Name: Chandrakant B Thakur

Statistics Assignment 4

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

covariance is a measure of the relationship between two random variables. The metric evaluates how much – to what extent – the variables change together. In other words, it is essentially a measure of the variance between two variables. However, the metric does not assess the dependency between variables.

- **Positive covariance**: Indicates that two variables tend to move in the same direction.
- **Negative covariance**: Reveals that two variables tend to move in inverse directions.

$$\text{Cov}(X, Y) = \frac{\sum(X_i - \bar{X})(Y_j - \bar{Y})}{n}$$

For a sample covariance, the formula is slightly adjusted:

$$\text{Cov}(X, Y) = \frac{\sum(X_i - \bar{X})(Y_j - \bar{Y})}{n - 1}$$

Where:

- $\mathbf{X_i}$ – the values of the X-variable
- $\mathbf{Y_j}$ – the values of the Y-variable
- $\bar{\mathbf{X}}$ – the mean (average) of the X-variable
- $\bar{\mathbf{Y}}$ – the mean (average) of the Y-variable
- **n** – the number of data points

2. What makes Correlations better than Covariance?

Name: Chandrakant B Thakur

Basis	Covariance	Correlation
Meaning	Covariance is an indicator of how two random variables are dependent on each other. A higher number denotes higher dependency.	Correlation indicates how strongly these two variables are related, provided other conditions are constant. The maximum value is +1, representing a perfect dependent relationship.
Relationship	We can deduct correlation from a covariance.	Correlation provides a measure of covariance on a standard scale. It is deduced by dividing the calculated covariance by standard deviation.
Values	The value of covariance lies in the range of -∞ and +∞.	Correlation is limited to values between the range -1 and +1.
Scalability	Covariance is affected.	Correlation is not affected by a change in scales or multiplication by a constant.
Units	Covariance has a definite unit as deduced by the multiplication of two numbers and their units.	Correlation is a unitless absolute number between -1 and +1, including decimal values.

Correlation and covariance are very closely related to each other, and yet they differ a lot. Covariance defines the type of interaction, but correlation represents the type and the strength of this relationship. Due to this reason, correlation is often termed the special case of covariance. However, if one must choose between the two, most analysts prefer correlation as it remains unaffected by the changes in dimensions, locations, and scale. Also, since it is limited to a range of -1 to +1, it is useful to draw comparisons between variables across domains. However, an important limitation is that these concepts measure only the linear relationship.

3. Explain the process as well as Pearson and Spearman Correlation.

Pearson Correlation Coefficient (PCC):

Pearson Correlation is the coefficient that measures the degree of relationship between two random variables. The coefficient value ranges between +1 to -1. Pearson correlation is the normalization of covariance by the standard deviation of each random variable.

$$PCC(X,Y) = \frac{COV(X,Y)}{SD_x * SD_y}$$

Name: Chandrakant B Thakur

Spearman Rank Correlation Coefficient (SRCC):

SRCC covers some of the limitations of PCC. It does not carry any assumptions about the distribution of the data. SRCC is a test that is used to measure the degree of association between two variables by assigning ranks to the value of each random variable and computing PCC out of it. Given two random variable X, Y. Compute rank of each random variable, such that the least value has rank 1. Then apply the Pearson correlation coefficient on Rank(X), Rank(Y) to compute SRCC.

$$SRCC(X, Y) = PCC(rank(X), rank(Y))$$

SRCC ranges between -1 to +1 and works well with monotonically increasing or decreasing functions.

4. What are the advantages of Spearman Correlation over Pearson Correlation?

The Spearman rank correlation coefficient is only to be used to describe the relationship between linear data. Also, it can be used for data at the ordinal level and it is easier to calculate by hand than the Pearson correlation coefficient.

The Spearman rank correlation coefficient is only to be used to describe the relationship between nonlinear data. Also, it can be used for data at the ordinal level and it is easier to calculate by hand than the Pearson correlation coefficient.

The Spearman rank correlation coefficient can be used to describe the relationship between linear or nonlinear data. Also, it can be used for data at the ordinal level and it is easier to calculate by hand than the Pearson correlation coefficient.

5. Describe the Central Limit Theorem.

The Central Limit Theorem (CLT) states that the distribution of a sample mean that approximates the normal distribution, as the sample size becomes larger, assuming that all the samples are similar, and no matter what the shape of the population distribution.

Name: Chandrakant B Thakur

Central Limit Theorem Formula

The central limit theorem is applicable for a sufficiently large sample size ($n \geq 30$). The formula for central limit theorem can be stated as follows:

$$\mu_{\overline{x}} = \mu$$

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

Where,

μ = Population mean

σ = Population standard deviation

μ_x = Sample mean

σ_x= Sample standard deviation

n = Sample size