1. What is the definition of covariance? Create formula for it.

→ Covariance is the measure of the relationship between two random variables and to what extent, they change together.

or in other words we can say, it defines changes between two variables, such that change in one variable is equal to change in another variables. ~~This~~

There are two kinds of Covariance:-

i) Positive Covariance

ii) Negative Covariance.

Formula :-

For Population:-

$$Cov(x,y) = \sum_{i=1}^{N} \frac{(x-\bar{x})(y-\bar{y})}{N}$$

For Sample :-

$$Cov(x,y)$$

2. What makes Correlation better than Covariance?

⇒ Covariance and correlation are two terms which are exactly opposite to each other, they both are used in statistics and regression analysis, Covariance shows us how the two variance variable vary from each other whereas correlation shows us the relationship between the two variables and how are they related.

Correlation and covariance are very closely related to each other and yet they differ a lot covariance defines the type of interaction but correlation defines not the type only but also the strength of this relationship. Due to this reason correlation is often termed as the special case of covariance. However if one must choose between the two, most will choose correlation as it remain unaffected by the

06 SUNDAY

changes in dimensions, location and scale. Also, since it range between -1 to +1 it is useful to draw comparisons between variables across domains.

3. Explain the process as well as Pearson and spearman correlation.

⇒ Correlation is a degree to which two variables are linearly related. This is an important step in bi-variant data analysis. In the broadest sense correlation is actually any statistical relationship whether casual or not.

The correlation coefficient is a statistical measure of the strength of the relationship between the relative movement of two variable. The value range between -1.0 and 1.0 A correlation of -1.0 shows negative correlation, while a correlation of 1.0 shows a positive correlation. A

correlation of 0.0 shows no linear relationship between the movement of two variable.

There are two important Correlation Coefficient :-

1. Pearson Correlation Coefficient :-

The Pearson correlation coefficient is a measure of the strength of linear association between two variables and is denoted by r. Basically, a Pearson coefficient draw a line of best fit through the data of two variable, and the Pearson correlation coefficient, r indicates how far away all data points are to this line of best fit.

Formula :-

$$\rho(x,y) = \frac{Cov(x,y)}{\sigma_x * \sigma_y}$$

ii) Spearman's correlation coefficient:

It measures the strength and direction of the association between two ranked variables. But before we talk about the Spearman's correlation coefficient. it is a statistical measure of the strength of a monotonic relationship between paired data. In a sample it is denoted by r_s and in by design constraints:

$$-1 \leq r_s \leq 1$$

Formula:

$$r_s = \frac{Cov(R_x, R_y)}{\sigma R_x * \sigma R_y}$$

4. What are the advantage over Spearman correlation over pearson correlation?

→ It is not able to capture the non-linear properties, that's why we prefer of using Spearman Rank correlation formula which focuses on rank of variable.

→ Spearman's coefficient as a significant measure of the strength of the associations between two variables. much better than pearson Correlation.

→ Pearson correlation evaluates the ~~monotonic~~ linear relationship between two continuous variables.

whereas Spearman correlation evaluates the monotonic relationship. The Spearman correlation coefficient based on the ranked values for each variable rather than the raw data.

5. Describe the central Limit theorem.

→

It states that regardless of the shape of the population distribution. The distribution of Sample means will always be approximately normal.

→ If we take sample of any population distribution, the mean of those samples would be in normal distribution.

→ The distribution of Sample means will become more normal as its sample size increases.

→ Sample distribution mean will be approximatly normal if their Sample size is greather than 30.