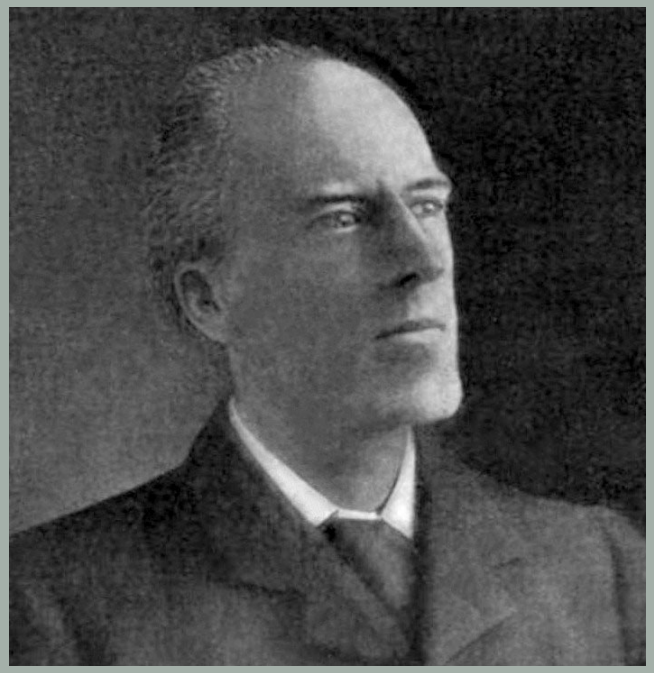
ICCPP-STATISTICS

- Pearson Product Moment Correlation

Vishal Lohchab

Scientific Assistant of
Prof. Dr. Hans-Werner Gessmann
Director ICCPP International



Carl Pearson (1857-1936)

Pearson Product Moment Correlation Coefficient

Pearson Product Moment Correlation

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

The Pearson Product Moment Correlation Coefficient is a parametric measurement

- **Pearson Product Moment Correlation (PPMC)** shows the linear relationship between two sets of data.
- Pearson's correlation coefficient calculates the effect of change in one variable when the other variable changes.

Assumptions

1.

Two or more continuous variables
(i.e., interval or ratio level)

Assumptions

2.

Cases must have non-missing values on both variables

Assumptions

3.

Linear relationship between the variables

Assumptions

4.

Independent cases
(i.e. independence of observations)

Assumptions

This means

No relationship between the values of variables of cases.

- the values for all variables across cases are unrelated
- for any case, the value for any variable cannot influence the value of any variable for other cases
- no case can influence another case on any variable

Assumptions

5.

Bivariate normality

- Each pair of variables is bivariate normally distributed
- This assumption ensures that the variables are linearly related; violations of this asumption may indicate that non-linear relationships among variables exist.

Assumptions

6.

Random sample of data from the population

Assumptions

7.

No outliers

Regression

Regression is a statistical method for calculating the line of best fit through a scatter plot of data points. The regression line uses the "independent variables" to predict the outcome or "dependent variable".

The dependent variable represents the output or response. The independent variables represent inputs or predictors, or they are variables that are tested to see if they predict the outcome.

Linear Regression

Linear regression is a linear approach to modeling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables).

Difference linear and non-linear

Linear Equations	Non-Linear Equations
It forms a straight line or represents the equation for the straight line	It does not form a straight line but forms a curve.
It **has only one degree**. Or we can also define it as an equation having the maximum degree 1.	A nonlinear equation **has the degree as 2 or more than 2**, but not less than 2.
All these equations form a straight line in XY plane. These lines can be extended to any direction but in a straight form.	It forms a curve and if we increase the value of the degree, the curvature of the graph increases.

Correlation Coefficient Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

Correlation Coefficient Formula

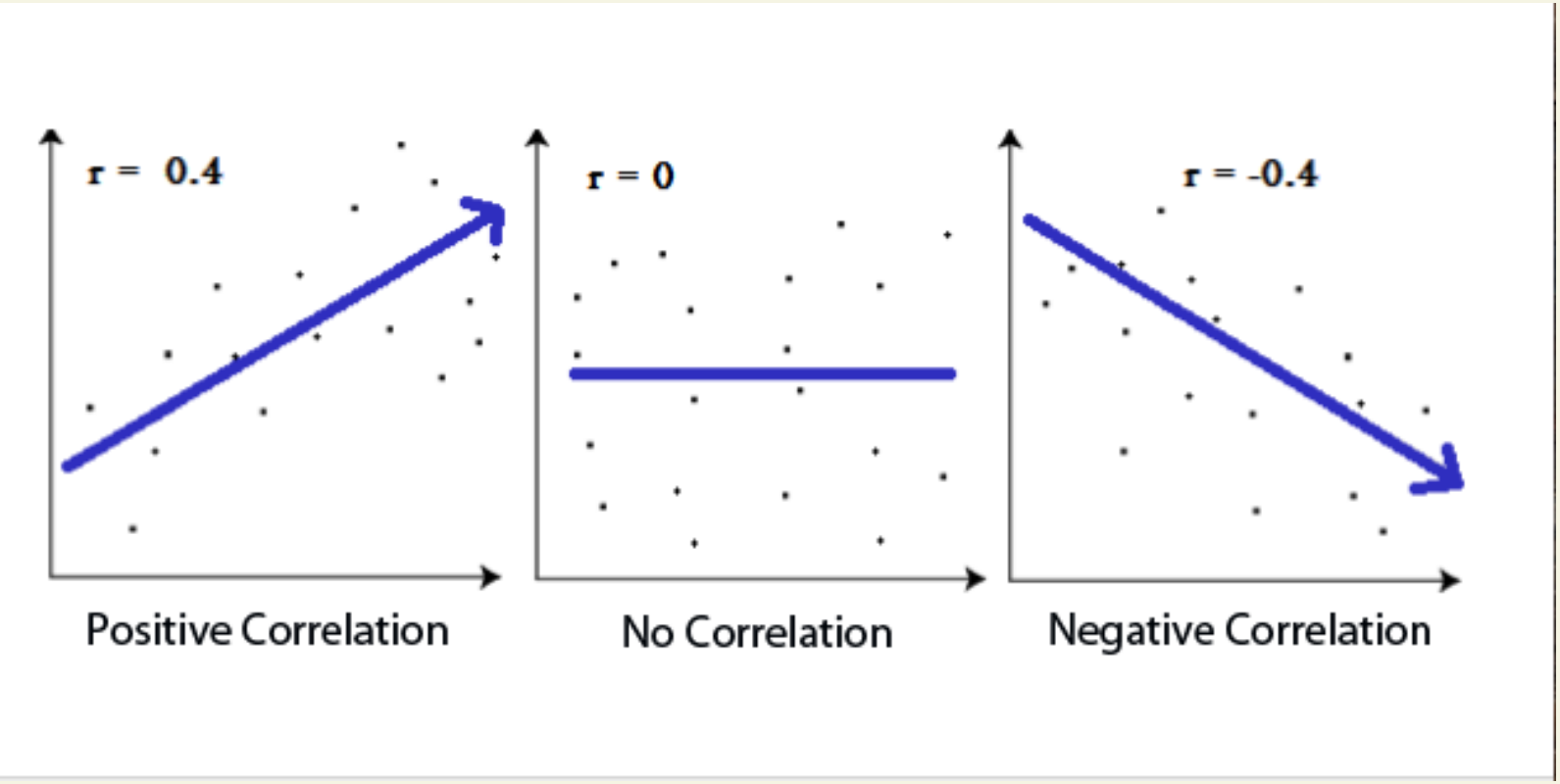
Correlation coefficient formulas are used to find how strong the relationship is between data. The formulas return a value between +1 and -1, where

+1 indicates a maximum strong positive relationship

-1 indicates a maximum strong negative relationship

0 indicates no relationship at all

Graphs showing a correlation of +0.4, 0 and -0.4



Meaning

- A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other.

Example for a positive correlation:

Shoe sizes go up in (almost) perfect correlation with foot length.

Meaning

- A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other.

  Example for a negative correlation:
  Amount of gas in a tank decreases in (almost) perfect correlation with speed.

- Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.

Solution Step Wise

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			

Solution Step Wise

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		

Solution Step Wise

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	

Solution Step Wise

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

Solution Step Wise

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" or summation.

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

Solution Step Wise

Step 6 Insert the values into the formula.

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

Solution Step Wise

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

2868 / 5413.27 = 0.529809

From our table

$\Sigma x = 247$
$\Sigma y = 486$
$\Sigma xy = 20{,}485$
$\Sigma x^2 = 11{,}409$
$\Sigma y^2 = 40{,}022$
n is the sample size, in our case = 6

Correlation coefficient

The Correlation coefficient

$$r = \frac{(20{,}485) - (247 \times 486)}{\sqrt{[6(11{,}409) - (247^2)] \times [6(40{,}022) - 486^2]}} = 0.5298$$

The range of the correlation coefficient is from -1 to 1. Our result is 0.5298 which means the variables have a moderate positive correlation.

How to convert correlation coefficient into percentage

$$r = 0.5298$$

$$0.5298 * 0.5298 = 0.2806$$

$$0.28 * 100 \Rightarrow 28\%$$

References

[1] Acton F S (1966) Analysis of Straight-Line Data. New York: Dover.

[2] Edwards A L (1976) "The Correlation Coefficient." Ch. 4 in An Introduction to Linear Regression and Correlation. San Francisco, CA: W. H. Freeman, pp. 33-46.

[3] Gonick L, Smith W (1993) "Regression." Ch. 11 in The Cartoon Guide to Statistics. New York: Harper Perennial, pp. 187-210.

[4] Knill O (2011) Lecture 12: Correlation. Retrieved April 16, 2021 from: http://people.math.harvard.edu/~knill/teaching/math19b_2011/handouts/lecture12.pdf

[5] Glen S "Welcome to Statistics How To!" From StatisticsHowTo.com: Elementary Statistics for the rest of us! https://www.statisticshowto.com/