Correlation

Prepared By: Dr. N.V. Ravi,
Sr. Executive Officer, BOS,
ICAI.

Quantitative Aptitude & Business Statistics

Correlation

- Correlation is the relationship that exists between two or more variables.
- If two variables are related to each other in such a way that change increases a corresponding change in other, then variables are said to be correlated.

Examples

- **Relationship between the heights and weights.**
- **Relationship between the quantum of rainfall and the yield of wheat.**
- **Relationship between the Price and demand of commodity.**
- **Relationship between the dose of insulin and blood sugar.**

Uses of Correlation

- **Economic theory and business studies relationship between variables like price and quantity demand.**
- **Correlation analysis helps in deriving precisely the degree and the direction of such relationships.**

- **The effect of correlation is to reduce the range of uncertainty of our prediction .**
- **The prediction based on correlation analysis will more reliable and near to reality.**

Positive correlation

- If both the variables are vary in the same direction ,correlation is said to be positive .
- If one variable increases ,the other also increases or ,if one variable decreases ,the other also decreases ,then the two variables are said to be positive.

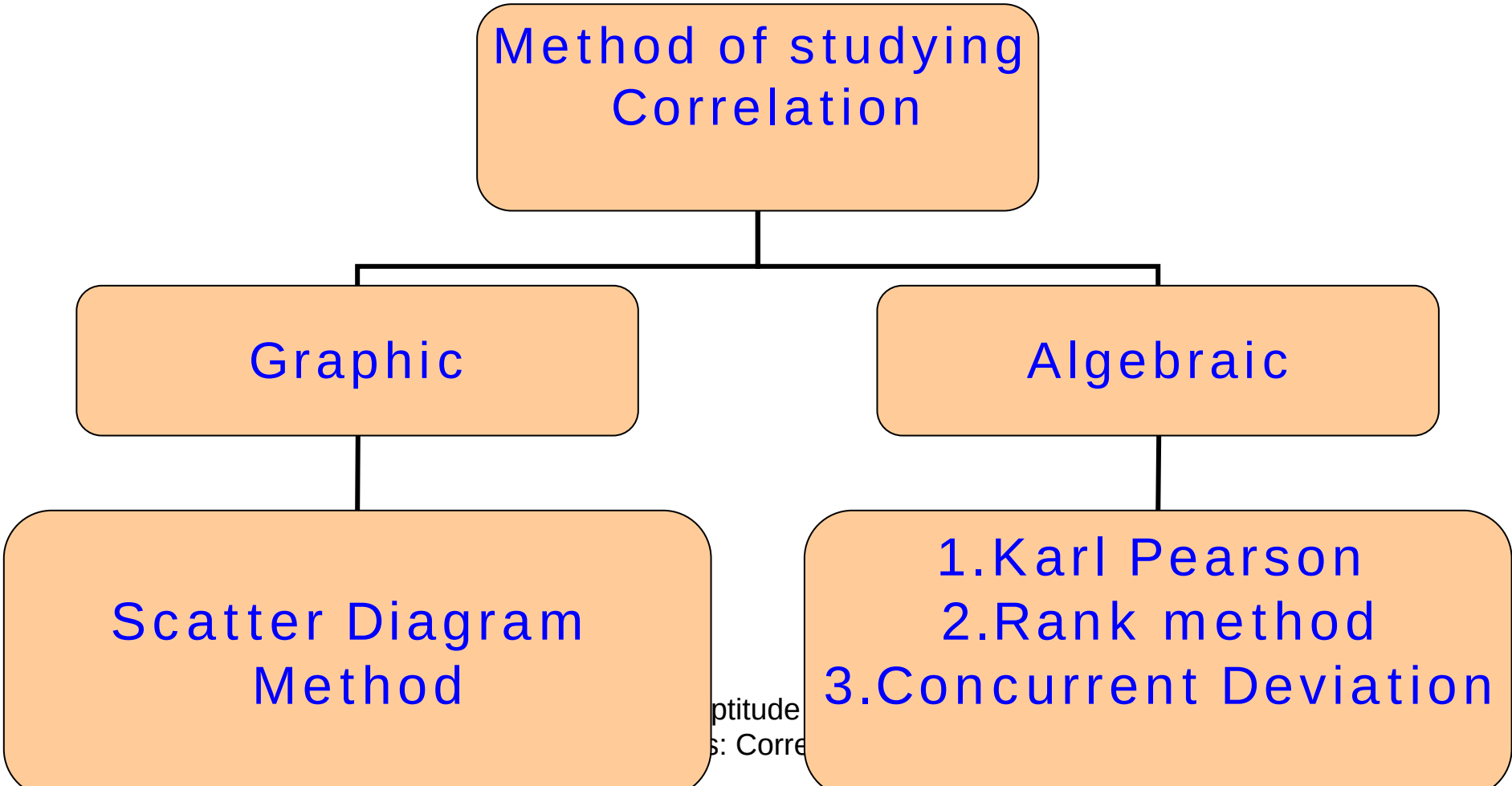
Negative correlation

- If both the variables are vary in the opposite direction ,correlation is said to be Negative.
- If one variable increases ,the other decrease or ,if one variable decreases ,the other also increases ,then the two variables are said to be Negative .

Types of Correlation

- **Simple correlation**
- **Multiple correlation**
- **Partial Multiple correlation**

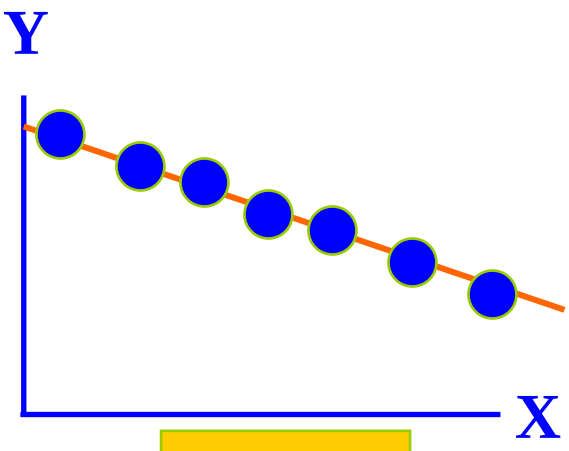
Methods of studying correlation



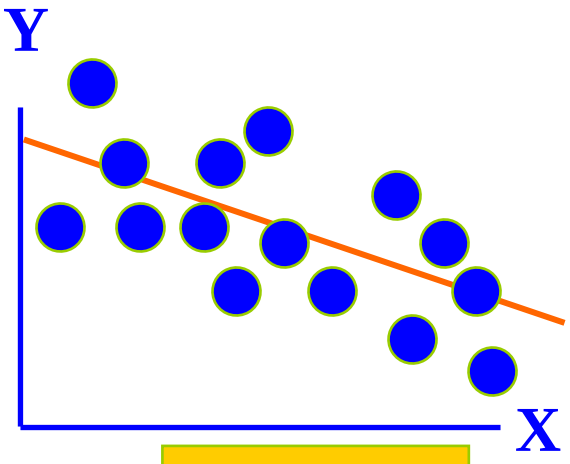
Scatter Diagram Method

- Scatter diagrams are used to demonstrate correlation between two quantitative variables.

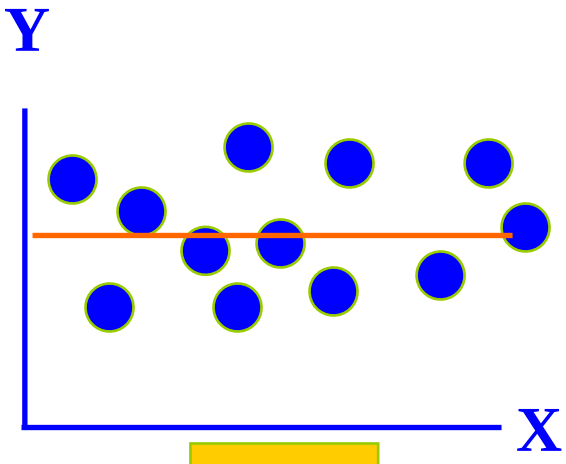
Scatter Plots of Data with Various Correlation Coefficients



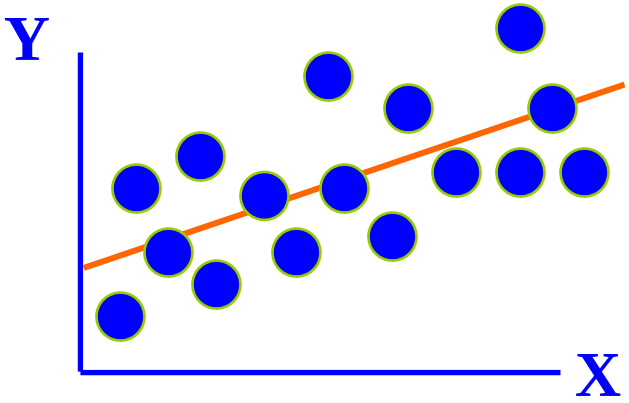
r = -1



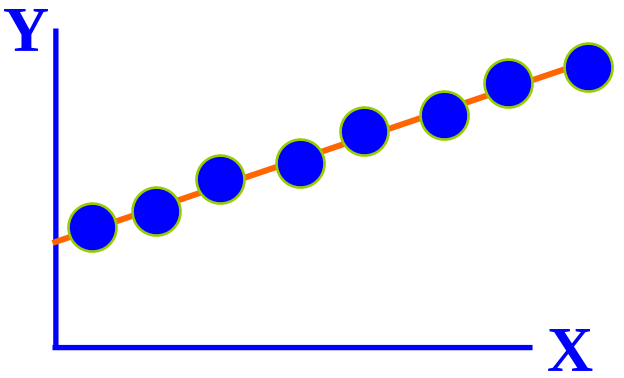
r = -Ve



***r* = 0**



***r* = +Ve**



***r* = 1**

Features of Correlation Coefficient

- Ranges between –1 and 1
- The closer to –1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker any positive linear relationship

The value of r lies between -1 and +1

- **If r=0 There exists no relationship between the variables**
- **If +0.75 ≤r ≤ +1 There exists high positive relationship between the variables .**
- **If -0.75 ≥ r ≥ -1 There exists high negative relationship between the variables**

- **If +0.5 ≤r ≤ 0.75 There exists Moderate positive relationship between the variables .**
- **If -0.50 ≥ r >-0.75 There exists moderate negative relationship between the variables.**
- **If r > -0.50 There exists low negative relationship between the variables**
- **If r <0.5 There exists low positive relationship between the variables .**

Covariance

- Definition : Given a n pairs of observations $(X_1,Y_1),(X_2,Y_2)$.,,,,,, (X_n,Y_n) relating to two variables X and Y ,the Covariance of X and Y is usually represented by Cov(X,Y)

$$Cov(X,Y) = \frac{\sum \left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}{N}$$

$$= \frac{\sum xy}{N}$$

Properties of Co-Variance

- Independent of Choice of origin
- not Independent of Choice of Scale.
- Co-variance lies between negative infinity to positive infinity.
- In other words co-variance may be positive or negative or Zero.

From the following Data Calculate Co-Variance

X	1	2	3	4	5
Y	10	20	30	50	40

Calculation of Covariance

X	$X-\bar{X}=x$	Y	$Y-\bar{Y}=y$	x.y
1	-2	10	-20	40
2	-1	20	-10	10
3	0	30	0	0
4	1	50	20	20
5	2	40	10	20
=15	=0	=150	=0	=90

- N= number of pairs =5

$$\overline{X} = \frac{\sum X}{N} = \frac{15}{5} = 3$$

$$\overline{Y} = \frac{\sum Y}{N} = \frac{150}{5} = 30$$

$$Cov(X,Y) = \frac{\sum \left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}{N}$$

$$= \frac{\sum xy}{N} = \frac{90}{5} = 18$$

Karl Pearson's Correlation

- The most widely used mathematical method for measuring the intensity or the magnitude of linear relationship between two variables was suggested by Karl Pearson's

Coefficient of Correlation

- Measures the strength of the linear relationship between two quantitative variables

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}$$

Properties of KralPear son's Coefficient of Correlation

- **Independent of choice of origin**
- **Independent of Choice Scale**
- **Independent of units of Measurement**

Assumptions of Karl Pearson's Coefficient of Correlation

- **Linear relationship between variables.**
- **Cause and effect relationship.**
- **Normality.**

- The correlation coefficient lies between -1 and +1
- The coefficient of correlation is the geometric mean of two regression coefficients.

Merits of Karl Pear son's Coefficient of Correlation

- **Coefficient of Correlation gives direction as well as degree of relationship between variables**
- **Coefficient of Correlation along with other information helps in estimating the value of the dependent variable from the known value of independent variable.**

Limitations of KralPear son's Coefficient of Correlation

- **Assumptions of Linear Relationship**
- **Time consuming**
- **Affected by extreme values**
- **Requires careful Interpretation**

From the following Data Calculate Coefficient of correlation

X	1	2	3	4	5
Y	10	20	30	50	40

X	$X-\bar{X}=x$	x^2
1 2 3 4 5	-2 -1 0 1 2	4 1 0 1 4
=15	=0	=10

Y	$Y-\bar{Y}=y$	y^2	x.y
10	-20	400	40
20	-10	100	10
30	0	0	0
50	20	400	20
40	10	100	20
=150	=0	=1000	=90

- N= number of pairs =5

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}}$$

$$= \frac{90}{\sqrt{10000}} = \frac{90}{100} = +0.9$$

$$\overline{X} = \frac{\sum X}{N} = \frac{15}{5} = 3$$

$$\overline{Y} = \frac{\sum Y}{N} = \frac{150}{5} = 30$$

- r=0.9 there exists high degree of positive correlation

Correlation for Bivariate analysis

$$r = \frac{\sum f d_x . d_y - \frac{\left(\sum f . d_x\right)\left(\sum f . d_y\right)}{N}}{\sqrt{\sum f . d_x^2 - \frac{\left(\sum f . dx\right)^2}{N}} \sqrt{\sum f . d_y^2 - \frac{\left(\sum f . dx\right)^2}{N}}}$$

Standard error

- **Standard error of co efficient of correlation is used foe ascertaining the probable error of coefficient of correlation**
- **Where r=Coefficient of correlation**
- **N= No. of Pairs of observations**

$$SE = \frac{1 - r^2}{\sqrt{N}}$$

Probable Error

- The Probable error of coefficient of correlation is an amount which if added to and subtracted from value of r gives the upper and lower limits with in which coefficients of correlation in the population can be expected to lie. It is 0.6745 times of standard error.

$$\Pr obable Error = 0.6745.\frac{1-r^2}{\sqrt{N}}$$

Uses of Probable Error

- PE is used to for determining reliability of the value of r in so far as it depends on the condition of random sampling.

Case	Interpretation
1.If |r |< 6 PE	The value of r is not at all significant. There is no evidence of correlation.
2. 1.If |r | >6 PE	The value of r is significant. There is evidence of correlation

Example

- **If r=-0.8 and N=36 ,Calculate a) Standard Error ,b) Probable Error and C) Limits of Population correlation .Also State whether r is significant**
- **Solution**
- **A)**

$$SE = \frac{1-r^2}{\sqrt{N}} = \frac{1-(-0.8)^2}{\sqrt{36}}$$

$$= \frac{1-0.64}{6} = \frac{0.36}{6} = 0.06$$

- b) Probable Error=0.6745.SE=0.6745*0.06=0.04
- c) Limits of Population Correlation
- =r± PE (r)= -0.8±0.04
- =-0.84 to -0.76
- d) Ratio of r to PE of r =
- |r |/PE(r)=0.8/0.04=20times
- Since the value of r is more than 6 times the Probable error ,the value of r is significant .Hence the existence of correlation

Coefficient of determination

- The coefficient of determination is defined as the ratio of the explained variance to the total variance
- Calculation: The coefficient determination is calculated by squaring the coefficient of correlation

Example

- If r=0.8 ,what is the proportion of variation in the dependent variable which is explained the independent variable?
- Solution :
- If r=0.8 ,r^2=0.64,
- It means 64% variation in the dependent variable explained by independent variable.

Coefficient of non-determination

- The coefficient of non determination is defined as the ratio of the unexplained variance to the total variance
- Calculation: The coefficient non determination is calculated by subtracting the Coefficient of determination from one.

Example

- If r=0.8 ,what is the proportion of variation in the dependent variable which is not explained the independent variable?
- Solution; Coefficient of determination $=r^2=0.64$
- Coefficient of non-determination
- $=1-r^2=0.36$,It means 36% variation in the dependent variable not explained by independent variable.

Spearman's Rank Correlation

Spearman's Rank Correlation uses ranks than actual observations and make no assumptions about the population from which actual observations are drawn.

$$r = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)}$$

Spearman's Rank Correlation for repeated ranks

- Where m=the no of times ranks are repeated
- n=No of observations
- r= Correlation Coefficient

$$r = 1 - \frac{6\left[\sum D^2 + \frac{m^3 - m}{12} + ...\right]}{n\left(n^2 - 1\right)}$$

Calculation of Rank Correlation

- Two judges in a beauty contest ranked the entries as follows

X	1	2	3	4	5
Y	5	4	3	2	1

X	Y	$d=r_1-r_2$	d^2
1	5	-4	16
2	4	-2	4
3	3	0	0
4	2	2	4
5	1	4	16
n=5			$\sum d^2$ =40

$$r = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)}$$

$$= 1 - \frac{6 \times 40}{5\left(5^2 - 1\right)}$$

$$= -1$$

Features of Spearman's Rank Correlation

- Spearman's Correlation coefficient is based on ranks rather than actual observations .
- Spearman's Correlation coefficient is distribution –free and non-parametric because no strict assumptions are made about the form of population from which sample observation are drawn.

Features of Spearman's Rank Correlation

- The sum of the differences of ranks between two variables shall be Zero
- It can be interpreted like Karl Pearson's Coefficient of Correlation.
- It lies between -1 and +1

Merits of Spearman's Rank Correlation

- Simple to understand and easy to apply
- Suitable for Qualitative Data
- Suitable for abnormal data.
- Only method for ranks
- Appliacble even for actual data.

Limitations of Spearman's Rank Correlation

- Unsuitable data
- Tedious calculations
- Approximation

When is used Spearman's Rank Correlation method

- The distribution is not normal
- The behavior of distribution is not known
- only qualitative data are given

Meaning of Concurrent Deviation Method

- **Concurrent Deviation Method is based on the direction of change in the two paired variables .The coefficient of Concurrent Deviation between two series of direction of change is called coefficient of Concurrent Deviation .**

- r_c=Coefficient of Concurrent deviation
- C= no of positive signs after multiplying the change direction of change of X-series and Y-Series
- n=no. of pairs of observations computed

$$r_c = \pm\sqrt{\pm\frac{2c-n}{n}}$$

Limitations of Concurrent Deviation Method

- This method does not differentiate between small and big changes .
- Approximation

Merits of Concurrent Deviation

- **Simple to understand and easy to calculate.**
- **Suitable for large N**

Calculation of coefficient of concurrent deviation

X	59	69	39	49	29
Y	79	69	59	49	39

X	Direction of Change of X (D_x)	Y	Direction of Change of X (D_y)	$D_x * D_y$
59		79		
69	+	69	-	-
39	-	59	-	+
49	+	49	-	-
29	-	39	-	+
	n=4			C=2

$$r_c = \pm\sqrt{\pm\frac{2c - n}{n}}$$

$$= 0$$

- 1___ is a relative measure of association between two or more variables

  (a)coefficient of correlation

  (b)coefficient of regression

  (c) both

  (d) none of these

- 1___ is a relative measure of association between two or more variables

  (a)coefficient of correlation

  (b)coefficient of regression

  (c) both

  (d) none of these

- **2.The correlation coefficient lies between**

  (a) –1 and +1

  (b) 0 and +1

  (c) –1 and 0

  (d) none of these

- **2.The correlation coefficient lies between**

  (a) –1 and +1

  (b) 0 and +1

  (c) –1 and 0

  (d) none of these

- **3. r is independent of __**

(a) choice of origin and not of choice of scale

(b) choice of scale and not of choice of origin

(c) both choice of origin and choice of scale

(d) none of these

- **3. r is independent of __**

(a) choice of origin and not of choice of scale

(b) choice of scale and not of choice of origin

(c) both choice of origin and choice of scale

(d) none of these

- **4.Probable error is ___**

(a) 0.6475 standard error

(b) 0.6745 standard error

(c) 0.6457 standard error

(d) 0.6547 standard error

- **4.Probable error is ___**

(a) 0.6475 standard error

(b) 0.6745 standard error

(c) 0.6457 standard error

(d) 0.6547 standard error

- **5.The product moment correlation coefficient is obtained by the formula**

(a) r = $\dfrac{\sum XY}{N\ \sigma X\ \ \sigma Y}$

(b) r = $\dfrac{\sum xy}{N\sigma x\,\sigma y}$

(c) r = $\dfrac{\sum xy}{N\ \sqrt{\sigma\ x\ \ \sigma\ y}}$

(d) r = $\dfrac{\sum xy}{\sqrt{N\ \sigma\ x\ \sigma\ y}}$

- **5.The product moment correlation coefficient is obtained by the formula**

(a) $r = \dfrac{\sum XY}{N \sigma X \sigma Y}$

(b) $r = \dfrac{\sum xy}{N \sigma x \sigma y}$

(c) $r = \dfrac{\sum xy}{N \sqrt{\sigma x \sigma y}}$

(d) $r = \dfrac{\sum xy}{\sqrt{N \sigma x \sigma y}}$

- **6. Correlation between Temperature and Sale of Woolen Garments.**
- **A) Positive**
- **B) 0**
- **C) negative**
- **D) none of these**

- **6. Correlation between Temperature and Sale of Woolen Garments.**
- **A) Positive**
- **B) 0**
- **C) negative**
- **D) none of these**

- **7.Covarince can vary from**
- **A)-1 to +1**
- **B)- infinity to + infinity**
- **C)-1 to 0**
- **D) 0 to +1**

- **7.Covarince can vary from**
- **A)-1 to +1**
- **B)- infinity to + infinity**
- **C)-1 to 0**
- **D) 0 to +1**

- **8.Karl Pearson’ s coefficient is defined from**
- **A) Ungrouped data**
- **B) grouped data**
- **C) Both**
- **D) none**

- **8.Karl Pearson’ s coefficient is defined from**
- **A) Ungrouped data**
- **B) grouped data**
- **C) Both**
- **D) none**

- 9. The coefficient of non determination is 0.36 ,the value of r will be
- A)0.64
- B)0.60
- C)0.80
- D)0.08

- 9. The coefficient of non determination is 0.36 ,the value of r will be
- A)0.64
- B)0.60
- C)0.80
- D)0.08

- **10.What is Spurious correlation**
- **A) It is bad relation between variables**
- **B) It is low correlation between variables**
- **C) It is the correlation between two variables having no causal relation**
- **D) It is a negative correlation**

- **10.What is Spurious correlation**
- **A) It is bad relation between variables**
- **B) It is low correlation between variables**
- **C) It is the correlation between two variables having no causal relation**
- **D) It is a negative correlation**

- **11.Rank coefficient correlation was developed by**
- **A) Karl Pearson**
- **B) R.A.Fisher**
- **C) Spearman**
- **D) Bowley**

- **11.Rank coefficient correlation was developed by**
- **A) Karl Pearson**
- **B) R.A.Fisher**
- **C) Spearman**
- **D) Bowley**

- **12. If r=0.9 probable error = 0.032 ,**
- **Value of N will be**
- **A)14**
- **B)15**
- **C)16**
- **D)17**

- **12. If r=0.9 probable error = 0.032 ,**
- **Value of N will be**
- **A)14**
- **B)15**
- **C)16**
- **D)17**

- **13.If the value of $r^2$for a particular situation is 0.49.what is the coefficient of correlation**
- **A)0.49**
- **B)0.7**
- **C)0.07**
- **D) cannot be determined**

- **13.If the value of r^2 for a particular situation is 0.49.what is the coefficient of correlation**
- **A)0.49**
- **B)0.7**
- **C)0.07**
- **D) cannot be determined**

- **14.What is the Quickest method to find correlation between variables .**
- **A) Scatter method**
- **B) Method of Concurrent Deviation**
- **C) Method of Rank correlation**
- **D) Method of Product moment correlation**

- **14.What is the Quickest method to find correlation between variables .**
- **A) Scatter method**
- **B) Method of Concurrent Deviation**
- **C) Method of Rank correlation**
- **D) Method of Product moment correlation**

- **15 If r=0.6 ,then the coefficient of non determination is**
- **A)0.4**
- **B)-0.6**
- **C)0.36**
- **D)0.64**

- **15 If r=0.6 ,then the coefficient of non determination is**
- **A)0.4**
- **B)-0.6**
- **C)0.36**
- **D)0.64**

- 17. If the relationship between two variables x and y is given by $2x + 3y + 4 = 0$, then the value of the correlation coefficient between x and y is
- A) 0
- B) 1
- C) **–1**
- D) Negative

- **17. If the relationship between two variables x and y is given by 2x + 3y + 4 = 0, then the value of the correlation coefficient between x and y is**
- **A) 0**
- **B) 1**
- **C) –1**
- **D) Negative**

- **18 When r = 0 then cov(x,y) is equal to**
- **A) + 1**
- **B) – 1**
- **C) 0**
- **D) None of these.**

- **18 When r = 0 then cov(x,y) is equal to**
- **A) + 1**
- **B) – 1**
- **C) 0**
- **D) None of these.**

- **19. For finding the degree of agreement about beauty between two Judges in a Beauty Contest, we use_____ .**
- **A) Scatter diagram**
- **B) Coefficient of rank correlation**
- **C) Coefficient of correlation**
- **D) Coefficient of concurrent deviation**

- **19. For finding the degree of agreement about beauty between two Judges in a Beauty Contest, we use_____ .**
- **A) Scatter diagram**
- **B) Coefficient of rank correlation**
- **C) Coefficient of correlation**
- **D) Coefficient of concurrent deviation**

- **20. Coefficient of determination is defined as**
- **A) r^3**
- **B) $1–r^2$**
- **C) $1+r^2$**
- **D) r^2**

- 20. Coefficient of determination is defined as
- A) r^3
- B) $1–r^2$
- C) $1+r^2$
- D) r^2

THE END

Correlation