

**Statistics Assignment 4**

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

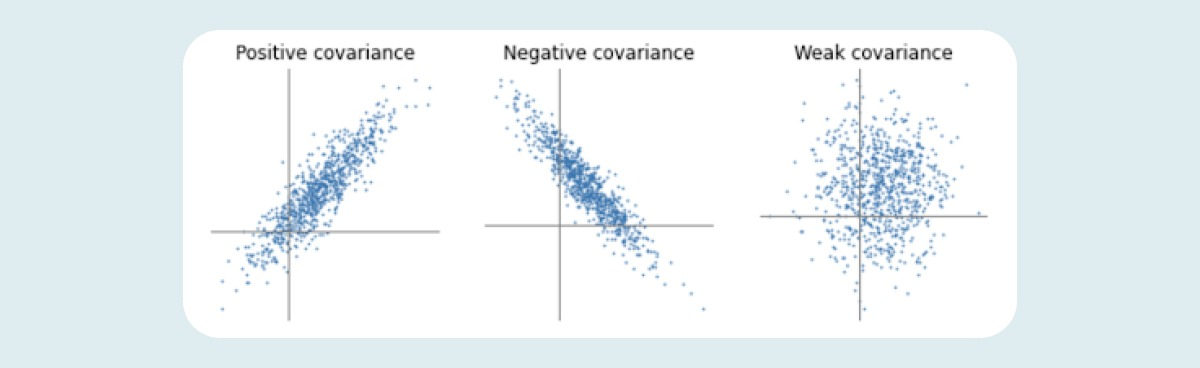
**Answer:**

Covariance measures the direction of relationship between two variables in a data set.

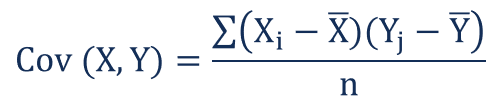
If variance is positive it means both the variables have positive relationship it means if value of one variable increases others variable also increase and vice versa. Both variable goes in same direction

If variance is negative it means both the variables have a negative relationship. It means if the value of one variable increases, others value decreases.

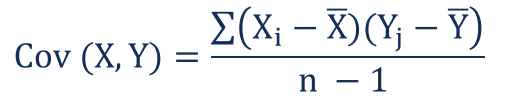
Both variable goes in opposite direction



Population Covariance:



Sample Covariance:



**Xi** – the values of the X-variable

**Yj**– the values of the Y-variable

**X̄** – the mean (average) of the X-variable

**Ȳ** – the mean (average) of the Y-variable

**n** – the number of data points

**2. What makes Correlations better than Covariance?**

**Answer:**

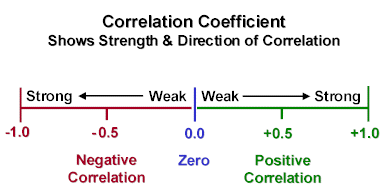
Covariance indicates only the direction of relationship between two variables, positive/negative. However correlation informs us strength and direction between two variables.

The Correlation coefficient strength value indicates within -1 to 1, below chart explain the strength.

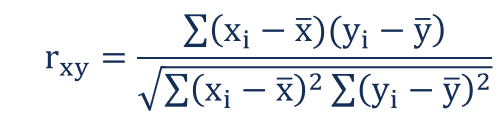
**3. Explain the process as well as Pearson and Spearman Correlation.**

**Answer:**

**Pearson Correlation Coefficient** – This statistic process been used to measures linear relationship between two variables/features in data set. It is ratio between variance of two variables It is also called as bivariate correlation as its result always has in range of -1 to 1. Below is the chart shows the strength & direction as per their ratio result.



To describe the correlation coefficient strength between two variables below formula and process used.



Let’s understand this process with below example:-

As an investor it would be interesting to find the relation between “S&P Index” with “Apple stock” price movements.

To understand this relation we found below data for the last five years.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Correlation Coefficient of S&P Index and Apple Price | | | | |  |  |  |
| Year | S&P 500 Index Price | Apple Price | C | D |  |  |  |
|  | |  | | --- | |  | | |  | | --- | |  | |
|  | (A) | (B) | (A - Mean) | (B - Mean) | (CxD) |  |  |
| 2013 | 1691.75 | 68.96 | -353.19 | -39.93 | 14101.62 | 124746.00 | 1594.09 |
| 2014 | 1977.80 | 100.11 | -67.14 | -8.78 | 589.26 | 4508.32 | 77.02 |
| 2015 | 1884.50 | 109.06 | -160.44 | 0.17 | -27.92 | 25742.28 | 0.03 |
| 2016 | 2151.31 | 112.18 | 106.37 | 3.29 | 350.37 | 11313.73 | 10.85 |
| 2017 | 2519.36 | 154.12 | 474.42 | 45.23 | 21459.73 | 225070.54 | 2046.11 |
| Mean | 2044.944 | 108.886 | Sum | | 36473.07 | 391380.86 | 3728.099 |
|  |  |  |  |  |  |  |  |
| Pearson Correlation Coefficient (rab) | | | 36473.07 |  | 0.955 |  |  |
|  |  | |  | | --- | |  | |  |  |  |  |
|  |  |  |  |  |  |  |  |

The above process indicates that price of Apple and S&P500 has high positive Correlation, it means if S&P500 index increase it has 95.5% chances to increase in price of Apple price. Here 95.5% is indicates the strength of correlation coefficient between both.

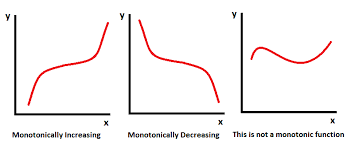
**Spearmen Correlation** - This statistical process used to measure strength and direction associated between two ranked variables. Spearman correlation used to measure strength and monotonic relationship between variables.

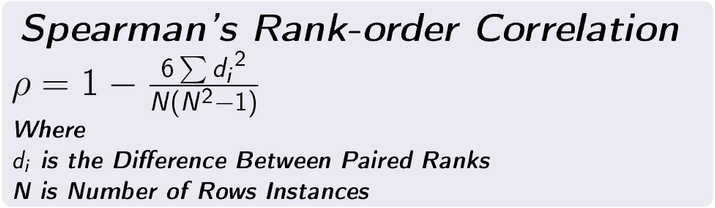
It means

If one variables increases other variable decreases – Monotonic Increasing.

If one variables decreases other variable increases – Monotonic Decreasing.

If once variables increases other sometime increase sometime decreases – Non-Monotonic





Lets understand from below example – Below two data set of two subject Math & Physic of 11 students

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Spearman Correlation Coefficient of Marks in Math & Physic Subject | | | | | | |
| Student | Math | Physic | Rank (A) | Rank (B) | Difference(E) |  |
| |  | | --- | |  | |
|  | (A) | (B) | C | D | (C-D) |  |
| Alen | 68.00 | 55.00 | 9 | 6 | 3.00 | 9.00 |
| Carl | 55.00 | 33.00 | 5 | 1 | 4.00 | 16.00 |
| David | 48.00 | 48.00 | 4 | 4 | 0.00 | 0.00 |
| Don | 35.00 | 71.00 | 3 | 10 | -7.00 | 49.00 |
| John | 26.00 | 41.00 | 1 | 3 | -2.00 | 4.00 |
| Matt | 77.00 | 73.00 | 10 | 11 | -1.00 | 1.00 |
| Mike | 80.00 | 62.00 | 11 | 8 | 3.00 | 9.00 |
| Neal | 29.00 | 39.00 | 2 | 2 | 0.00 | 0.00 |
| Rick | 58.00 | 51.00 | 7 | 5 | 2.00 | 4.00 |
| Rob | 62.00 | 59.00 | 8 | 7 | 1.00 | 1.00 |
| Bob | 56.00 | 63.00 | 6 | 9 | -3.00 | 9.00 |
|  |  |  | Sum | |  | 102 |
|  |  | |  | | --- | |  | |  |  |  |  |
| Spearman Correlation (*p)* | | |  |  | 0.5364 |  |
|  |  |  |  |  | 0.5364 |  |
|  |  |  |  |  |  |  |

It means Math & Physic mark of students has positive correlation since its value falls between -1 to 1.

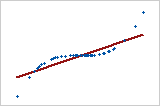
4. What are the advantages of Spearman Correlation over Pearson Correlation?

**Answer:** Although both statistics method informs correlation and strength between two variables, however most prominent difference between both is

In Pearson correlation we can find linear relationship between two variables whether positive or negative. It means once increase other increase or once decrease other also decrease.

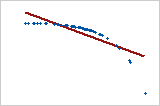
In Spearman correlation evaluate the monotonic relationship between tow variables, it means once increase other increase some time and sometime decrease as well, The variable changes together but not at constant rate.

If you find correlation value of one data set in Pearson and Spearman the value will be shown differently as below



**Value will be shown as : Pearson = +0.85 (Indicates in Red line) and Spearman = +1 (Indicates in Blue dot line)**

If the relationship is that one variable increases when the other increases, but the value is not consistent



**Value will be shown as : Pearson = - 0.799 (Indicates in Red line) and Spearman = - 1 (Indicates in Blue dot line)**

5. Describe the Central Limit Theorem

**Answer:**

CLT states that no matter what will be shape of your population distribution, if you draw large amount of samples (normally largest than 30), the average of those sample approach to normal distribution. The average of those sample means equal to that population mean.

In other words, add up the means from all of your samples, find the average and that average will be your actual population mean. Similarly, if you find the average of all of the standard deviations in your sample, you’ll find the actual standard deviation for your population. It’s a pretty useful phenomenon that can help accurately predict characteristics of a population.

**Example** - A certain group of welfare recipients receives SNAP benefits of $110 per week with a standard deviation of $20. If a random sample of 25 people is taken, what is the probability their mean benefit will be greater than $120 per week?

CLTzvalue

Step 1: Insert the information into the z-formula: = (120-110)/20 /√25 = 10/ (20/5) = 10/4 = 2.5.  
Step 2: Value in the z-score table of 2.5 has an area of roughly 49.38%. Adding 50% (for the left half of the curve), we get 99.38%.