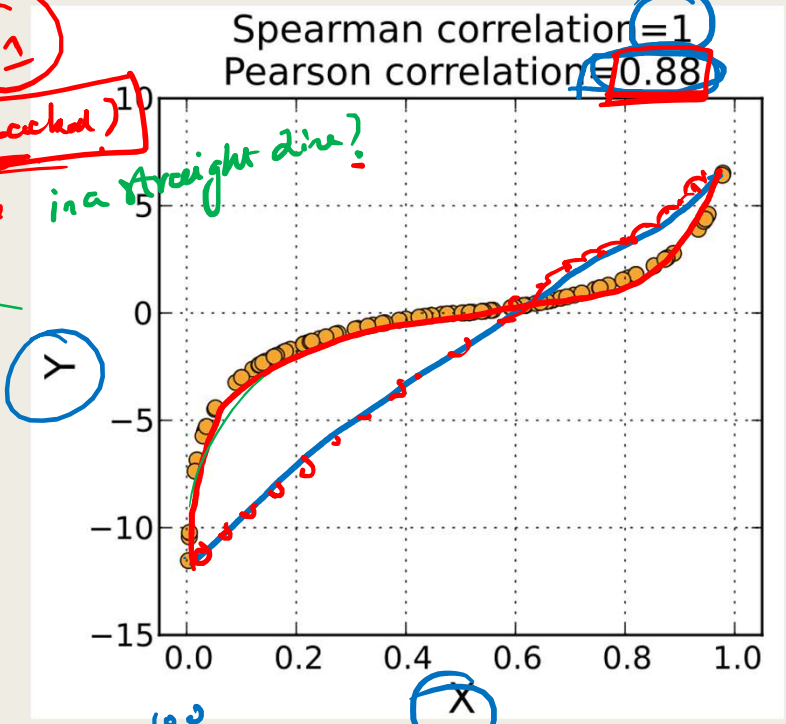
## Spearman Rank Correlation

- Used for Non-Linear Variables
- Spearman Corr Coeff = Pearson Corr coeff (rank varaibles)
- In Python: *DataFrame.corr(method='spearman')*

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

- Denoted by rho.



Steps for Spearman Correlation Coefficient

1. Create a new column for rank(x) and assign the rank of each variable.
2. Assign the rank of 2nd variable in a new column rank(y).
3. Calculate the difference in rank of both the variables = d.
4. Calculate the d-squared.
5. Add up d-squared score.
6. Put in the formula provided:

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

diff. in Rank.

of values

# of observations.

Question: The scores for 10 students in English and Maths are as follows:

| | Marks | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| English | 56 | 75 | 45 | 71 | 62 | 64 | 58 | 80 | 76 | 61 |
| Maths | 66 | 70 | 40 | 60 | 65 | 56 | 59 | 77 | 67 | 63 |

Compute the Spearman rank correlation.

Solution:

Step 1,2,3 and 4:

| English (mark) | Maths (mark) | Rank (English) | Rank (maths) | d | $d^2$ |
|---|---|---|---|---|---|
| 56 | 66 | 9 | 4 | 5 | 25 |
| 75 | 70 | 3 | 2 | 1 | 1 |
| 45 | 40 | 10 | 10 | 0 | 0 |
| 71 | 60 | 4 | 7 | −3 | 9 |
| 62 | 65 | 6 | 5 | 1 | 1 |
| 64 | 56 | 5 | 9 | -4 | 16 |
| 58 | 59 | 8 | 8 | 0 | 0 |
| 80 | 77 | 1 | 1 | 0 | 0 |
| 76 | 67 | 2 | 3 | -1 | 1 |
| 61 | 63 | 7 | 6 | -1 | 1 |

Solution Contd.

Step 5:

$$\sum d_i^2 = 25 + 1 + 9 + 1 + 16 + 1 + 1 = 54$$

Step 6:

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

$$\rho = 1 - \frac{6 \times 54}{10(10^2 - 1)}$$

$$\rho = 1 - \frac{324}{990}$$

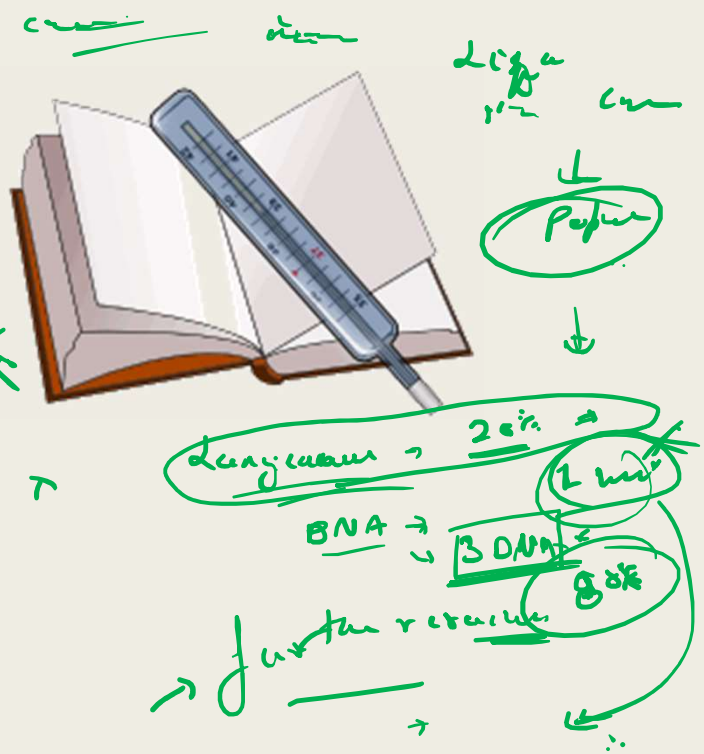
$$\rho = 1 - 0.33$$

$$\rho = 0.67$$

Hence, the Spearman Rank Coefficient is 0.67.

# Points to Ponder (Correlation):

**Example:** A few years ago a survey of employees found a strong positive correlation between "Studying an external course" and Sick Leave Days.



Does this mean?

- Studying makes them sick?
- Sick people study a lot?
- Or did they lie about being sick so they can study more?

Without further research we can't be sure why. :-D

**Example:** Poor suburbs are more likely to have high pollution.

Why?

- Do poor people make pollution?
- Are polluted suburbs the only place poor people can afford?
- Is it a common link, such as factories with low paying jobs and lots of pollution?

# Recap – Descriptive Statistics

- Statistics? Its Importance. Population vs Sample.
- Types of variable – *(Quantitative, Categorical), - (Ordinal, Nominal), (discrete, continuous).*
- Types of charts – Pie, Donut, Line, Scatterplot, Histogram, Bar chart, Box-plot
- Descriptive Stats:

  *Measure of central tendency -- Mean , Median & Model*
  *Measures of Dispersion/spread – Standard Deviation, Variance, range & IQR*
  *Measures of symmetricity – Skewness and Kurtosis*

- 5 Number Summary – Box Plot (Box and Whiskers)
- Effect of transformation on central tendency and spread.
- Outliers? How to detect? Modified Boxplot. (IQR Method)
- Covariance, Correlation. Pearson's Correlation Coefficient. Nature & Strength of Correlation. How to calculate Pearson's Correlation Coefficient and Spearman's Rank Correlation Coefficient.