PROBABILITY AND STATISTICS

PRESENTAION ON

CORRELATION COEFFICIENT

SUBMITTED BY: AYUSH MENDIRATTA (2K21CSUN01057)

ADITYA GAUR(2K21CSUN01052)

DEVANSH TIWARI(2K21CSUN01065)

CLASS: CSE-4B

SUBMITTED TO: DR SAVITA SAINI

CORRELATION COEFFICIENT

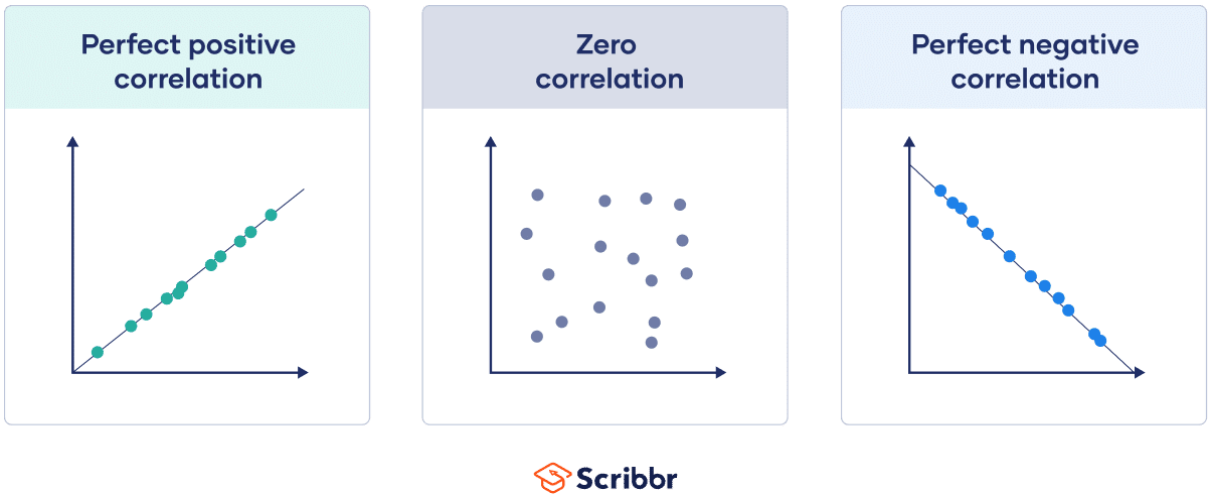
A correlation coefficient is a number between -1 and 1 that tells you the strength and direction of a relationship between variables.

In other words, it reflects how similar the measurements of two or more variables are across a dataset.

A correlation coefficient is a numerical measure of some type of correlation, meaning a statistical relationship between two variables. The variables may be two columns of a given data set of observations, often called a sample, or two components of a multivariate random variable with a known distribution.

Correlation coefficient value	Correlation type	Correlation meaning
1	Perfect positive correlation	When one variable changes, the other variables change in the same direction.
0	Zero correlation	There is no relationship between the variables.

-1	**Perfect negative correlation**	**When one variable changes, the other variables change in the opposite direction.**



THERE ARE TWO MAIN WAYS TO FIND OUT THE COEFFICIENT CORRELATION

- **PEARSON**
- **RANK CORRELATION**

PEARSON'S CORRELATION COEFFICIENT

The Pearson's product-moment correlation coefficient, also known as Pearson's r, describes the linear relationship between two quantitative variables.

These are the assumptions your data must meet if you want to use Pearson's r:

- **Both variables are on an interval or ratio level of measurement**
- **Data from both variables follow normal distributions**
- **Your data have no outliers**
- **Your data is from a random or representative sample**
- **You expect a linear relationship between the two variables**

The Pearson's r is a parametric test, so it has high power. But it's not a good measure of correlation if your variables have a nonlinear relationship, or if your data have outliers, skewed distributions, or come from categorical variables. If any of these assumptions are violated, you should consider a rank correlation measure.

The formula for the Pearson's r is complicated, but most computer programs can quickly churn out the correlation coefficient from your data. In a simpler form, the formula divides the covariance between the variables by the product of their standard deviations.

Pearson sample vs population correlation coefficient formula

$$r_{xy} = \frac{cov(x, y)}{s_x s_y}$$

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

- r_{xy} **= strength of the correlation between variables x and y**
- n **= sample size**

- $\sum$ = sum of what follows…
- X = every x-variable value
- Y = every y-variable value
- XY = the product of each x-variable score and the corresponding y-variable score

RANK CORRELATION OR SPEARMAN'S RHO CORRELATION COEFFICIENT

Spearman's rho, or Spearman's rank correlation coefficient, is the most common alternative to Pearson's *r*. It's a rank correlation coefficient because it uses the rankings of data from each variable (e.g., from lowest to highest) rather than the raw data itself.

You should use Spearman's rho when your data fail to meet the assumptions of Pearson's *r*. This happens when at least one of your variables is on an ordinal level of measurement or when the data from one or both variables do not follow normal distributions.

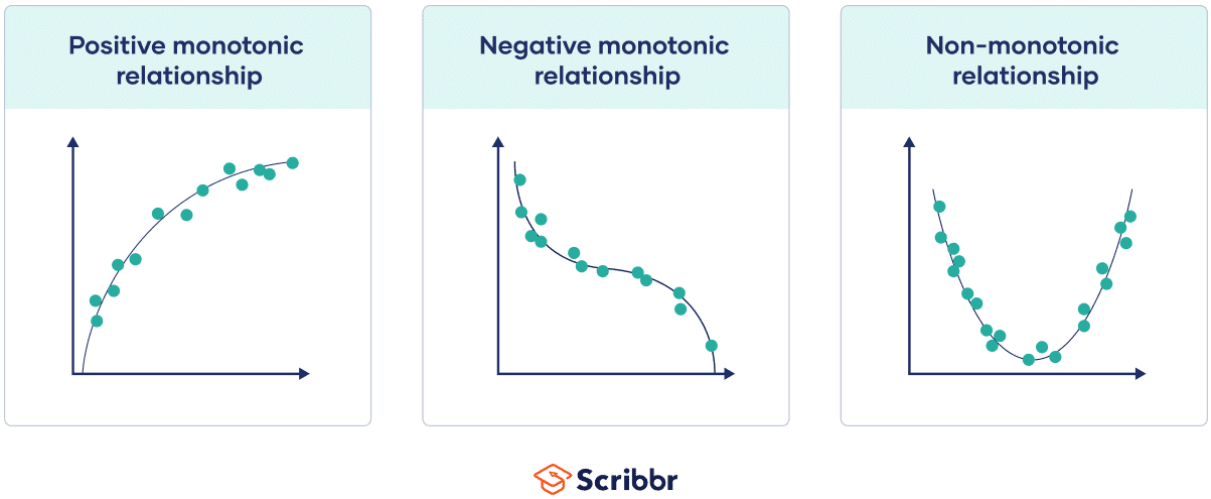
While the Pearson correlation coefficient measures the linearity of relationships, the Spearman correlation coefficient measures the monotonicity of relationships.

In a linear relationship, each variable changes in one direction at the same rate throughout the data range. In a monotonic relationship, each variable also always changes in only one direction but not necessarily at the same rate.

- **Positive monotonic: when one variable increases, the other also increases.**

- **Negative monotonic: when one variable increases, the other decreases.**

Monotonic relationships are less restrictive than linear relationships.



Spearman's rank correlation coefficient formula

The symbols for Spearman's rho are ρ for the population coefficient and r_s for the sample coefficient. The formula calculates the Pearson's *r* correlation coefficient between the rankings of the variable data.

To use this formula, you'll first rank the data from each variable separately from low to high: every datapoint gets a rank from first, second, or third, etc.

Then, you'll find the differences (d_i) between the ranks of your variables for each data pair and take that as the main input for the formula.

$$r_s = 1 - \frac{6 \sum d_i^2}{(n^3 - n)}$$

- **r_s= strength of the rank correlation between variables**
- **d_i = the difference between the x-variable rank and the y-variable rank for each pair of data**
- **$\sum d^2_i$ = sum of the squared differences between x- and y-variable ranks**
- ***n* = sample size**