STATISTICS FOR DATA SCIENCE

Power Test &
Simple Linear Regression

Dr. Karthiyayini

Department of Science and Humanities

STATISTICS FOR DATA SCIENCE

Unit 5 : Power Test & Simple Linear Regression

Session : 5 (Continued Session)

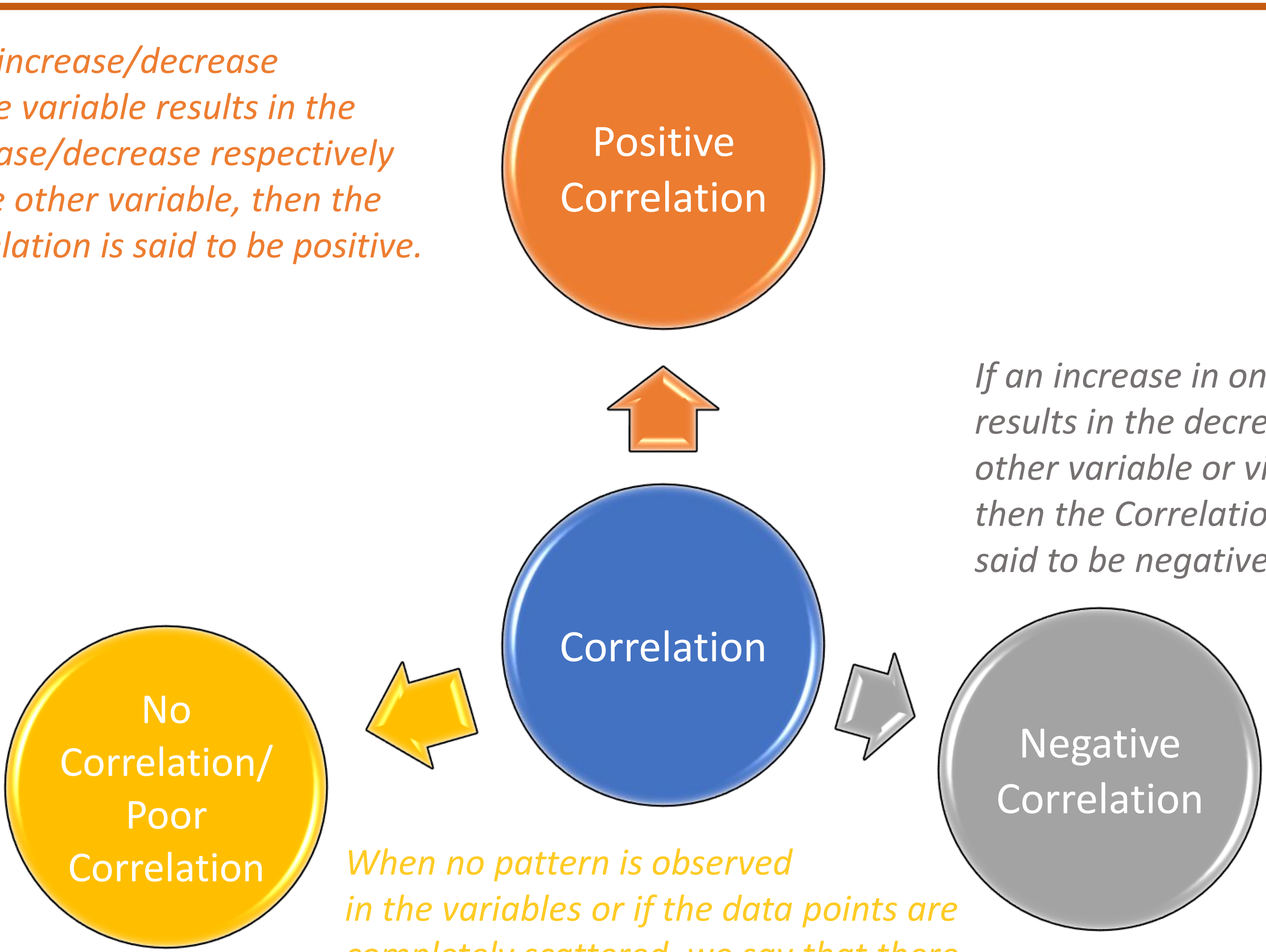
Sub Topic : Correlation

Dr. Karthiyayini

Department of Science & Humanities

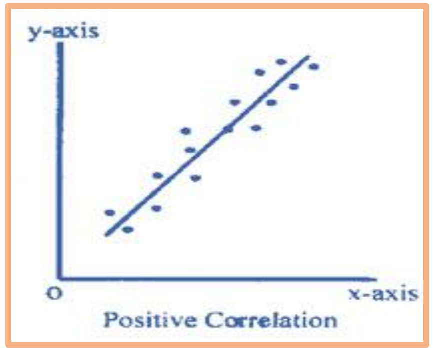
Classification of Correlation

If an increase/decrease in one variable results in the Increase/decrease respectively in the other variable, then the Correlation is said to be positive.

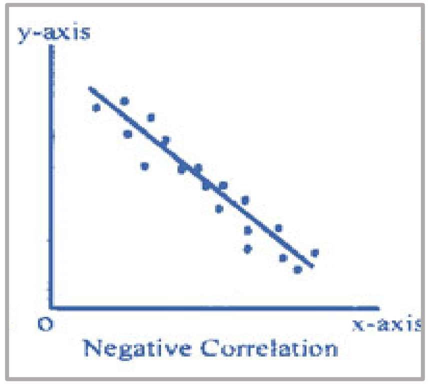


If an increase in one variable results in the decrease in the other variable or vice-versa, then the Correlation is said to be negative.

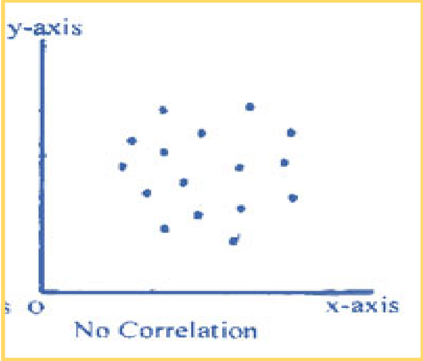
When no pattern is observed in the variables or if the data points are completely scattered, we say that there is no Correlation between the variables.



Positive Correlation



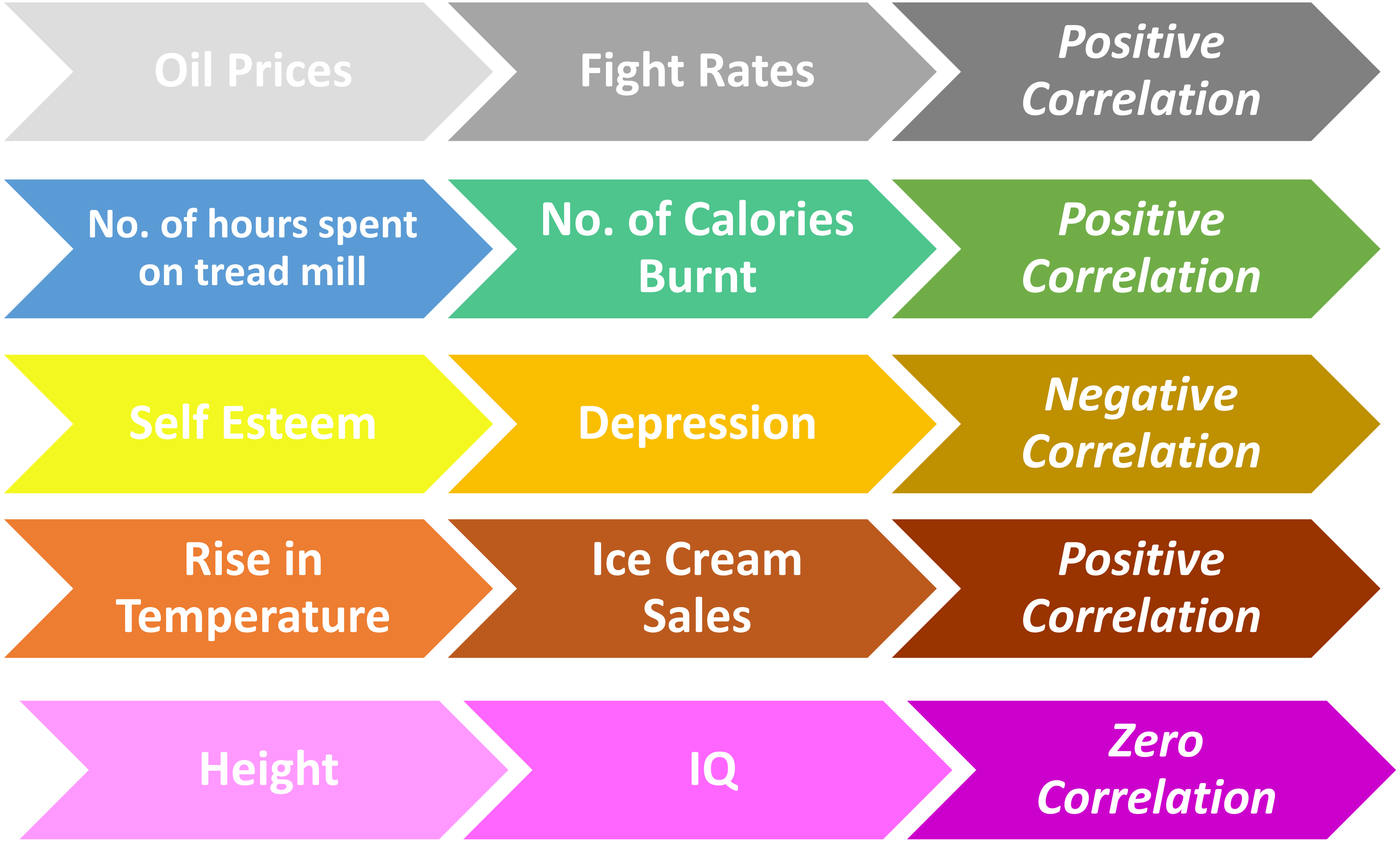
Negative Correlation



No Correlation

Source : emathzone.com

Some Examples :



Oil Prices → Fight Rates → *Positive Correlation*

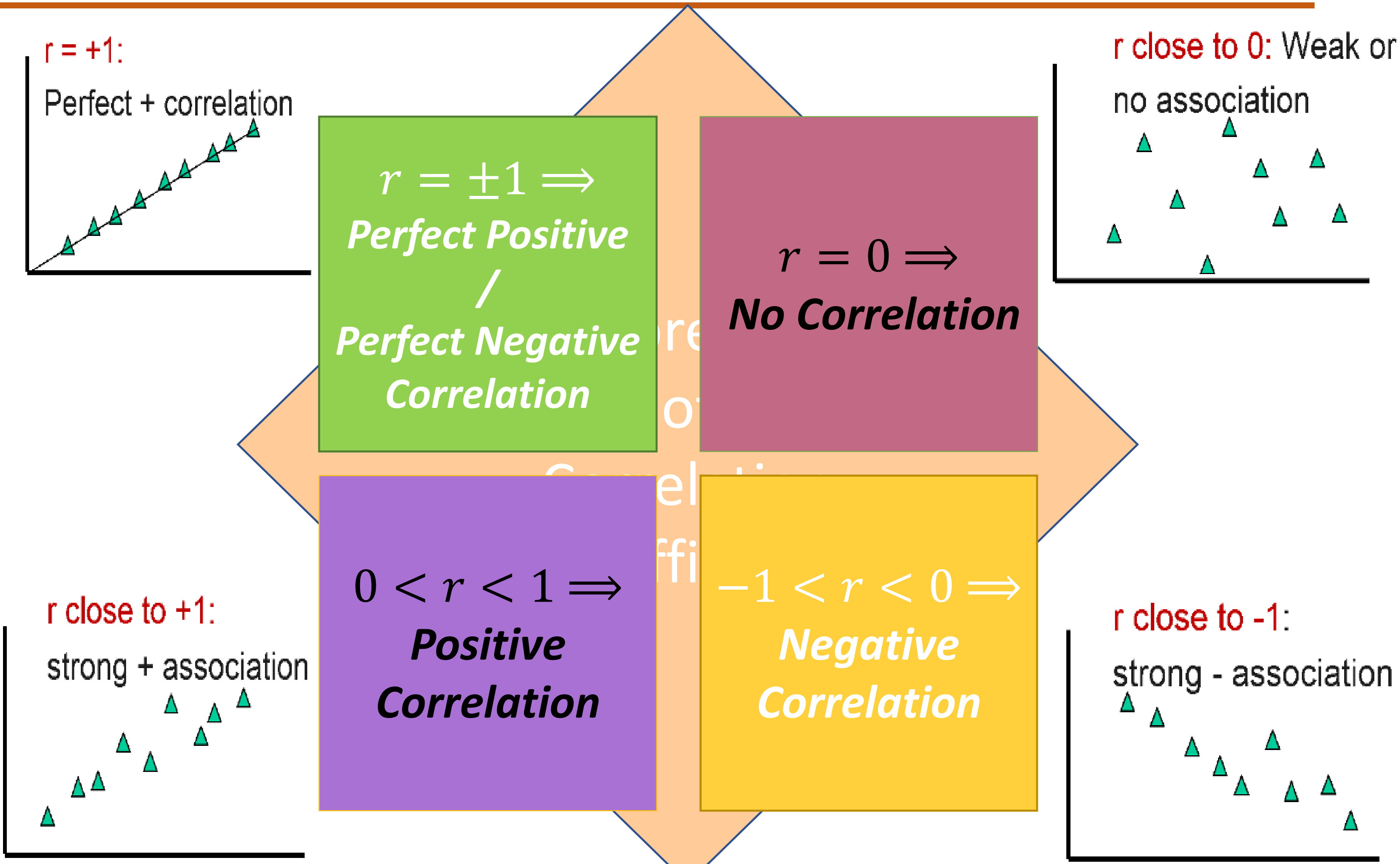
No. of hours spent on tread mill → No. of Calories Burnt → *Positive Correlation*

Self Esteem → Depression → *Negative Correlation*

Rise in Temperature → Ice Cream Sales → *Positive Correlation*

Height → IQ → *Zero Correlation*

Interpretation of Correlation Coefficient



r = +1:

Perfect + correlation

$r = \pm 1 \Longrightarrow$ ***Perfect Positive / Perfect Negative Correlation***

$r = 0 \Longrightarrow$ ***No Correlation***

r close to 0: Weak or no association

r close to +1:

strong + association

$0 < r < 1 \Longrightarrow$ ***Positive Correlation***

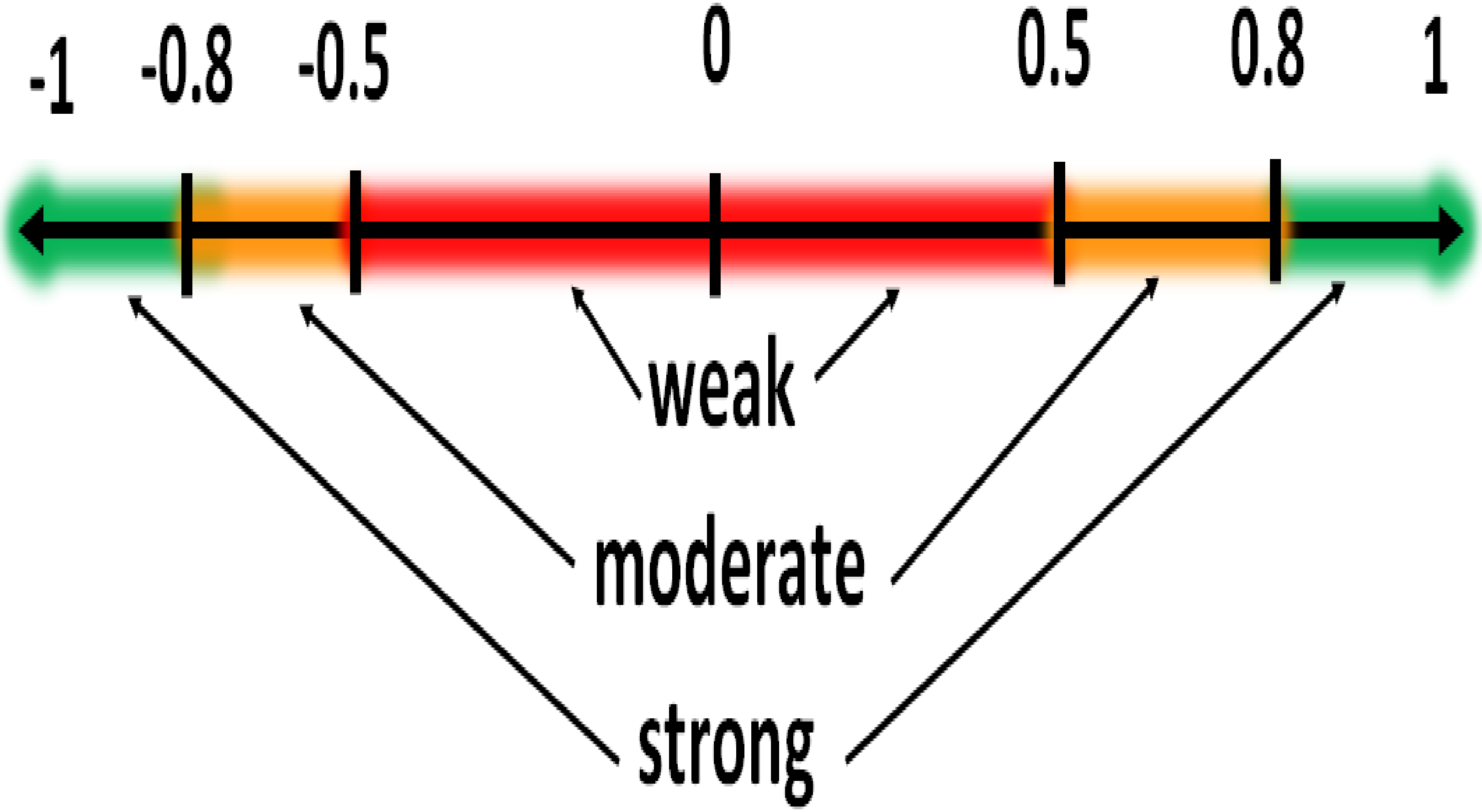
$-1 < r < 0 \Longrightarrow$ ***Negative Correlation***

r close to -1:

strong - association

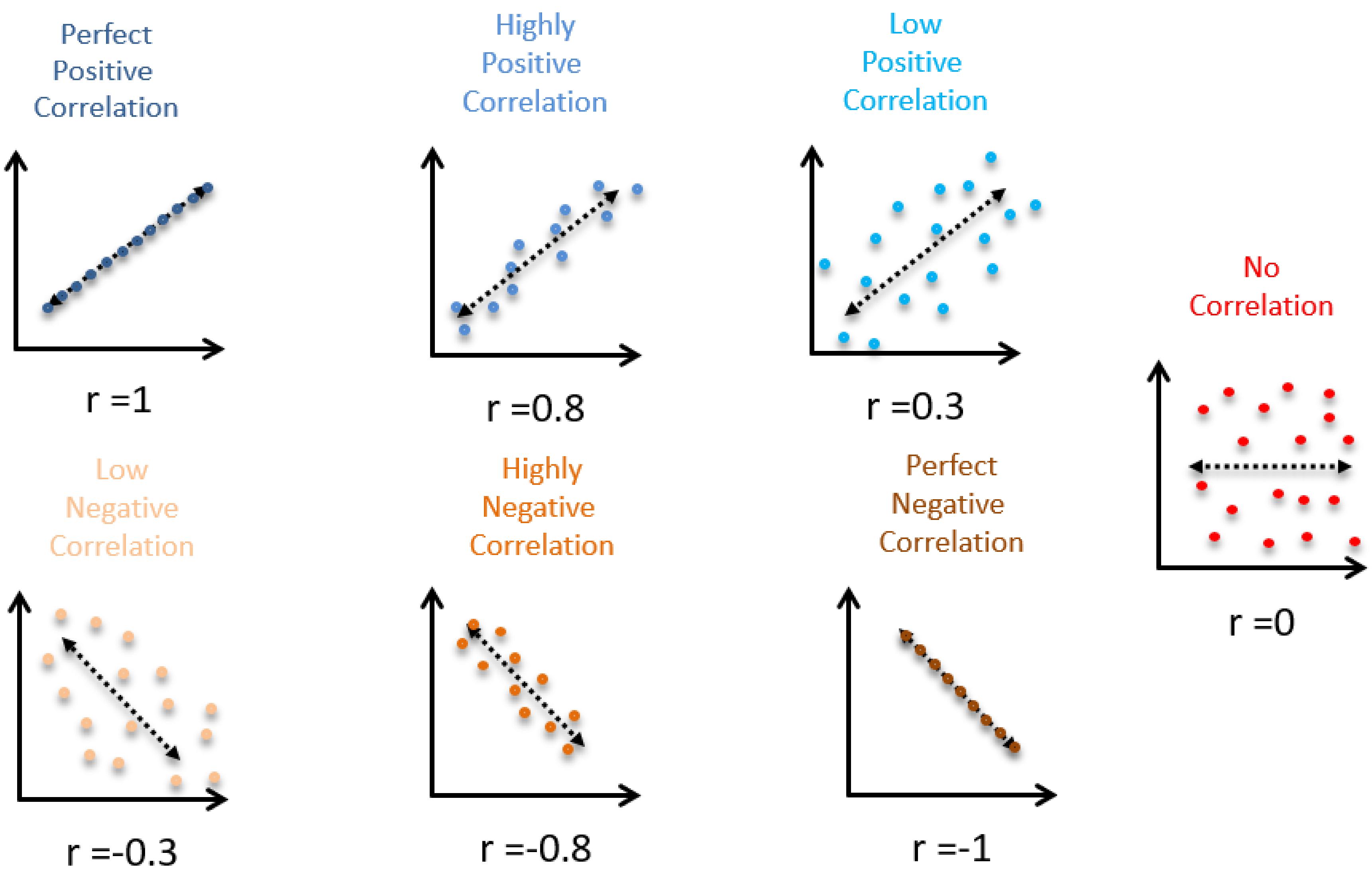
Source : sphweb.bumc.bu.edu

Interpretation of Correlation Coefficient



Source : datavedas.com

Examples of various levels of correlation



Perfect Positive Correlation

r =1

Highly Positive Correlation

r =0.8

Low Positive Correlation

r =0.3

No Correlation

r =0

Low Negative Correlation

r =-0.3

Highly Negative Correlation

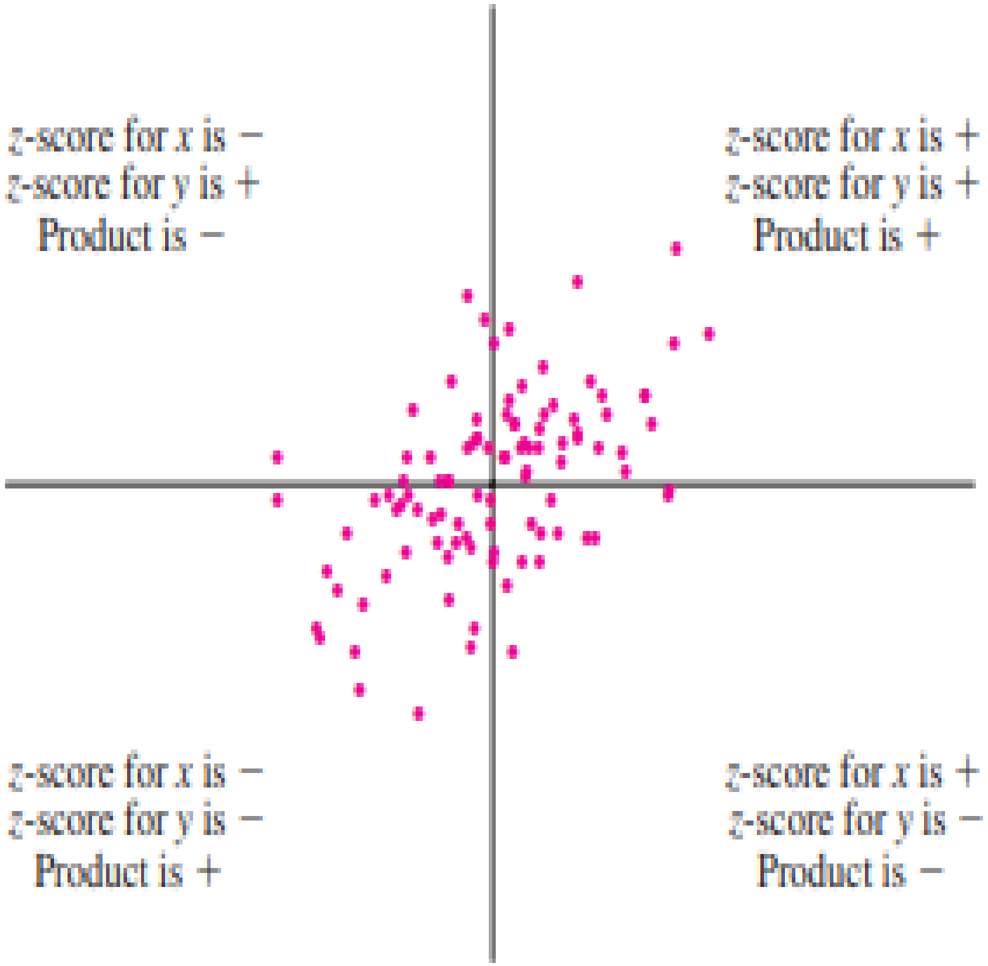
r =-0.8

Perfect Negative Correlation

r =-1

Source : https://lytongblog.wordpress.com/

How the Correlation coefficient works!!



- The origin is placed at the point of averages $(\bar{x}, \bar{y})$.
- The z – scores $\frac{x_i - \bar{x}}{S_x}$ and $\frac{y_i - \bar{y}}{S_y}$ are both positive in the first quadrant, hence their product is also positive.
- Hence the points in the first quadrant contributes a positive amount to the sum in the formula for Correlation coefficient
- On similar lines, we can say that the points in the second and fourth quadrants contribute a negative amount whereas the points in the third quadrant contribute to a positive amount.
- In this scatter plot it can be observed that the number of points in the first and third quadrants are more than that in the second and fourth quadrants.
- Hence the Correlation in this case is “Positive”.

Source : Statistics for Engineers & Scientists, William Navidi

More about Correlation Coefficient

Correlation Coefficient

Sample Correlation ‘r’: If the Correlation coefficient is computed by taking a random sample from a population of points, then it is referred to as “Sample Correlation” and it is an estimate of the population Correlation

Population Correlation $'\rho'$: The population Correlation can be computed by replacing the sample means by population means in the Pearson’s Correlation coefficient formula.

Intuitively, you may imagine the population to consist of large finite collection of points.

Note that Sample Correlation is not only used to measure the strength of a relationship but is also used *to construct Confidence intervals and perform Hypothesis testing* on the population correlation.

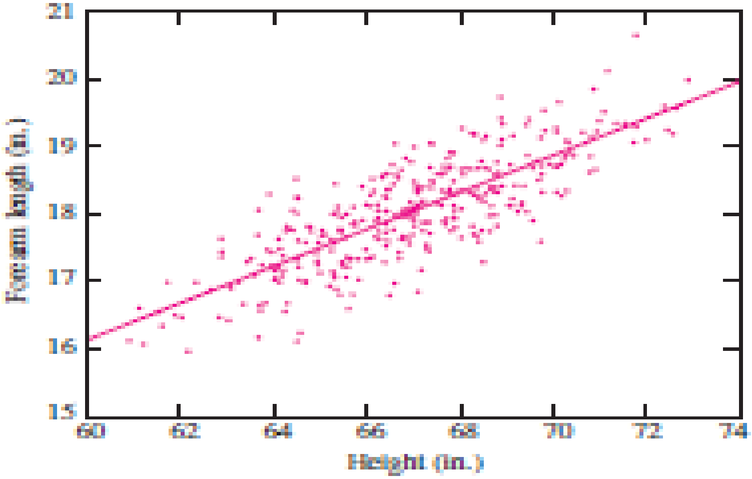
Source : shutterstock.com, clker.com

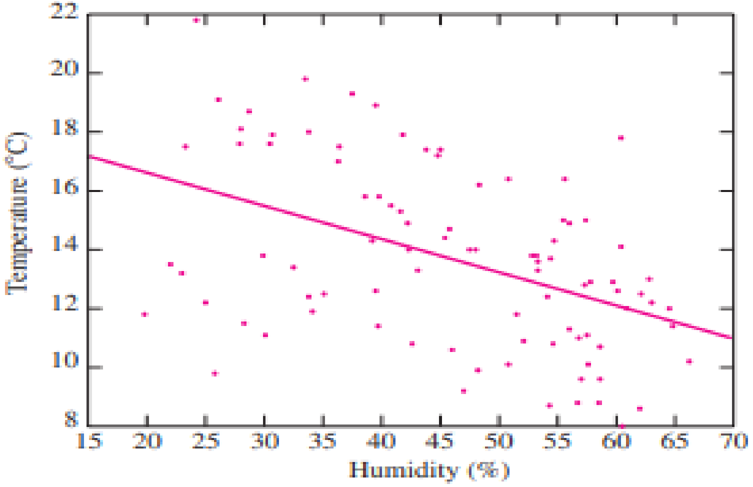
The Correlation Coefficient is unitless!!

- Consider the Correlation Co-efficient given by,

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

- Note that the units of the numerator and the denominators of $\left(\frac{x_i-\bar{x}}{s_x}\right)$ will be the same as that of ‘x’ and hence gets cancelled. The same is true even in the case of $\left(\frac{y_i-\bar{y}}{s_y}\right)$. Therefore the Correlation coefficient ‘r’ is “unitless”.
- This allows the comparison of between 2 different sets of Data.
- In the Galton’s case study of correlation between the heights and forearm lengths of 348 men, the correlation coefficient = 0.80
- In another study of the relationship between the humidity and temperature for a certain period of days in a certain city, the correlation coefficient = - 0.46
- Though in the *Galton’s case study the measurements are made in inches* and in the second case *measurements are in degree Celsius*, we can compare the 2 and say that the relationship between heights of men and their forearm lengths is more strongly linear than the relationship between humidity and temperature.[See the scatter plots]





Source : Statistics for Engineers & Scientists, William Navidi

Some more properties of the Correlation Coefficient

The Correlation coefficient remains unaltered in the following cases :

❖ Interchanging the values of x and y.

- Since the Correlation coefficient is computed taking the product of the z – scores, it does not matter which variable is represented by ‘x’ and which by ‘y’.

- For instance, in the Galton’s case study the correlation coefficient will remain unaltered even if the forearm is represented by ‘x’ and the heights by ‘y’.

Some more properties of the Correlation Coefficient

❖ Adding a constant to each value of a variable

- In the Galton’s case study, suppose the heights of each man is measured by making him stand on a 2 inches high platform.
- Then each x_i would increase by ‘2’. Also $\bar{x}$ will increase by ‘2’.
- Hence the z – score will remain unchanged.
- So the Correlation coefficient also remains unaltered .

This implies that adding a constant to each value of a variable does not change the Correlation coefficient.

Some more properties of the Correlation Coefficient

❖ Multiplying each value of a variable by a positive constant.

- In the Galton’s case study, suppose the heights of each man is measured in centimeters rather than inches.
- Then each x_i would multiplied by ‘2.54’.
- This would cause both $\bar{x}$ and to be multiplied ‘2.54’.
- Hence the z – score will remain unchanged.
- So the Correlation coefficient also remains unaltered .

This implies that Multiplying each value of a variable by a positive constant does not change the Correlation coefficient.

PES
UNIVERSITY
ONLINE

THANK YOU

Dr. Karthiyayini
Department of Science & Humanities

Karthiyayini.roy@pes.edu

+91 80 6618 6651