**Statistics Assignment 4**

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

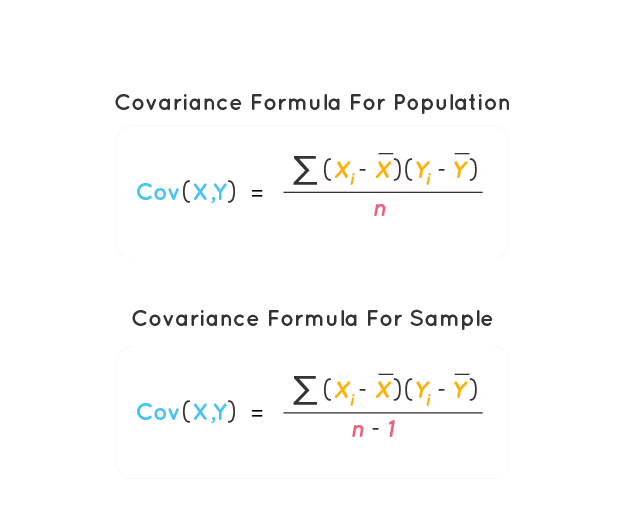
Answer:-

- The covariance formula is used to assess the relationship between two variables. It is essentially a measure of the variance between two variables. Covariance is measured in units and is calculated by multiplying the units of the two variables. The variance can be any positive or negative values. Following are the interpreted values:

- When two variables move in the same direction, it results in a positive covariance

Contrary to the above point is two variables in opposite directions, it results in a negative covariance.

- Covariance is a measure of the relationship between two random variables, in statistics. The covariance indicates the relation between the two variables and helps to know if the two variables vary together. In the covariance formula, the covariance between two random variables X and Y can be denoted as Cov(X, Y).



Where,

Xi is the values of the X-variable

Yi is the values of the Y-variable

--

X is the mean of the X-variable

--

Y is the mean of the Y-variable

n is the number of data points

---------------------------------------------------------------------------------------------------------------------------

1. What makes Correlations better than Covariance?

Answer:-

|  |  |
| --- | --- |
| Correlation | Covariance |
| 1. Correlation is an indicator of how strongly these 2 variables are related, provided other conditions are constant. The maximum value is +1, denoting a perfect dependent relationship. | 1. Covariance is an indicator of the extent to which 2 random variables are dependent on each other. A higher number denotes higher dependency. |
| 2. Correlation provides a measure of covariance on a standard scale. It is deduced by dividing the calculated covariance with standard deviation. | 2. Correlation can be deduced from a covariance. |
| 3. Correlation is limited to values between the range -1 and +1. | 3. The value of covariance lies in the range of and -∞ to +∞. |
| 4. Correlation is not affected by a change in scales or multiplication by a constant. | 4. Affects covariance. |
| 5. Correlation is a unit less absolute number between-1 and +1, including decimal values | 5. Covariance has a definite unit as it is deduced by the multiplication of two numbers and their units. |

--------------------------------------------------------------------------------------------------------------------------

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

Answer:-

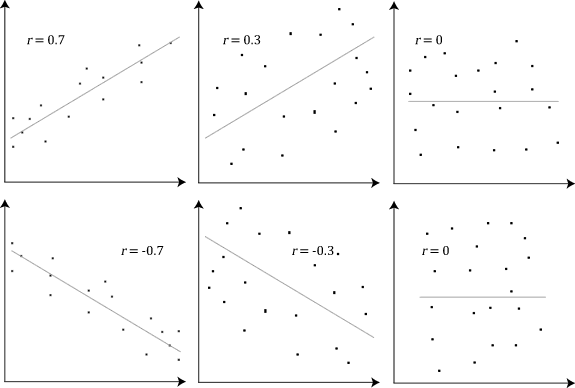
- Both Pearson and Spearman are used for measuring the correlation but the difference between them lies in the kind of analysis we want.

- Pearson correlation:

- Pearson correlation evaluates the linear relationship between two continuous variables.

- The Pearson correlation coefficient also referred to as Pearson’s *r*or the bivariate correlation is a statistic that measures the linear correlation between two variables *X* and *Y*. It has a value between +1 and −1. A value of +1 is a total positive linear correlation, 0 is no linear correlation, and −1 is a total negative linear correlation.

- Below is an example of how the Pearson correlation coefficient (r) varies with the strength and the direction of the relationship between the two variables. Note that when no linear relationship could be established (refer to graphs in the third column), the Pearson coefficient yields a value of zero.



- Spearman correlation:

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

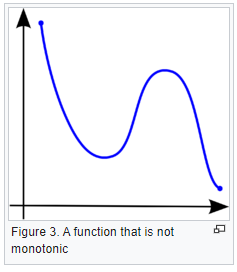
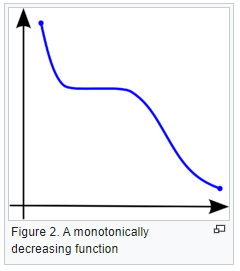
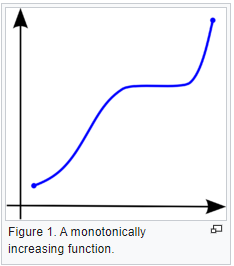
- Spearman’s rank correlation coefficient or Spearman’s ρ, named after Charles Spearman is a nonparametric measure of rank correlation (statistical dependence between the rankings of two variables). It assesses how well the relationship between two variables can be described using a monotonic function.

- A monotonic relationship is a relationship that does one of the following:

(1) As the value of one variable increases, so does the value of the other variable, OR,

(2) As the value of one variable increases, the other variable value decreases.

- But, not exactly at a constant rate whereas in a linear relationship the rate of increase/decrease is constant.



------------------------------------------------------------------------------------------------------------------------

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

Answer:-

- Pearson correlation coefficients measure the linear relationship between the variables while Spearman correlation coefficients measure only monotonic relationships, relationship in which the variables tend to move in the same/opposite direction but not necessarily at a constant rate whereas the rate is constant in a linear relationship.

-------------------------------------------------------------------------------------------------------------------------

1. Describe the Central Limit Theorem.

Answer:-

- The central limit theorem (CLT) states that the distribution of a sample variable approximates a normal distribution (i.e., a “bell curve”) as the sample size becomes larger, assuming that all samples are identical in size, and regardless of the population's actual distribution shape.

- CLT is a statistical premise that, given a sufficiently large sample size from a population with a finite level of variance, the mean of all sampled variables from the same population will be approximately equal to the mean of the whole population.

