**Covariance, Correlation & Central Limit Theorem**

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

Covariance is a measure of the joint variability of two random variables. It quantifies the relationship between two random variables.

*For a Population:*

For a Sample:

# **What makes Correlations better than Covariance?**

Correlation is preferred over covariance because it does not get affected by the change in scale. Correlation provides a standardized, absolute measure of the strength of the relationship between two variables, bounded by -1.0 and 1.0 such that the closer the correlation to -1.0 or 1.0 the stronger the relationship.

# **Explain the process as well as Pearson and Spearman Correlation.**

Pearson's correlation coefficient:

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The absolute values of both the sample and population Pearson correlation coefficients are on or between −1 and 1. When graphically represented on a scatter plot, using the Pearson Correlation coefficient, we can draw a straight line on the graph where the slope of the line is equal to this coefficient. The Pearson correlation coefficient is also symmetric, i.e. corr(X,Y) = corr(Y,X).

Pearson's Correlation Coefficient Formula for a Population:

Text

Description automatically generatedText

Description automatically generated

Spearman’s Rank Correlation Coefficient:

The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables; while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not).

The process & calculations are shown in the below file:

<https://github.com/RupertSymss/Statistics1/blob/dcc914e5a39c9fd780729234f953017fc91b68ed/Correlation%20Example.xlsx>

# **What are the advantages of Spearman Correlation over Pearson Correlation?**

While the Pearson's Correlation Coefficient quantifies the relationship between two variables, it is not able to capture a nonlinear relationship. The Spearman's Rank Correlation Coefficient has an advantage that it is not impacted by non-linear variations as it converts all values to an equivalent rank. Thus, the Spearman's Rank Correlation Coefficient expresses better insight into the relationship between two variables.

# **Describe the Central Limit Theorem.**

The central limit theorem (CLT) states that if you have a population with mean μ and standard deviation σ and we take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be approximately normally distributed.

This will hold true regardless of whether the source population is normal or skewed, provided the sample size is sufficiently large (usually n > 30).

If the population is normal, then the theorem holds true even for samples smaller than 30. In fact, this also holds true even if the population is binomial, provided that min(np, n(1-p))> 5, where n is the sample size and p is the probability of success in the population.

A key aspect of CLT is that the average of the sample means and standard deviations will equal the population mean and standard deviation. **Therefore, the advantage of CLT is that we can use the normal probability model to quantify uncertainty when making inferences about a population mean based on the sample mean.**