INTER-QUARTILE RANGE AND QUARTILE DEVIATION:

The distance between first & third Quartile is called Inter-Quartile range.

Inter Quartile range $= Q_3 - Q_1$

Semi Inter Quartile range (or) Quartile deviation

$$= \frac{Q_3 - Q_1}{2}$$

co-efficient of Quartile deviation $= \frac{Q_3 - Q_1}{Q_3 + Q_1}$

MEAN DEVIATIONS (OR) AVERAGE DEVIATIONS:

Average deviations is the average amount of scatter of the items in a distribution from either the mean or the Median, ignoring the science of the deviation. The relative measure of deviation of mean is called co-efficient of Mean deviation.

$$\text{co-efficient} = \frac{\text{Mean deviation}}{\text{Mean (or) Median}}$$

INDIVIDUAL:

$$M.D. = \frac{\Sigma |D|}{n}$$

DISCRETE:

$$M.D. = \frac{\Sigma f|D|}{N}$$

CONTINUOUS:

$$M.D. = \frac{\Sigma f|D|}{N}$$

STANDARD DEVIATION:

Standard deviation is also called ROOT MEAN SQUARE DEVIATION or mean error or Mean square error. It is denoted by greek letter σ (sigma).

INDIVIDUAL:

* Deviation taken from actual mean and assumed mean.

by ACTUAL MEAN:

- $x = x - \bar{x}$
- $\sigma = \sqrt{\frac{\Sigma x^2}{N}}$ (or) $\sqrt{\frac{\Sigma (x-\bar{x})^2}{N}}$

by ASSUMED MEAN:

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

Step deviation Method:

$$\sigma = \sqrt{\frac{\Sigma f d'^2}{N} - \left(\frac{\Sigma f d'}{N}\right)^2} \times C$$

COMBINED STANDARD DEVIATION

$$\bar{x}_{12} = \frac{N_1 \bar{x}_1 + N_2 \bar{x}_2}{N_1 + N_2}$$

$$\sigma_{12} = \sqrt{\frac{N_1 \sigma_1^2 + N_2 \sigma_2^2 + N_1 d_1^2 + N_2 d_2^2}{N_1 + N_2}}$$

$$\sigma_{12} = \sqrt{\frac{N_1 (\sigma_1^2 + d_1^2) + N_2 (\sigma_2^2 + d_2^2)}{N_1 + N_2}}$$

For, 'n' natural numbers, the standard deviation can be computed by the formula,

$$\sigma = \sqrt{\frac{1}{12}(n^2 - 1)}$$

VARIANCE : square of standard deviation is called variance

$$\sigma = \sqrt{\text{variance}}$$

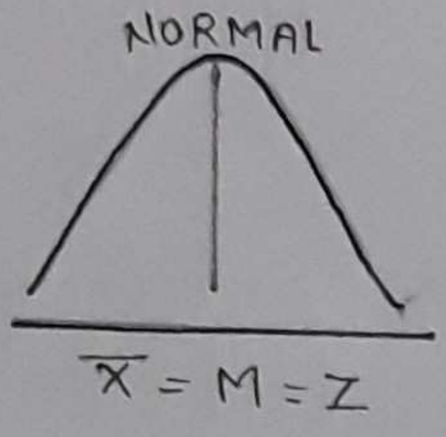
Co-efficient of variation (C.V.) = $\frac{\sigma}{\bar{x}} \times 100$.

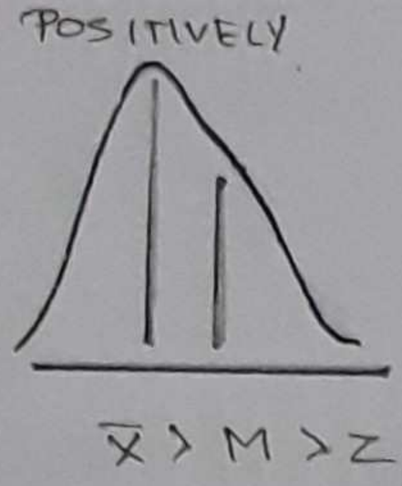
SKEWNESS AND KURTOSIS

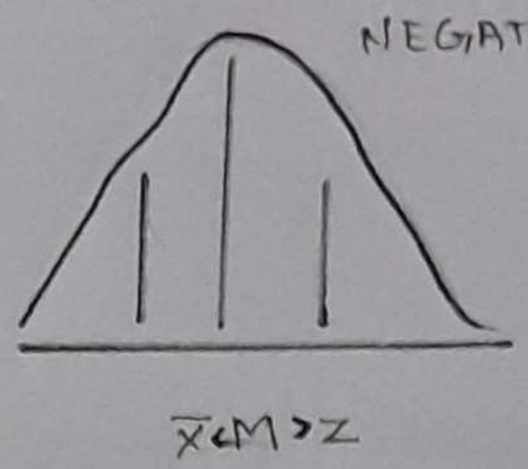
DEFINITION:

Skewness or asymmetry is the attribute of a frequency distribution that extends further on one side of the class with the highest frequency that on the other - SIMPSON & KALFA.

When a series is not symmetrical, it is said to asymmetrical or skewed:-







CHARACTERISTICS

* $\bar{X} \neq$ Median $\neq Z$
* Q_3 - Median $\neq$ Median - Q_1
* sum of +ve deviation $\neq$ sum of -ve deviation.
* The frequencies are unequal on all sides.
* The Plotted graph have unequal halves.

CO-EFFICIENT OF SKEWNESS:

1. KARL PEARSON $Sk_p = \frac{\bar{X} - Mode}{\sigma}$

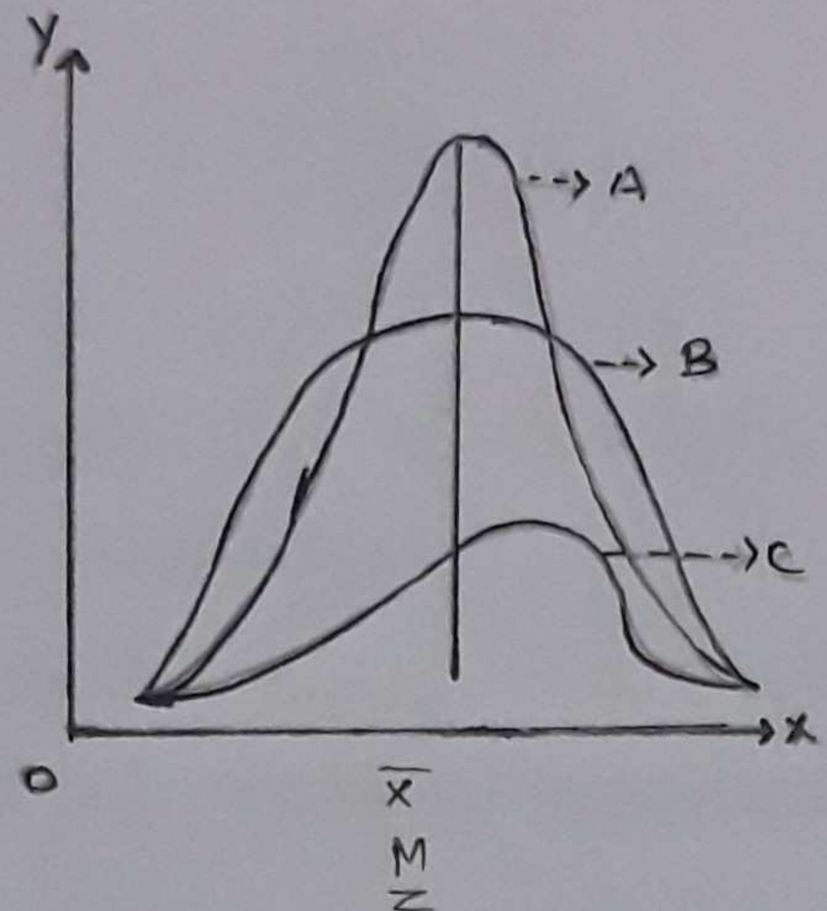
$$Sk_p = \frac{3(Mean - Median)}{\sigma}$$

2. BOWLEY $SK = \frac{Q_3 + Q_1 - 2Median}{Q_3 - Q_1}$

KURTOSIS

DEFINITION:

The degree of kurtosis of a distribution is measured relative to the peakedness or a normal - SIMPSON & KALFA



A - PEAKED (LEPTO)

B - NORMAL (MESO)

C - FLAT TOPPED (PLATY)

MEASURES OF KURTOSIS:

It is based on the fourth moments about the mean of distribution.

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

If $\beta_2 = 3$, the distribution is normal

$\beta_2 > 3$, the curve is lepto

$\beta_2 < 3$, the curve is platy

MOMENTS

DEFINITION:

Moment is a familiar term for the measure of a force with reference to its tendency to produce rotation. The strength of this tendency depends. obviously, upon the amount of the force and the distance from the origin of the point at which the force is exerted. - FEDERIC MILLS.

UNIT - IV

CORRELATION

DEFINITION:

Correlation is a analysis of co-variation between two or more variable.

Correlation Analysis attempts to determine the degree of relation ship between variable. — YA KUN CHOU

TYPES OF CORRELATION

There are 3 important types of correlation.

(i) Positive and Negative
(ii) Simple, Partial & Multiple
(iii) Linear & Non-linear

POSITIVE & NEGATIVE:

This type of correlation depends upon the direction of change of variables.

If two variable tend to move together in same direction, there is a increase in those two variables or decrease in those variables is called positive correlation.

If two variable tend to move opposite direction together, there is a increase on one variable and decrease in another and vice versa is called Negative.

SIMPLE, PARTIAL, MUILTIPLE:

This type depends on no. of variables.

If only two variables, it is called simple.

Eg: Demand & price.

In Multiple, we study more two variables simultaneously.

Eg: Price, Demand and commodity.

If we study two variables excluding the effect of other variables is called partial.

eg: The study of Price and demand eliminating the supply.

LINEAR OR NON-LINEAR

If the ratio of change between two variables is uniforms then it will be linear. If we plot the graph, it will be the straight line.

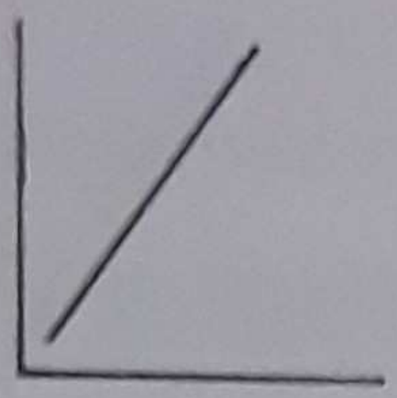
In NON-linear, we get a curve.

METHODS OF MEASURING CORRELATION:

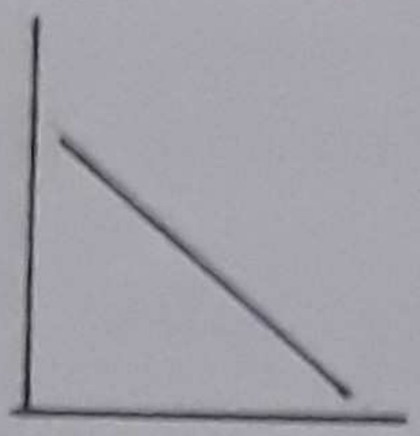
* scattered diagram
* Karl pearson co-efficient of correlation
* spearman's correlation
* RANK
* co-efficient of concurrent Deviation

SCATTERED DIAGRAM:

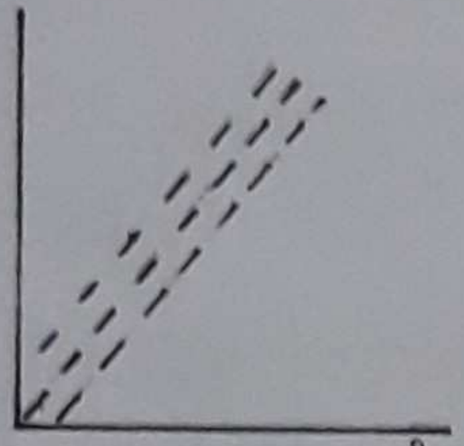
This is the simplest method of finding out whether there is any relationship present between 2 variables by plotting the values on a chart, known as scatter diagram. In this method, the given data are plotted on a graphed paper in the form of dots. This will show the types of correlation.



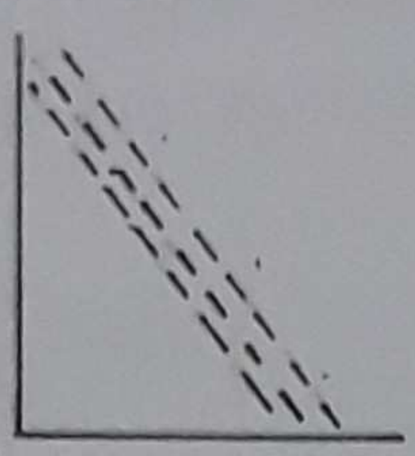
Perfect Positive correlation
(ie, $r = +1$ diagram)



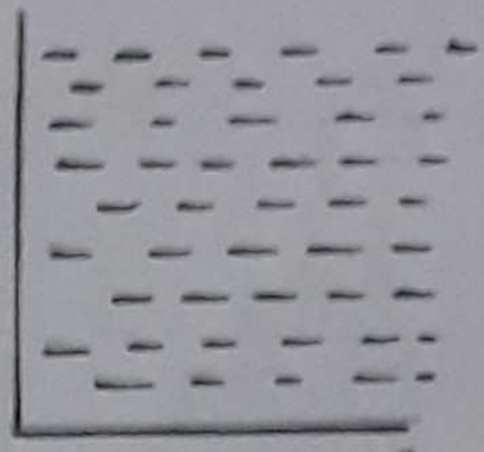
Perfect Negative correlation
(ie, $r = -1$ diagram)



High degrees of +ve correlation diagram



High degrees of -ve correlation



No correlation

KARL PEARSON'S CO-EFFICIENT OF CORRELATION:

This method is the most widely used method in Practice and is known as pearsonian co-efficient of correlation It is denoted by the symbol 'r', the formula,

(1) $r = \dfrac{\text{Covariance of } xy}{\sigma x \times \sigma y}$ (2) $r = \dfrac{\Sigma xy}{N\sigma x \times \sigma y}$

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

σx = standard deviation of series x.

σy = standard deviation of series y.

The value of co-efficient will always lie between -1 and $+1$.

$r = +1 \rightarrow$ Perfect +ve correlation

$r = -1 \rightarrow$ Perfect $-$ve correlation

$r = 0 \rightarrow$ No correlation.

RANK CORRELATION CO-EFFICIENT:

This method is based on rank. This measure is useful in dealing with Qualitative characteristic.

RANK ARE NOT GIVE:

When no rank is given but actual data are given then we must assing ranks. We can give ranks by the highest value has rank 1 and next value has rank 2 for both variable seperately.

$$\text{The correlation co-efficient} = 1 - \frac{6\Sigma D^2}{n^3 - n}$$

TIED RANKS:

When two or more items have equal values it is difficult to give ranks to the in that case, the items are given the average of the ranks. they could received if they are not tied.

REGRESSION

DEFINITION:

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

REGRESSION EQUATION:

There are two regression equation namely,

(i) R.E. of X on Y

(ii) R.E. of Y on X

There are two methods of finding regression co-efficient

DEVIATION TAKEN FROM ASSUMED MEAN:

When the actual is not a whole number then deviation are taken from the assumed mean and correlation is found out by using,

$$r = \frac{n\Sigma dx\,dy - (\Sigma dx)(\Sigma dy)}{\sqrt{n\Sigma d^2 - (\Sigma dx)^2}\ \sqrt{n\Sigma d^2 - (\Sigma dy)^2}}$$

where,

$dx = (x - A)$; $dy = (Y - B)$; A, B are assumed means of X & Y.

SPEARSMAN RANK CORRELATION:

- Take the deviation of x series and find Σdx
- Similarly, take the deviation of Y series and find Σdy
- Square dx and dy and get total Σdx^2 and Σdy^2

respectively.

- Multiply dx and dy and get $\Sigma dx.dy$.
- Then the formula,

$$= \frac{\Sigma dx\,dy - \frac{\Sigma dx\,\Sigma dy}{N}}{\sqrt{\left[\Sigma dx^2 - \frac{(\Sigma dx)^2}{N}\right]}\sqrt{\left[\Sigma dy^2 - \frac{(\Sigma dy)^2}{N}\right]}}$$

MATHEMATICAL PROPERTIES:

1) The co-efficient lies between -1 & $+1$.

$$-1 \leq r \leq +1$$

Moreover, this provides check on our calculations.

2) co-efficient is independant of change of origin and scale.

In change in scale, all the variables are multiplied or divided by same constant.

namely deviations from original mean of x and Y and deviations from assumed mean:

DEVIATION FROM ORIGINAL MEAN:

Regression of x on Y

$$(x - \bar{x}) = b_{xy} (Y - \bar{Y})$$

where $b_{xy} = r \frac{\sigma x}{\sigma y} = \frac{\Sigma xy}{\Sigma y^2}$

DEVIATION FROM ASSUMED MEAN:

If the actual mean is in fraction, this method can be used.

$$X - \bar{Y} = r \frac{\sigma x}{\sigma y} (Y - \bar{Y})$$

We can find out the value of $r \frac{\sigma x}{\sigma y}$ (by appling formula)

or, $b_{xy} = \frac{n \Sigma dx\, dy - (\Sigma dx)(\Sigma dy)}{n \Sigma dy^2 - \Sigma dy^2}$

MATHEMATICAL PROPERTIES:

(i) The Geometric means between two regression co-efficient is the co-efficient of correlation.

$$r = \pm \sqrt{b_{xy} \times b_{yx}}$$

(ii) Arithmetic mean of b_{xy} and b_{yx} is equal to or greater than r.

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

(iii) Regression co-efficients are independent of change of origin but not to side.

(iv) If one of the regression co-efficient is greater than unity the other one will less than unity because correlation co-efficient lie between -1 & +1.

(v) In regression co-efficient if any of the three values are given the fourth one can be found out.

DIFFERENCE BETWEEN CORRELATION AND REGIRESSION

CORRELATION	REGIRESSION
• Correlation is the relationship between two or more variable, which vary in sympathy with the other in the same (or) the opposite direction.	• Regression means going back and it is a mathematical measure showing the average relationship between two variables.
• Both the variables x and y are random variables.	• Here x is a random variable and y is a fixed variable sometimes both variables may be random variables.
• It finds out the degree of relationship between two variables are not their cause and effect of the variables.	• It indicates the cause and effect relationship the variables and establishes a function relationship.
• Their may be non-sense correlation b/n two variables.	• There is no such non-sense regression.

• It is used for Testing and verifying the relationships between two variables and series limited information.	• Besides, Verifications it is used for the production of one value in relationship to the other value.
• The co-efficient of correlation is a relative measure. The range of relationship lies b/n the variable.	• The regression co-efficient is an absolute figure. If we their value of the independent variable. we can find the value of the dependent variable.
• It is not veryful for further mathematical treatment.	• It is widely used for further mathematical treatment
• If the co-efficient of correlation is positive, then their to variables are positively correlated and viceversa.	• Regression co-efficient explicits that decrease in one variable is associated with the increase in the other variables.
• It has limited application because, it is confined only to linear relation ship b/n the variables.	• It has wider application as it studies linear and non-linear relationship b/n the variables.
• It is immediately whether x depends upon y (or) y depends upon x.	• There is a functional relation ship b/n the two variables. so, that we may identify b/n the independent and dependent variables.

UNIT- V

ASSOCIATION OF ATTRIBUTE

NOTATIONS AND DEFINITIONS:

(i) POSITIVE AND NEGATIVE CLASSES

The attributes may be positive or negative. If the attribute is present if termed as positive class.

They are denoted by A, B, C . . .

If the attribute is not present, it is termed as negative class.

They are denoted by greek letters α, β, γ. . .

(ii) ULTIMATE CLASS FREQUENCIES:

The number of values which posses a particular attribute is called the class frequencies.

Those classes which specify the attributes of the highest order are known as the ultimate class and their frequencies are known as the ultimate class frequencies.

In the study of Attribute (A) and (B) are the ultimate class frequency (AB) (αB) (Aβ) ($\alpha\beta$) formula 2^n.

In the study formula, the number of ultimate class frequency = 2^n.

(iii) ORDER OF CLASSES:

The Order of a class depends upon the no. of

attributes under study. A class having one attribute is known as the class of 1st order ; A class having two attributes is known as the second order class. N denotes Zero order class.

FIRST ORDER - A , B, α , β

SECOND ORDER - AB, αB , $\alpha\beta$, Aβ

N - ZERO ORDER

CONTINGENCY TABLE (OR) NINE SQUARE TABLE :

The class frequency in a study of two attributes A and B are represented by following table.

	A	α	
B	(AB)	(αB)	(B)
β	(Aβ)	($\alpha\beta$)	(β)
	(A)	(α)	N

Since it has parts q^2 which represent the class frequency which also called nine square table.

RELATIONSHIP :

(i) (A) = (AB) + (Aβ)

(ii) (α) = (αB) + ($\alpha\beta$)

(iii) (B) = (AB) + (αB)

(iv) (β) = (Aβ) + ($\alpha\beta$)

(v) N = (A) + (α) (or) N = (B) + (β) (or) N = (AB) + (αB) + (Aβ) + ($\alpha\beta$)

PROBLEM

Calculate the missing value and fillup the contingency table:

$(AB) = 35$, $(A) = 55$, $N = 100$ & $(B) = 65$

	A	α	
B	35	30	65
β	20	15	35
	55	45	100

$(\alpha) = (N - A) = 100 - 55 = 45$

$(\beta) = (N - B) = 100 - 65 = 35$

$(A\beta) = A - (AB) = 55 - 35 = 20$

$(\alpha B) = (B) - (AB) = 65 - 35 = 30$

$(\alpha\beta) = (\alpha) - (\alpha B) = 45 - 30 = 15$

TYPES OF ASSOCIATION:

There are 3 types,

i) Positive Association

(ii) Negative Association

(iii) Independent Association

POSITIVE ASSOCIATION

when two attributes are present (or) absent together in the data, and actual frequency is more than

the expected frequency is called positive association.

$$\text{ie, } (AB) > \frac{(A) \times (B)}{N}$$

(actual) (expected)

NEGATIVE ASSOCIATION

when the existance of one attribute causes absents of another attribute and actual frequency is less than the expected frequency is called Negative association.

$$\text{ie, } (AB) < \frac{(A) \times B)}{N}$$

(actual) (expected)

INDEPENDENT ASSOCIATIUN

when their exist no association two attributes (or) when they have no tendency to be present together (or) the Presence of one attribute are set to be independent actual frequency is equal to the expected frequency.

$$\text{ie, } AB = \frac{(A) \times (B)}{N}$$

(actual) (expected)

METHODS OF DETERMINING ASSOCIATION:

YULE'S CO-EFFICIENT OF ASSOCIATION:

The above mentioned methods will give us as a

rough idea about their associate but the association cannot be found out. Proff. yule, has suggested a formula to measure the association (Q) false between ±1.

If Q=0, no association; If Q=+1, their is Perfect positive association; If Q=-1, their is perfect negative association.

PROBLEM:

Find association between literacy and unemployment from the following frequencies.

TOTAL ADULTS = 10,000 (N)

LITERATE = 1290 (A)

UNEMPLOYED = 1390 (B)

LITERATE UNEMPLOYED = 820 (AB)

soln:

	A	α	
B	820		1390
β			
	1290		10,000

$\alpha = N - A = 10,000 - 1290 = 8710$

$\beta = N - B = 10,000 - 1390 = 8610$

$(A\beta) = A - AB = 1290 - 820 = 470$

$(\alpha B) = B - AB = 1390 - 820 = 570$

$(\alpha\beta) = \alpha - \alpha B = 8710 - 570 = 8140$

Since it is a positive, and consistent.

$$Q = \frac{(820)(8140) - (470)(570)}{(820)(8140) + (470)(570)}$$

$$Q = \frac{6,406,900}{69,42,700} = 0.9228.$$

There is a positive association b/n the attributes of A and B.

	A	α	
B	820	570	1390
β	470	8140	8610
	1290	8710	10,000