Dimensionless Technologies Private Limited
Visit us at: www.dimensionless.in
info@dimensionless.in
9923170071, 8108094992

Notes For Students - Lesson 14
Correlation

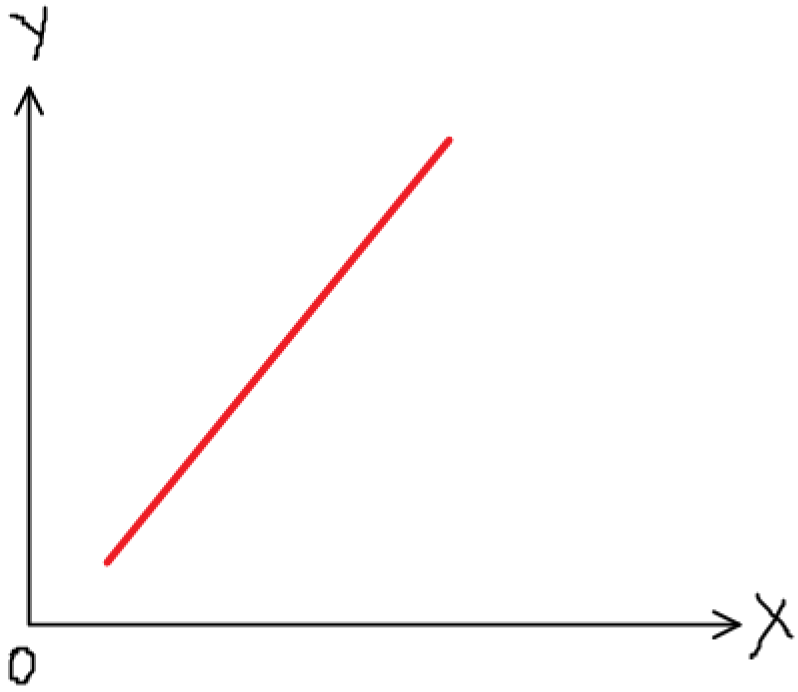
Till now, we have studied how to analyse if the population parameter is different from what is believed to be or if two or more population parameters are significantly different from each other.

However, there may be instances where rather than testing if population parameters are significantly different to each other or not, we want to analyse if they are related to each other or not and to what extent.

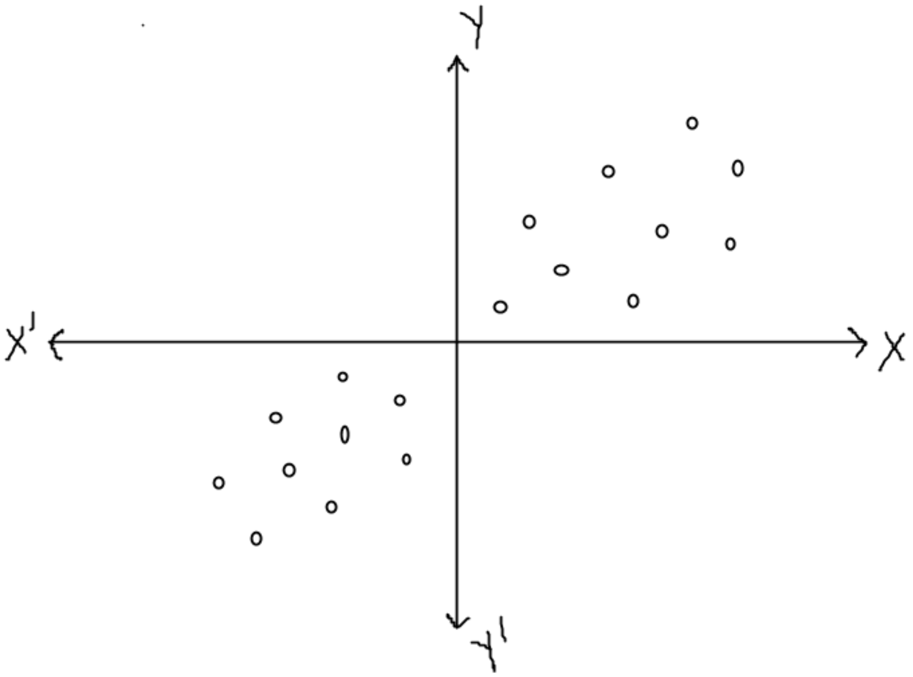
Q. Do you think there is a relationship between the two variables?

X	Y
1	2
2	4
3	6
4	8
5	10

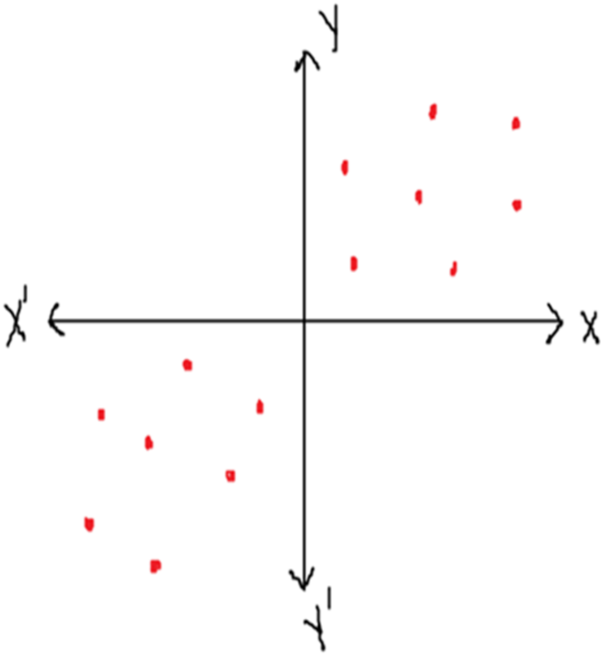
When we plot the values on X-Y plane, it forms a straight line.

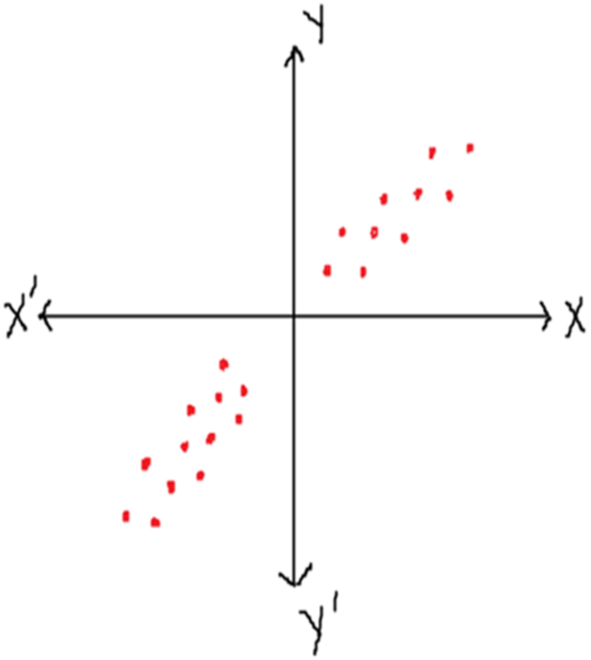


Q. Do you think there is a relationship between the two variables?

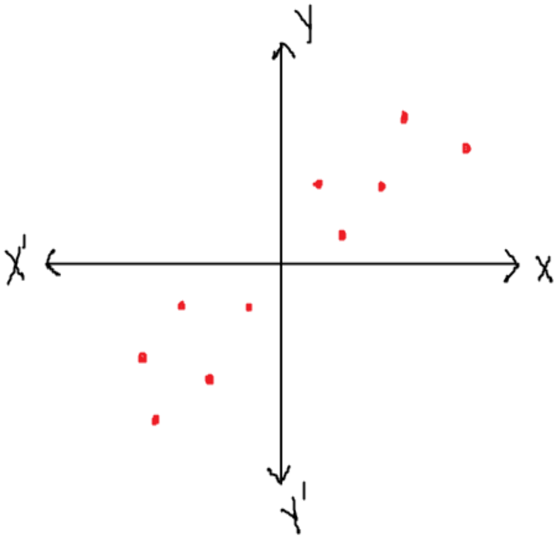


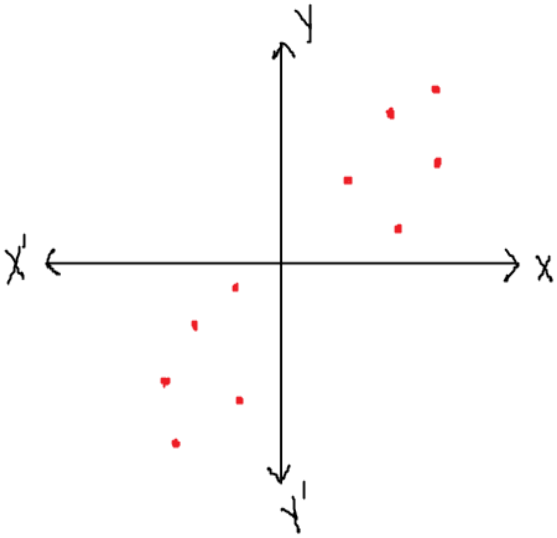
Q. Which of the two diagrams depict a stronger relationship between the two variables?





Q. Which of the two diagrams depict a stronger relationship between the two variables?





With the help of scatter plot, we may not be able to depict always if the variables have a relationship or not and the extent to which they are related to each other.

So, we use correlation coefficient. It tells the extent to which two variables fluctuate together. It is denoted by *r*.

$$r = \frac{Cov(x, y)}{\sigma_x \cdot \sigma_y} = \frac{Cov_{x,y}}{\sigma_x \cdot \sigma_y}$$

$$r = \frac{1}{n-1} \left(\frac{\sum_x \sum_y (x - \bar{x})(y - \bar{y})}{\sigma_x \cdot \sigma_y} \right)$$

Q. Area of House (sq. feet)

Price of House (in million $)

Area of House	Price of House
1500	1
2000	1.5
3500	2
4000	4

Q. What do you think the direction of relationship between these variables would be?

POSITIVE

Q. Calculate and interpret the value of sample correlation coefficient.

r = 0.87

As the Area of House increases, the Price of House tends to increase and vice-versa.

To Calculate Correlation coefficient in Excel, use function

PEARSON (array1 , array2)

Q. How would the correlation coefficient change if we add a data point (5000 , 1)?

r = 0.29

Q. Would the correlation coefficient be affected if X = Price of House and Y = Area of House.

NO

Q. Would the r value be affected if we measure the price of House in Rupees rather than dollar?

NO

Q. What do you think would be the range of the correlation coefficient?

-1 to 1

A correlation coefficient of -1 represents perfect negative correlation

while that of 1 represents perfect positive correlation.

Q. What would be the correlation coefficient for the following variables?

a) **Proportion of income spent on consumption, proportion of Income saved.**

-1

b) **Number of matches won in a series, Number of matches lost in the series by a particular team.**

-1

c) **Number of people who go to watch movie, Amount of money spent on tickets**

1

d) **Price of Milk, Price of Pens**

0

Q. No. of Years of Education

Salary in dollars ($)

No. of Years of Education	Salary ($)
12	10000
12	80000
15	20000
17	40000
20	1000000
15	15000

Q. Calculate and interpret the correlation coefficient.

Q. Do you think correlation implies causation?

Correlation does not imply causation. Causation can be established only through a scientific study. For eg : Correlation can tell that Depression and lack of sleep are highly correlated. However, we cannot say that they cause each other unless established by scientific theory as causation would imply that it would definitely cause depression which may not always be true.

Q. A study reported that the correlation coefficient for the age of a

person and time since he acquired a driving license was 0.97.

Why do you think the value of r is so close to 1?

A high positive correlation coefficient implies the higher the age of the person, the greater the time since he/she has acquired a driving license. Since most of the people attain their driving license early, so as age increases, the time since they acquired license would be greater. Therefore, r is close to 1.

Q.

Age	No. of Runs Scored
21	30
27	80
20	120
29	5
33	15
22	10
32	150

Q. Calculate and interpret the correlation coefficient.

0.38

Hypothesis Testing

Whatever be the sample correlation coefficient, it might be possible that the population correlation coefficient is significantly different from it due to sampling error. To decide whether it is significantly different or not, we perform hypothesis testing.

ρ (pronounced as rho) represents the population correlation coefficient while r represents the sample correlation coefficient.

$H_o : \rho = 0$

$H_a : \rho \neq 0$

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

Q. Would you reject or retain the null hypothesis that runs scored are not dependent on age for the above example at a significance level of 5%?

t value = ?

= 0.086894

t critical = ?

t* = 2.57

Reject

Retain

Use TINV for calculating t critical value in Excel.

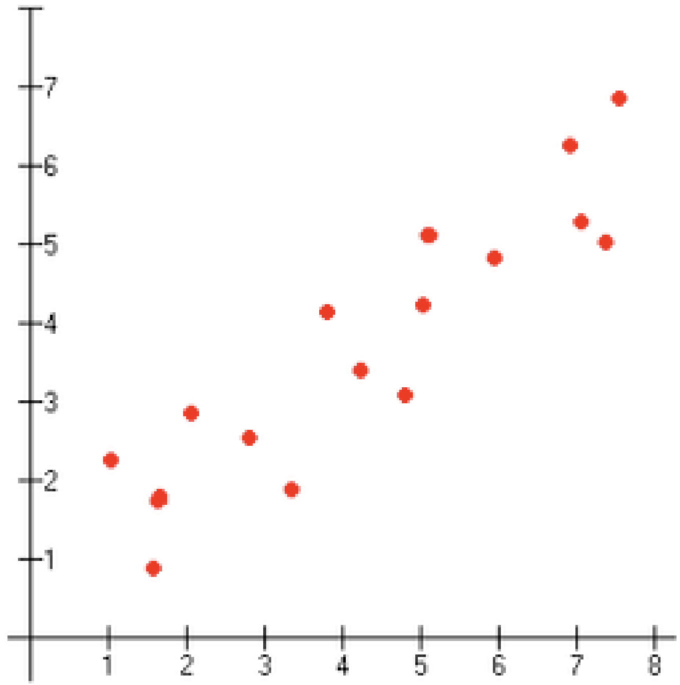
Use TDIST for calculating P-value.

Q. We want to see if there is a relationship between Husband's age and Wife's age.

Husband's Age	Wife's Age
35	34
27	23
49	47
66	63
21	22
32	25
26	26

Q. What is the correlation coefficient?

0.985



Q. What will be the null and alternate hypothesis?

$H_o : \rho = 0$

$H_a : \rho \neq 0$

Q. Would you reject or retain the null hypothesis for a significance level of 99%?

$t \ = 12.92348$

$t^* = 3.365$

Reject null

Do not reject null

Q. Would your decision change if we add (60 , 30) to our data?

t value = 3.26

t critical = 3.7

Reject null

Do not reject null