Correlation Analysis
Pearson Correlation & Spearman rank correlation

DR.ALAA MOHAMMED

3RD LEVEL, 2ND SEMESTER, BIOMEDICAL INSTRUMENTATION AND BIOMECHANICS BRANCHES, BIOMEDICAL ENG. DEPARTMENT

Correlation Analysis

PART 1

Correlation

A **correlation** is a relationship between two variables.

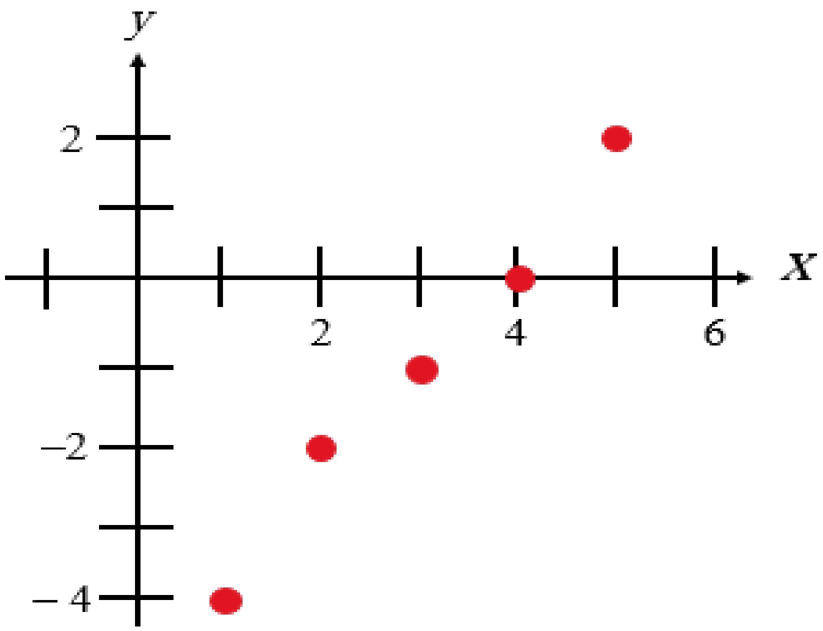
The data can be represented by the ordered pairs (x, y), where x is the **independent variable**, and y is the **dependent** (or **response**) **variable**.

A *scatter plot* can be used to determine whether a linear (straight line) correlation exists between two variables

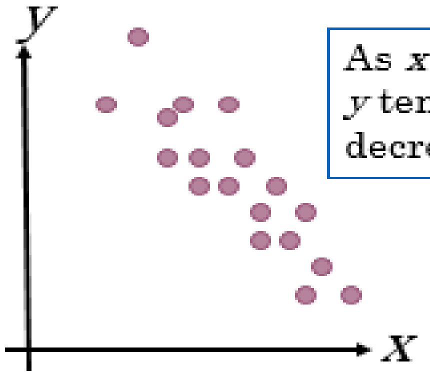
Mathematically, the strength and direction of a linear relationship between two variables is represented by the **correlation coefficient**.

Example

x	1	2	3	4	5
y	– 4	– 2	– 1	0	2

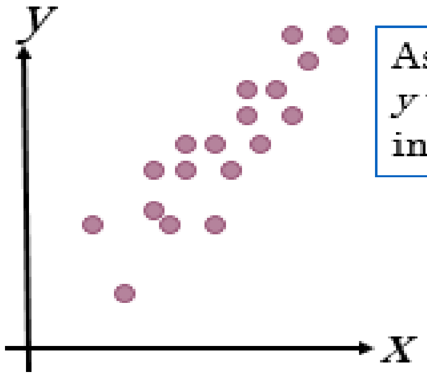


Types of Correlations



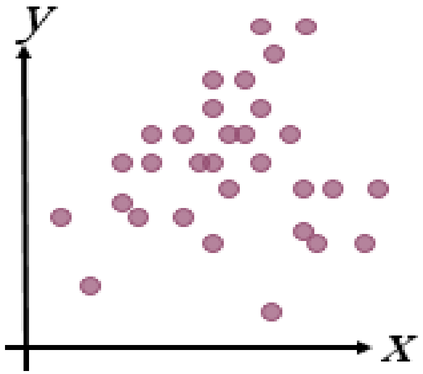
As x increases, y tends to decrease.

Negative Linear Correlation

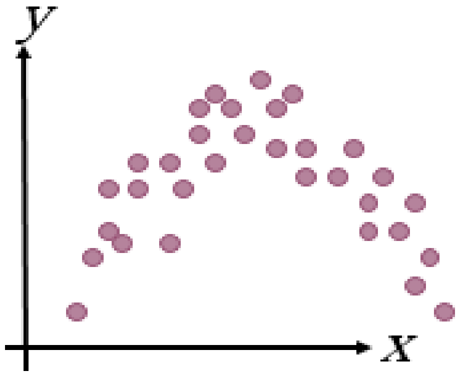


As x increases, y tends to increase.

Positive Linear Correlation



No Correlation



Nonlinear Correlation

Correlation Coefficient

The ***correlation coefficient*** is a measure of the strength and the direction of a linear relationship between two variables.

Statistic showing the degree of relation between two variables

The symbol $\boldsymbol{r}$ represents the sample ***correlation coefficient***.

Suppose that there are n ordered pairs $(\boldsymbol{x};\ \boldsymbol{y})$ that make up a sample from a population.

The correlation coefficient $\boldsymbol{r}$ is given by:

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}}.$$

Correlation Coefficient (cont.)

This will always be a number between **-1** and **1** (inclusive).

- If $\boldsymbol{r}$ is close to $\mathbf{1}$, we say that the variables are positively correlated.

  This means there is likely a strong linear relationship between the two variables, with a positive slope.

- If $\boldsymbol{r}$ is close to $-\mathbf{1}$, we say that the variables are negatively correlated.

  This means there is likely a strong linear relationship between the two variables, with a negative slope.

- If $\boldsymbol{r}$ is close to $\mathbf{0}$, we say that the variables are not correlated.

  This means that there is likely no linear relationship between the two variables, however, the variables may still be related in some other way.

Correlation Coefficient (cont.)

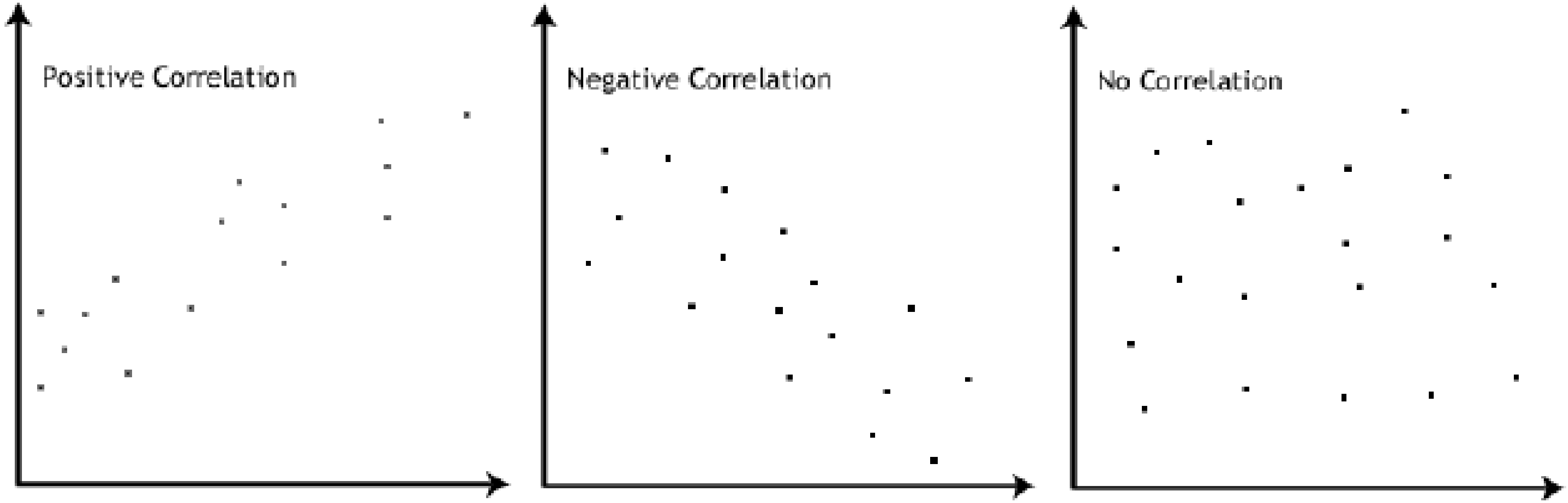
The sign $(+, -)$ of the correlation coefficient indicates the direction of the association.

The magnitude of the correlation coefficient indicates the strength of the association, e.g. A correlation of $r = -0.8$ suggests a strong, negative association (reverse trend) between two variables, whereas a correlation of $r = 0.4$ suggest a weak, positive association.

A correlation close to zero suggests no linear association between two continuous variables.

The correlation coefficient of the population is denoted by $\boldsymbol{\rho}$ and is usually unknown.

Correlation Coefficient (cont.)



Calculating a Correlation Coefficient

Procedure:

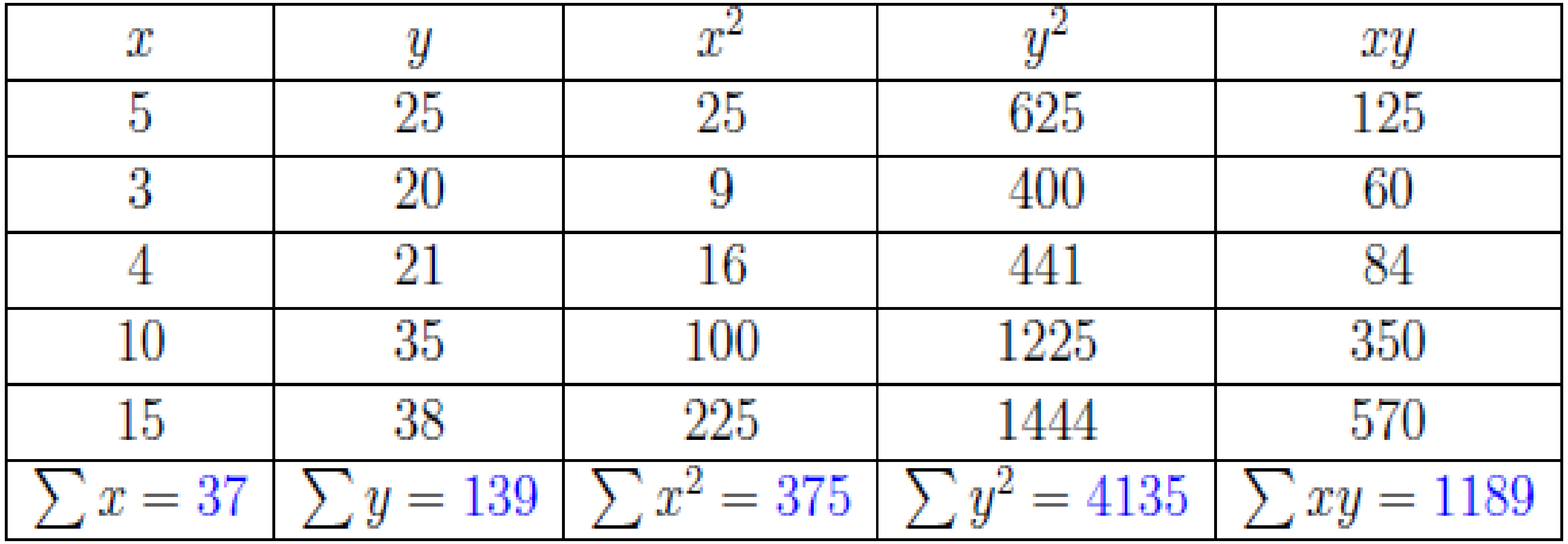
1. Find the sum of the *x*-values ($\sum x$)
2. Find the sum of the *y*-values ($\sum y$)
3. Multiply each *x*-value by its corresponding *y*-value and find the sum ($\sum xy$)
4. Square each *x*-value and find the sum ($\sum x^2$)
5. Square each *y*-value and find the sum ($\sum y^2$)
6. Use these five sums to calculate the correlation coefficient:

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}}.$$

Example 1

The time x in years that an employee spent at a company and the employee's hourly pay, y, for 5 employees are listed in the table below.

Calculate and interpret the correlation coefficient r. Include a plot of the data in your discussion.



x	y	x^2	y^2	xy
5	25	25	625	125
3	20	9	400	60
4	21	16	441	84
10	35	100	1225	350
15	38	225	1444	570
$\sum x = 37$	$\sum y = 139$	$\sum x^2 = 375$	$\sum y^2 = 4135$	$\sum xy = 1189$

Example 1 (cont.)

Hint: Calculate the numerator of r :

$$n\sum(xy)-\left(\sum x\right)\left(\sum y\right)=5\cdot 1189-37\cdot 139=802$$

Then calculate the denominator of r :

$$\sqrt{n\sum x^2-\left(\sum x\right)^2}\sqrt{n\sum y^2-\left(\sum y\right)^2}=\sqrt{5\cdot 375-(37)^2}\sqrt{5\cdot 4135-(139)^2}$$
$$=\sqrt{506}\sqrt{1354}\approx 827.72$$

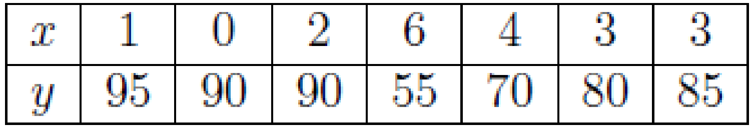
Now, divide to get

$$r\approx\frac{802}{827.72}\approx 0.97$$

Interpret this result: There is a strong positive correlation between the number of years and employee has worked and the employee's salary, since **r** is very close to 1.

Example 2

The table below shows the number of absences, x, in a Calculus course and the final exam grade, y, for 7 students. Find the correlation coefficient and interpret your result



x	1	0	2	6	4	3	3
y	95	90	90	55	70	80	85

You may use the facts that (double check this for practice)

$$\sum x = 19, \qquad \sum y = 565, \qquad \sum x^2 = 75, \qquad \sum y^2 = 46,775, \quad \sum xy = 1,380$$

Calculate the numerator:

$$n\sum(xy) - \left(\sum x\right)\left(\sum y\right) = 7 \cdot 1380 - 19 \cdot 565 = -1075$$

Example 2 (cont.)

Then calculate the denominator:

$$\sqrt{n\sum x^2-\left(\sum x\right)^2}\sqrt{n\sum y^2-\left(\sum y\right)^2}=\sqrt{7\cdot 75-(19)^2}\sqrt{7\cdot 46775-(565)^2}$$

$$=\sqrt{164}\sqrt{8200}\approx 1159.66$$

Now, divide to get $r\approx\dfrac{-1075}{1159.66}\approx -0.93.$

Interpret this result: There is a strong negative correlation between the number of absences and the final exam grade, since r is very close to -1.

Thus, as the number of absences increases, the final exam grade tends to decrease.

Significance level

In linguistic, "**significant**" means important, while in Statistics "significant" means probably true (not due to chance).

When statisticians say a result is "highly significant" they mean it is very probably true. They do not (necessarily) mean it is highly important.

Significance levels show you how likely a pattern in your data is due to chance.

The most common level, used to mean something is good enough to be believed, is "0.95". This means that the finding has a 95% chance of being true which also means that the finding has a confidence degree 95% of being true.

Instead it will show you ".05," meaning that the finding has a five percent (.05) chance of not being true "error", which is the converse of a 95% chance of being true.

Significance level (cont.)

To find the significance level, subtract the number shown from one. For example, a value of ".01" means that there is a confidence degree 99% (1-.01=.99) chance of it being true.

In other words the **significance level** α "alpha level" for a given hypothesis test is a value for which a ***ρ-value*** "calculated value" less than or equal to α is considered statistically significant.

The levels of value correspond to the probability of observing such an extreme value by chance.

For example, if the ***ρ-value*** is 0.0082, so the probability of observing such a value by chance is less that 0.01, and the result is significant at the 0.01 level.

The Population Correlation Coefficient ρ-*value*

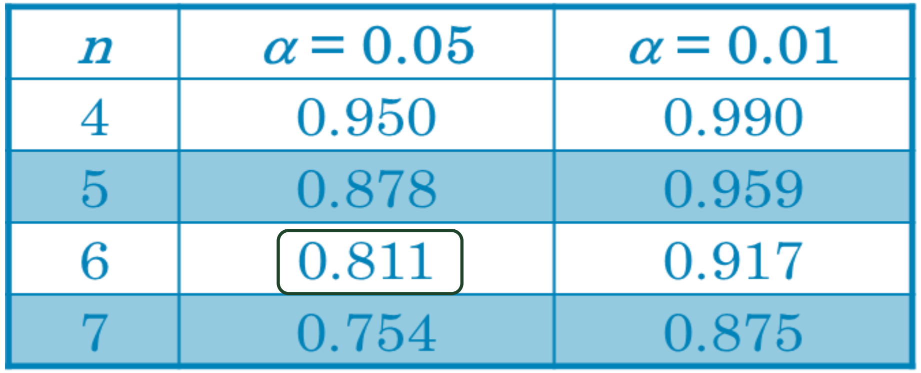
<u>Procedure:</u>

1. Determine the number of pairs of data in the sample (n).
2. Specify the level of significance (Identify α).
3. Find the critical value.
4. Decide if the correlation is significant.
5. Interpret the decision in the context of the original claim (If $|r| \ >$ critical value, the correlation is significant. Otherwise, there is not enough evidence to support that the correlation is significant).

Testing a Population Correlation Coefficient

Once the sample correlation coefficient *r* has been calculated, we need to determine whether there is enough evidence to decide that the population correlation coefficient ***ρ*** is significant at a specified level of significance.

If |*r*| is greater than the critical value, there is enough evidence to decide that the correlation coefficient *ρ* is significant.



n	$\alpha = 0.05$	$\alpha = 0.01$
4	0.950	0.990
5	0.878	0.959
6	$\boxed{0.811}$	0.917
7	0.754	0.875

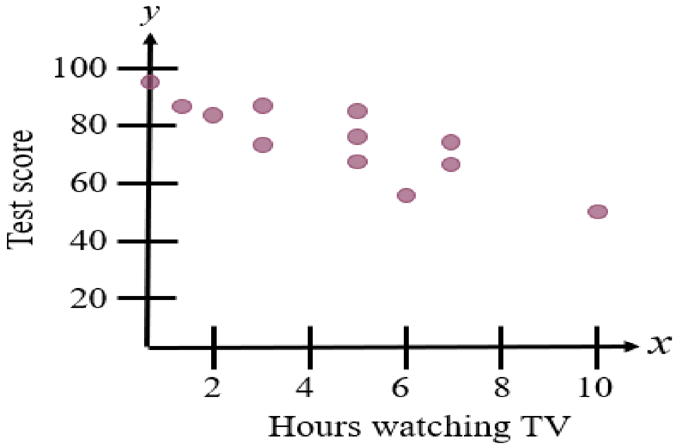
For a sample of size *n* = 6, *ρ* is significant at the 5% significance level, if |*r*| > 0.811.

Example 3

The following data represents the number of hours 12 different students watched television during the weekend and the scores of each student who took a test the following Monday.

Display the scatter plot.. And Calculate the correlation coefficient r.

Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50



Example 3 (cont.)

Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50
xy	0	85	164	222	285	340	380	420	348	455	525	500
x^2	0	1	4	9	9	25	25	25	36	49	49	100
y^2	9216	7225	6724	5476	9025	4624	5776	7056	3364	4225	5625	2500

$\Sigma x = 54$ $\quad \Sigma y = 908$ $\quad \Sigma xy = 3724$ $\quad \Sigma x^2 = 332$ $\quad \Sigma y^2 = 70836$

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}} = \frac{12(3724) - (54)(908)}{\sqrt{12(332) - 54^2}\sqrt{12(70836) - (908)^2}} \approx -0.831$$

There is a strong negative linear correlation.

As the number of hours spent watching TV increases, the test scores tend to decrease.

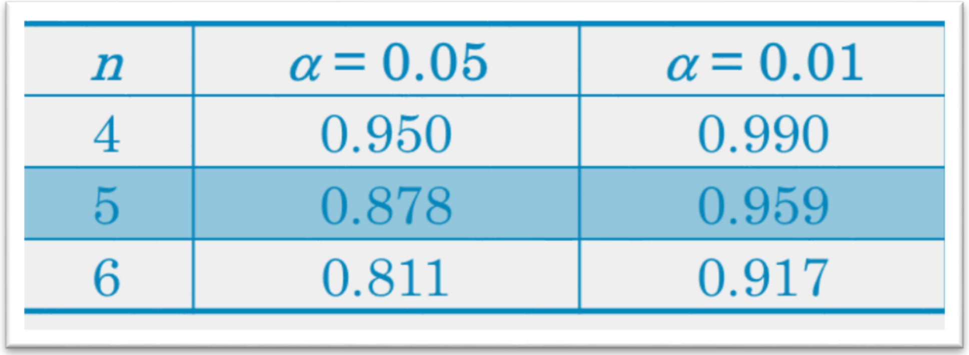
Example 3 (cont.)
Testing a Population Correlation Coefficient

The correlation coefficient $r \approx -0.831$.

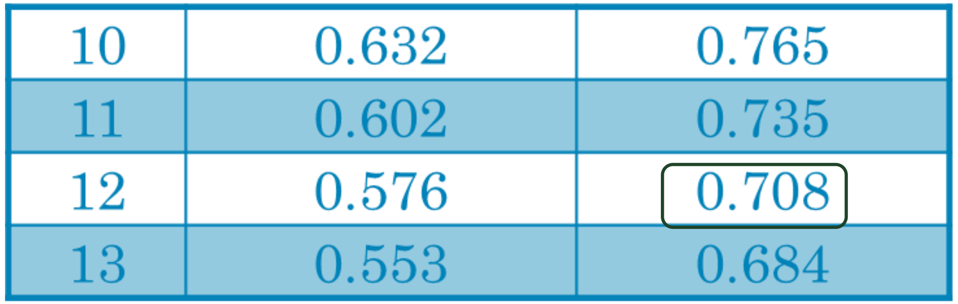
Is the correlation coefficient significant at $\alpha = 0.01$?

$r \approx -0.831$, $n = 12$, $\alpha = 0.01$

Because, the population correlation is significant, there is enough evidence at the 1% level of significance to conclude that there is a significant linear correlation between the number of hours of television watched during the weekend and the scores of each student who took a test the following Monday.



n	$\alpha = 0.05$	$\alpha = 0.01$
4	0.950	0.990
5	0.878	0.959
6	0.811	0.917



10	0.632	0.765
11	0.602	0.735
12	0.576	$\boxed{0.708}$
13	0.553	0.684

$|r| > 0.708$

Pearson Correlation
&
Spearman rank correlation

PART2

Pearson's r or the 'product-moment' correlation coefficient

The standard method (**Pearson correlation**) leads to a quantity called **r** which can take any value from **-1** to **+1**. This **correlation coefficient r** measures the degree of '**straight-line**’ association between the values of two variables.

Thus a value of **+1** or **-1** is obtained if all the points in a scatter plot lie on a perfect straight line.

A commonly employed correlation coefficient are **Pearson correlation, Kendall rank correlation and Spearman correlation**

The **correlation coefficient** usually calculated is called **Pearson's r** or the **'product-moment' correlation coefficient** (other coefficients are used for ranked data, etc.).

Correlation used to examine the **presence of a linear relationship between two variables** providing certain assumptions about the data are **satisfied**.

The **results of the analysis**, however, need to be **interpreted with** care, particularly when looking for a **causal relationship**.

Pearson correlation coefficient

The correlation coefficient is a "standardized score" of the covariance

Then the **Pearson correlation coefficient** is given by the following equation:

$$r = \frac{\text{covariance betwenx and y}}{\left(\begin{array}{c}\text{standard deviation}\\ \text{of x}\end{array}\right)\left(\begin{array}{c}\text{standard deviation}\\ \text{of y}\end{array}\right)}$$

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Where $\bar{x}$ is the mean of variable x values, and $\bar{y}$ is the mean of variable y values

Or

Bivariate correlation

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}}.$$

Spearman rank correlation

- **Spearman rank correlation:** Spearman rank correlation is a non-parametric test that is used to measure the degree of association between two variables. It was developed by Spearman, thus it is called the Spearman rank correlation.

- Spearman rank correlation test does not assume any assumptions about the distribution of the data and is the appropriate correlation analysis when the variables are measured on a scale that is at least ordinal.

The following formula is used to calculate the Spearman rank correlation coefficient:

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

Where:

- ρ = Spearman rank correlation coefficient
- di= the difference between the ranks of corresponding values $X_i$and Y_i
- n= number of value in each data set.

Spearman rank correlation (cont.)

- The Spearman correlation coefficient,ρ, can take values from $+1\ to\ -1$.
- A ρ of $+1$ indicates a perfect association of ranks, a ρ of zero indicates no association between ranks and a ρ of -1 indicates a perfect negative association of ranks.
- The closer ρ to zero, the weaker the association between the ranks.