Statistics Assignment 4

1. What is the definition of covariance? Create the formula for it.
Answer:
Covariance quantify the relationship between two different features (variables) of a dataset.
For Population: $Cov(X,Y) = (1/N) * \sum_{i=1}^{N}((Xi - \bar{X}) * (Yi - \bar{Y}))$
For Sample: $Cov(X,Y) = (1/n-1) * \sum_{i=1}^{n}((Xi - \bar{X}) * (Yi - \bar{Y}))$
where,

- Xi is the values of the X-variable
- Yi is the values of the Y-variable
- $\bar{X}$ is the mean of the X-variable
- $\bar{Y}$ is the mean of the Y-variable
- N is the total number of values in the population
- n is the total number of values in the sample

Higher the positive value of Cov(X,Y) means higher the directly proportional relationship between X and Y.
Higher the negative value of Cov(X,Y) means higher the inversely proportional relationship between X and Y.

2. What makes Correlations better than Covariance?
Answer:
In Covariance, there is no limit on its range. It means that covariance can give magnitude of any length i.e. no fixed value for magnitude. For example +1000, +10, -200, -2000 etc. Due to this we cannot determine how much a variable is positively or negatively correlated with other variable.
To overcome this problem, we use Correlation which restrict all the relationship values to a range of -1 to +1. According to this range [-1 to +1], a value close to +1 shows that the two variables are highly positively correlated while a value close to -1 shows that the two variables are highly negatively correlated and value = 0 means that both variables are not correlated at all.

3. Explain the process as well as Pearson and Spearman Correlation?
Answer:
The Pearson Correlation measures only the linear relationship between two continuous variables. A relationship is linear when a change in one variable is directly proportional to the change in the other variable. It ranges between [-1 to +1].

Pearson Correlation is denoted by ρ and calculated as shown below: (For Population)

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

The Spearman Correlation measures the monotonic relationship (whether liner or non-linear) between two continuous or ordinal variables. The Spearman correlation coefficient is based on the ranked values for each variable rather than the raw data.

Spearman Correlation can be calculated as shown below:

$$r_s = \rho_{\mathrm{R}(X),\mathrm{R}(Y)} = \frac{\operatorname{cov}(\mathrm{R}(X),\mathrm{R}(Y))}{\sigma_{\mathrm{R}(X)} \sigma_{\mathrm{R}(Y)}}$$

where, R(X) and R(Y) are rank of X and Y respectively.

4. What are the advantages of Spearman Correlation over Pearson Correlation?
Answer:
According to Pearson Correlation, there should be a linear relationship between two continuous variables(say X and Y). It means that with increase in value of variable X, the value of variable Y also increases at every step, then the Pearson's Correlation will be +1 with respect to X and Y.

It is been observed if the value of variable X is increases with increase in the value of variable Y but a very slower rate, then the Pearson Correlation does not give +1 correlation with respect to X and Y. It give value less than +1.
To overcome this problem, Spearman Correlation is used.

Spearman Correlation works on monotonic relationship between two continuous variables. In a monotonic relationship, the variables tend to change together, but not necessarily at a constant rate. And this correlation works on the ranked values for each variable rather than the raw data or rate of increase/decrease.

5. Describe the Central Limit Theorem.

Answer:

Central Limit Theorem (CLT) states that regarding of the shape of the population distribution, the distribution of sample means will approximately be in Normal Distribution.

2 Important points to consider in CLT:-

i. The distribution of sample means will become more and more normal as its sample size increases more and more.
ii. Good Thumb Rule: Sample distribution will approximately be Normal if Sample Size is greater than or equal to 30 i.e. ($n \geq 30$)

Basically it means that with more sample size, sample mean will be more closer to population mean and thus curvature of sample distribution will be more in Normal Distribution.