Day-7 **Statistics for Data Science**

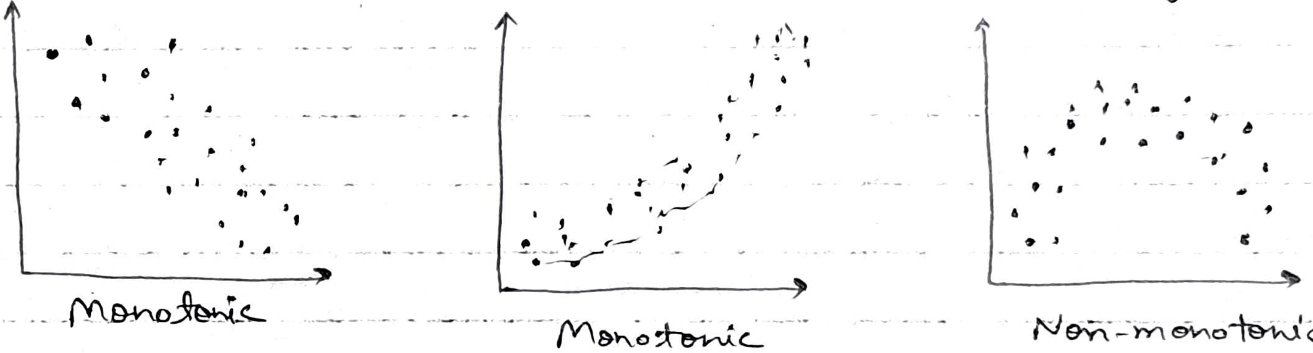
Spearman Rank Correlation :

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$$

Spearman's correlation determines the strength and direction of the monotonic relationship between two variables rather than the strength and direction of the linear relationship between two variables, which is what Pearson's correlation coefficient determines

What is a monotonic relationship?

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 other variable increases; or (2) as the value of one variable increases, the value of other variable decreases. Example of monotonic and non-monotonic relationships are represent in the diagram below:



Calculation of Spearman rank correlation :

Marks { Math : 30, 33, 45, 23, 8, 49, 12, 4, 31
Physics : 35, 23, 47, 17, 10, 43, 9, 6, 28

Step 1: Find the rank for each individual subject. (Rank from higher to lower)

Physics	Rank	Math	Rank
35	3	30	5
23	5	33	3
47	1	45	2
17	6	23	6
10	7	8	8
43	2	49	1
9	8	12	7
6	9	4	9
28	4	31	4

Step2: Add a third column 'd' to the data where d is the difference between ranks. Also add fourth column having value d^2.

Rank (Physics)	Rank (Math)	d	d^2
3	5	2	4
5	3	2	4
1	2	1	1
6	6	0	0
7	8	1	1
2	1	1	1
8	7	1	1
9	9	0	0
4	4	0	0

$\boxed{n=9}$

$\sum d_i^2 = 12$

$\therefore \rho = 1 - \frac{6 \times 12}{9(81-1)} = 1 - \frac{\cancel{72}\,8}{\cancel{9} \times \cancel{80}\,10} = 1 - 0.1 = 0.9.$

→ Central Limit Theoram

The central limit theoram states that the sampling distribution of the sample means approaches a normal distribution. This fact holds especially true for sample sizes over 30 or equal to 30.

In common words, it can be stated as —

As we take more samples, especially large ones, our graph of the sample means tends to look like a normal distribution graph.

Use case: Suppose we have a class consisting of 10,000 students. To find the mean height of students, we can use the CLT as —

1. Divide the students into groups randomly having sample size >= 30. (Divide multiple samples, may be around 150).
2. Now, calculate mean of each sample.
3. Now, calculate mean of these sample means.
4. This sampled mean of the (all divided samples) will give us approximate mean of the height of all students in class.

The histogram of the sample mean weights of students will resemble a bell curve (normal distribution).