Correlation Coefficient, Assumptions of Correlation Coefficient

The correlation coefficient is a statistical measure that calculates the strength of the relationship between the relative movements of the two variables. The range of values for the correlation coefficient bounded by 1.0 on an absolute value basis or between -1.0 to 1.0. If the correlation coefficient is greater than 1.0 or less than -1.0, the correlation measurement is incorrect. A correlation of -1.0 shows a perfect negative correlation, while a correlation of 1.0 shows a perfect positive correlation. A correlation of 0.0 shows zero or no relationship between the movements of the two variables.

While the correlation coefficient measures a degree of relation between two variables, it only measures the linear relationship between the variables. The correlation coefficient cannot capture nonlinear relationships between two variables.

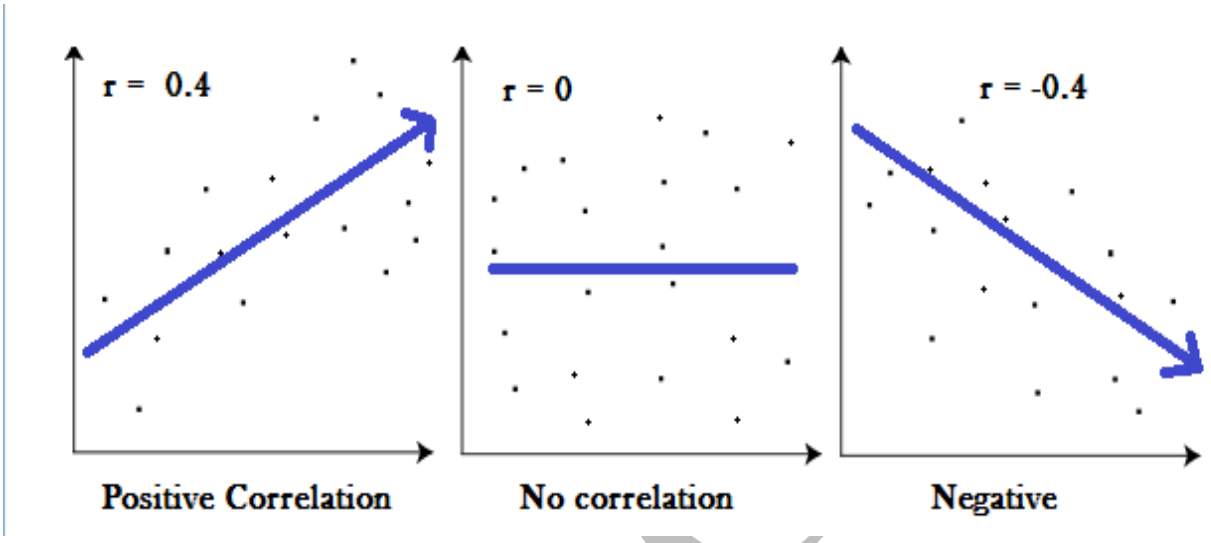
Correlation coefficients are used in statistics to measure how strong a relationship is between two variables. There are several types of correlation coefficient: Pearson's correlation (also called Pearson's *R*) is a correlation coefficient commonly used in linear regression. If you're starting out in statistics, you'll probably learn about Pearson's *R* first. In fact, when anyone refers to **the** correlation coefficient, they are usually talking about Pearson's.

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

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.

The strength of the relationship varies in degree based on the value of the correlation coefficient. For example, a value of 0.2 shows there is a positive relationship between the two variables, but it is weak and likely insignificant. Experts do not consider correlations significant until the value surpasses at least 0.8. However, a correlation coefficient with an absolute value of 0.9 or greater would represent a very strong relationship.

This statistic is useful in finance. For example, it can be helpful in determining how well a mutual fund performs relative to its benchmark index, or another fund or asset class. By adding a low or negatively correlated mutual fund to an existing portfolio, the investor gains diversification benefits.



- A **positive correlation** is a relationship between two variables in which both variables either increase or decrease at the same time. An example would be height and weight. Taller people tend to be heavier.
- A **negative correlation** is a relationship between two variables in which an increase in one variable is associated with a decrease in the other. An example would be height above sea level and temperature. As you climb the mountain (increase in height) it gets colder (decrease in temperature).
- A **zero correlation** exists when there is no relationship between two variables. For example there is no relationship between the amount of tea drunk and level of intelligence.

A correlation can be expressed visually. This is done by drawing a scatter gram - that is one can plot the figures for one variable against the figures for the other on a graph.

Correlation Coefficient Formulas

One of the most commonly used formulas in stats is Pearson's correlation coefficient formula. If you're taking a basic stats class, this is the one you'll probably use:

Where,

r = Pearson correlation coefficient

x = Values in first set of data

y = Values in second set of data

n = Total number of values.

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

Sample question:

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

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

Step 6: Use the correlation coefficient formula.

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

The correlation coefficient =

$6(20,485) – (247 \times 486) / [\sqrt{[[6(11,409) – (247^2)] \times [6(40,022) – 486^2]]}]$

$= 0.5298$

Assumptions of Correlation Coefficient are:

1. **Normality** means that the data sets to be correlated should approximate the normal distribution. In such normally distributed data, most data points tend to hover close to the mean.

2. **Homoscedascity** comes from the Greek prefix hom, along with the Greek word skedastikos, which means 'able to disperse'. Homoscedascity means 'equal variances'. It means that the size of the error term is the same for all values of the independent variable. If the error term, or the variance, is smaller for a particular range of values of independent variable and larger for another range of values, then there is a violation of homoscedascity. It is quite easy to check for homoscedascity visually, by looking at a scatter plot. If the points lie equally on both sides of the line of best fit, then the data is homoscedastic.

3. **Linearity** simply means that the data follows a linear relationship. Again, this can be examined by looking at a scatter plot. If the data points have a straight line (and not a curve) relationship, then the data satisfies the linearity assumption.

4. **Continuous variables** are those that can take any value within an interval. Ratio variables are also continuous variables. To compute Karl Pearson's Coefficient of Correlation, both data sets must contain continuous variables. If even one of the data sets is ordinal, then Spearman's Coefficient of Rank Correlation would be a more appropriate measure.

5. **Paired observations** mean that every data point must be in pairs. That is, for every observation of the independent variable, there must be a corresponding observation of the dependent variable. We cannot compute correlation coefficient if one data set has 12 observations and the other has 10 observations.

6. **No outliers** must be present in the data. While statistically there's no harm if the data contains outliers, they can significantly skew the correlation coefficient and make it inaccurate. When does a data point become an outlier? In general, a data point thats beyond +3.29 or -3.29 standard deviations away, it is considered to be an outlier. Outliers are easy to spot visually from the scatter plot.

Uses 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.

Strengths of Correlations

1. Correlation allows the researcher to investigate naturally occurring variables that maybe unethical or impractical to test experimentally. For example, it would be unethical to conduct an experiment on whether smoking causes lung cancer.

2. Correlation allows the researcher to clearly and easily see if there is a relationship between variables. This can then be displayed in a graphical form.

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. 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.