Online-Class 17-3-2021

Probability, Statistics and Reliability (MAT3003)

SLOT: B21 + B22 + B23

MODULE - 3

Topic: Correlation Coefficient

Contents

Correlation Coefficient (or Karl Pearson Correlation Coefficient)

Definition: Correlation coefficient is a statistical measure of the strength of the relationship between the relative movements of two variables. The values range between -1.0 and 1.0.

Remarks:

- **Correlation coefficients** are used in statistics to measure how strong a relationship is between two variables.
- There are several types of correlation coefficient, but the most popular is Karl Pearson's.
- Karl **Pearson's correlation coefficient** commonly used in linear relationship between two sets of data. In fact, when anyone refers to **the** correlation coefficient, they are usually talking about Karl Pearson's.

Correlation Coefficient Formula:

Formula 1:

The correlation coefficient (r) is given by

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

where n = number of data points,

Reference: Text Book: 1.Probability and Statistics for Engineers & Scientists by R.E.Walpole, Pearson (2012).

$Formula\ 2\ (Using\ Covariance)$:

Correlation coefficient r is given by:

$$r = \frac{Cov\ (X, Y)}{\sigma_X \sigma_Y}$$

where $Cov(X, Y)$ – covariance between the variables X and Y.

σ_X – standard deviation of the X-variable.

σ_Y – standard deviation of the Y-variable.

Formula for Covariance

$Cov(X, Y)$ for population: $\text{Cov}(X, Y) = \dfrac{\sum(X_i - \bar{X})(Y_j - \bar{Y})}{n}$

$Cov(X, Y)$ for sample: $\text{Cov}(X, Y) = \dfrac{\sum(X_i - \bar{X})(Y_j - \bar{Y})}{n - 1}$

where X_i – values of the X-variable,

Y_j – values of the Y-variable,

$\bar{X}$ – mean (or average) of the X-variable,

$\bar{Y}$ – mean (or average) of the Y-variable,

n – number of data points.

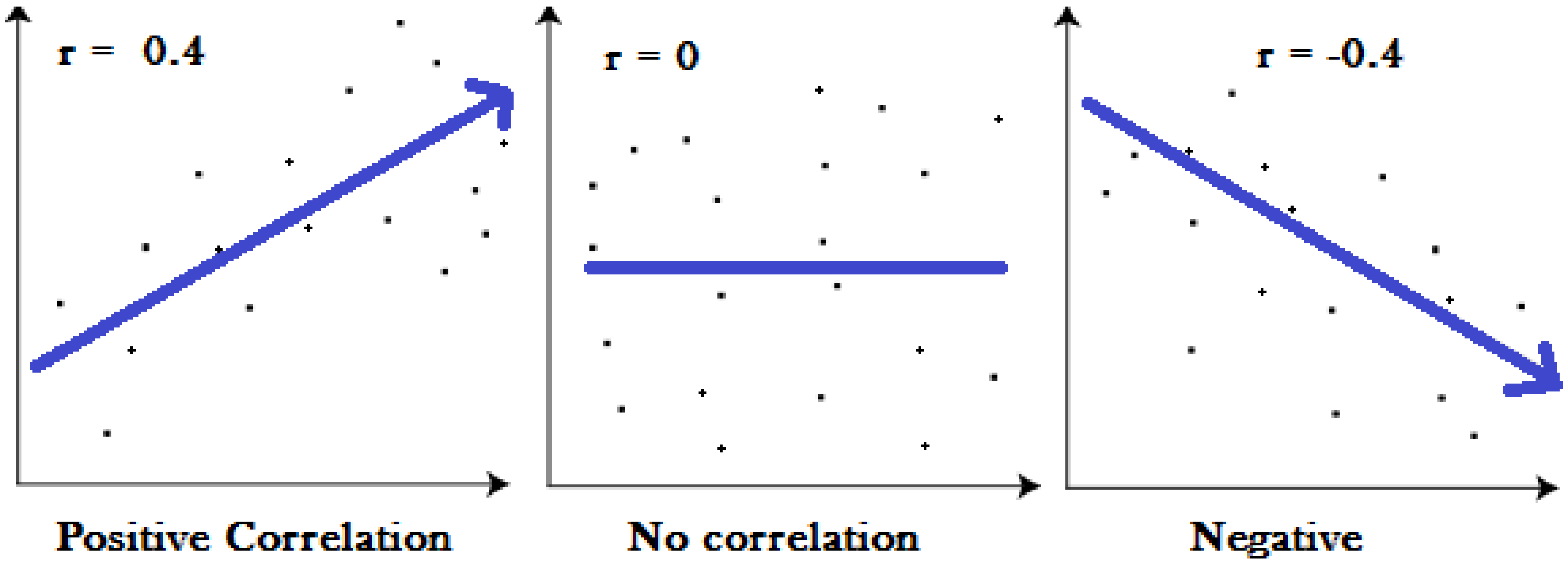
$$\text{Cov}(X, Y) = E\{[X - E(X)]\,[Y - E(Y)]\}$$

Reference: Probability and Statistics for Engineers & Scientists by R.E.Walpole, Pearson (2012).

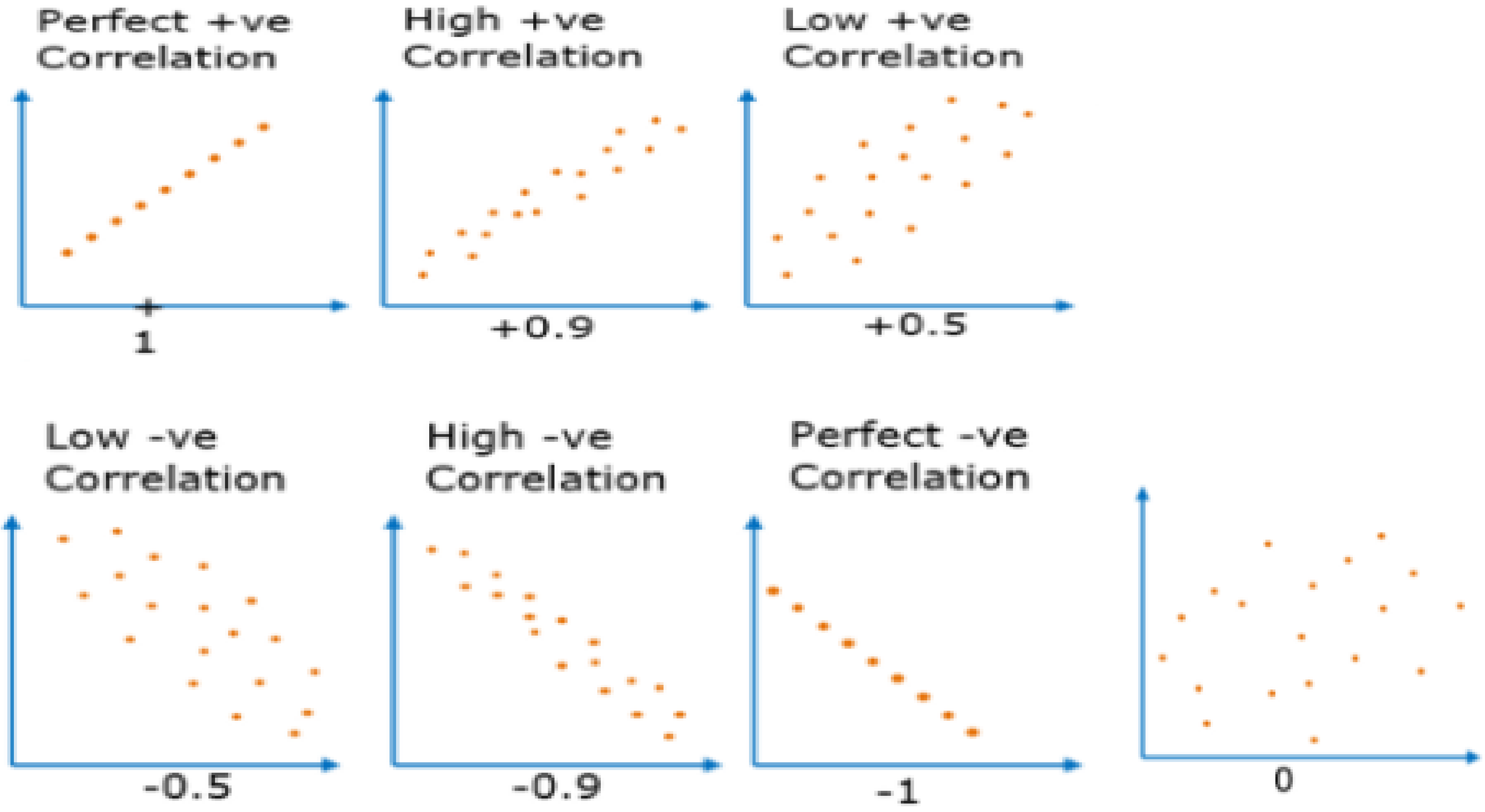
Properties of r

1. $-1 \leq r \leq 1$

2. r>0: Positive Correlation, r=0: No Correlation, r<0 : Negative Correlation



Reference: Text Book: 1.Probability and Statistics for Engineers & Scientists by R.E.Walpole, Pearson (2012).



Reference: Probability and Statistics for Engineers & Scientists
by R.E.Walpole, Pearson (2012).

3. r= +1 => Perfect Positive Correlation.

4. r= -1 => Perfect Negative Correlation.

5. r is near to + 1 => Strong Positive Correlation.

6. r is near to - 1 => Strong Negative Correlation.

7. r is Positive and close to zero => Weak Positive Correlation.

8. r is Negative and close to zero => Weak Negative Correlation.

Real Life Examples

- Shoe sizes go up in (almost) correlation with foot length (Positive Correlation).
- The amount of gas in a tank decreases in (almost) perfect correlation with speed (Negative Correlation).
- Their is no relationship between the amount of tea drunk and level of intelligence. This is done by drawing a scatter diagram (Zero Correlation or No Correlation).

Positive Correlations - Common Examples

- As attendance at school drops, so does achievement.
- When enrollment at college decreases, the number of teachers decreases.
- As a student's study time increases, so does his test average.
- As the temperature goes up, ice cream sales also go up.
- When an employee works more hours his paycheck increases proportionately.

Positive Correlations - Common Examples

- The more it rains, the more sales for umbrellas go up.
- As a person's level of happiness decreases, so does his level of helpfulness.
- People who suffer from depression have higher rates of suicide than those who do not.
- If any product is on demand, then its price also increases.

Positive Correlations - Common Examples

- As the number of trees cut down increases, the probability of AIR POLUTION increases.
- As the temperature decreases, the speed at which molecules move decreases.
- As the speed of a wind turbine increases, the amount of electricity that is generated, increases.
- As the amount of moisture increases in an environment, the growth of mold spores increases.

Some Applications of Correlations

Prediction

- If there is a relationship between two variables, we can make predictions about one from another.

Validity

- Concurrent validity (correlation between a new measure and an established measure).

Reliability

- Test-retest reliability (are measures consistent).
- Inter-rater reliability (are observers consistent).

Theory verification

- Predictive validity.

Theorem

- Two independent RV’s X and Y are uncorrelated, but two uncorrelated RV’s need not be independent.

Proof

When X and Y are independent, $E(XY) = E(X) \cdot E(Y)$.

$\therefore \qquad C_{XY} = 0$ and hence $r_{XY} = 0$

viz., X and Y are uncorrelated.

The converse is not true, since $E(XY) = E(X) \cdot E(Y)$, when $r_{XY} = 0$.
This does not imply that X and Y are independent, as X and Y are independent only when $f(x, y) = f_X(x) \cdot f_Y(y)$.

Question 1

Find the value of the correlation coefficient from the following table:

SUBJECT	AGE X	GLUCOSE LEVEL Y
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81

Solution

Step 1: *Make a chart.* Use the given data, and add three more columns: xy, x^2, and y^2.

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X^2	Y^2
1	43	99			
2	21	65			
3	25	79			
4	42	75			
5	57	87			
6	59	81			

Step 2: *Multiply x and y together to fill the xy column. For example, row 1 would be 43 × 99 = **4,257**.*

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X^2	Y^2
1	43	99	4257		
2	21	65	1365		
3	25	79	1975		
4	42	75	3150		
5	57	87	4959		
6	59	81	4779		

Step 3: *Take the square of the numbers in the x column, and put the result in the x^2 column.*

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X^2	Y^2
1	43	99	4257	1849	
2	21	65	1365	441	
3	25	79	1975	625	
4	42	75	3150	1764	
5	57	87	4959	3249	
6	59	81	4779	3481	

Step 4: *Take the square of the numbers in the y column, and put the result in the y^2 column.*

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X^2	Y^2
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561

Step 5: *Add up all of the numbers in the columns and put the result at the bottom of the column.* The Greek letter sigma (Σ) is a short way of saying "sum of."

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X^2	Y^2
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561
Σ	247	486	20485	11409	40022

From table: $\Sigma x = 247$, $\Sigma y = 486$

$\Sigma xy = 20{,}485$, $\Sigma x^2 = 11{,}409$

$\Sigma y^2 = 40{,}022$, n = sample size = 6.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}} = 2868 / 5413.27 = 0.529809$$

Question 2

Compute the coefficient of correlation between X and Y, using the following data:

X:	1	3	5	7	8	10
Y:	8	12	15	17	18	20

Solution

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	8	1	64	8
3	12	9	144	36
5	15	25	225	75
7	17	49	289	119
8	18	64	324	144
10	20	100	400	200
34	90	248	1446	352

Thus, $n = 6$

$$\Sigma x_i = 34,\ \Sigma y_i = 90$$
$$\Sigma x_i^{\,2} = 248,\ \Sigma y_i^{\,2} = 1446$$
$$\Sigma x_i y_i = 582$$

$$r_{XY} = \frac{n\Sigma xy - \Sigma x \cdot \Sigma y}{\sqrt{\{n\Sigma x^2 - (\Sigma x)^2\}\,\{n\Sigma y^2 - (\Sigma y)^2\}}}$$

$$= \frac{6 \times 582 - 34 \times 90}{\sqrt{\{6 \times 248 - (34)^2\}\,\{6 \times 1446 - (90)^2\}}}$$

$$= \frac{432}{\sqrt{332 \times 576}} = 0.9879$$

Question 3 (For Students)

A researcher wished to determine if a person's age is related to the number of hours he or she exercises per week. The data obtained from a sample is given. State your opinion based on Karl Pearson's coefficient of correlation for the data.

Age x:	18	26	32	38	52	59
Hours y:	10	5	2	3	1.5	1

Solution

	Age x	Hours y	xy	x^2	y^2
	18	10	180	324	100
	26	5	130	676	25
	32	2	64	1024	4
	38	3	114	1444	9
	52	1.5	78	2704	2.25
	59	1	59	3481	1
Total	225	22.5	625	9653	141.25

Correlation Coefficient (r) = -0.8320

Limitations of Correlations

1. Correlation is not and cannot be taken to imply causation. Even if there is a very strong association between two variables we cannot assume that one causes the other.

For example suppose we found a positive correlation between watching violence on T.V. and violent behavior in adolescence. It could be that the cause of both these is a third (extraneous) variable - say for example, growing up in a violent home - and that both the watching of T.V. and the violent behavior are the outcome of this.

2. Correlation does not allow us to go beyond the data that is given.

For example suppose it was found that there was an association between time spent on homework (1/2 hour to 3 hours) and number of G.C.S.E. (General Certificate of Secondary Education) passes (1 to 6). It would not be legitimate to infer from this that spending 6 hours on homework would be likely to generate 12 G.C.S.E. passes.

Practice Questions

Compute and interpret the correlation coefficient for the following grades of 6 students selected at random.

Mathematics grade:	70	92	80	74	65	83
English grade:	74	84	63	87	78	90

Find the coefficient of correlation between X and Y using the following data:

X:	5	10	15	20	25
Y:	16	19	23	26	30

Ans. 0.9907

Ten students got the following marks in Mathematics and Basic Engineering:

Marks in Mathematics	78	36	98	25	75	82	90	62	65	39
Marks in Basic Engg.	84	51	91	60	68	62	86	58	53	47

Calculate the coefficient of correlation.

THANK YOU