**CORRELATION VERSUS COVARIANCE**

Before, proceeding directly to difference between the two concepts correlation and covariance.

*It is Important to understand the concept of correlation and covariance separately.*

**Note – I have already made a separate note on covariance and variance (in detail), so kindly go and check out that first**.

**Here,**

**1. I will provide glimpse on covariance topic and**

**2. detail analysis of correlation**

**a. what is correlation**

**b. different types of correlation**

**c. difference between the types of correlation**

**d. methods to determine correlation.**

**(Including, diagrammatic and formula explanation)**

**3. difference between covariance and correlation.**

*Let us start with correlation topic.*

**CORRELATION**

"correlation" may indicate any type of association; in statistics it normally refers to the degree to which a pair of variables are *linearly* related.

**It is interesting to note that informally, we can synonymously use correlation term as dependence.**

**Now, you might be wondering why?**

As correlation means it is a statistical term describing the degree to which two variables move in coordination with one another.

If the two variables move in the **same direction**, then those variables are said to have a **positive correlation**.

If they move in **opposite directions**, then they have a **negative correlation**.

**Now, if x and y variables do not have any relation or any dependency among each other, then we do not have any specified relationship between x and y.**

Correlation talks about relationship; this means telling us about the **strength and direction** of the x and y variable**.**

**Strength** *means how much related the x and y variables and* **direction** *talks about position in which x and y variables move.*

**Important to Note:**

ASSUMPTION : Now, you might have assumed that correlation talks about the cause-and-effect relationship between x and y variable.

Yes, your assumption is wrong.

In statistics, Correlation studies and measures the direction and extent of relationship among variables, so the correlation measures co-variation, not causation.

Therefore, we should never interpret correlation as implying cause and effect relation. This means we cannot tell by seeing the relation data or diagram that what is a cause and what effect on x and y variable caused.

For example, there exists a correlation between two variables X and Y, which means the value of one variable is found to change in one direction, the value of the other variable is found to change either in the same direction (i.e., positive change) or in the opposite direction (i.e., negative change).

Furthermore, if the correlation exists, it is linear, i.e., we can represent the relative movement of the two variables by drawing a straight line on graph paper.

The correlation coefficient, r, is a summary measure that describes the extent of the statistical relationship between two interval or ratio level variables.

The correlation coefficient is scaled so that it is always between -1 and +1.

When **r is close to 0** this means that there is **little relationship between the variables** and the **farther away from 0 r is, in either the positive or negative direction**, the greater the relationship between the two variables.

The two variables are often given the symbols X and Y. In order to illustrate how the two variables are related, the values of X and Y are pictured by drawing the scatter diagram, graphing combinations of the two variables.

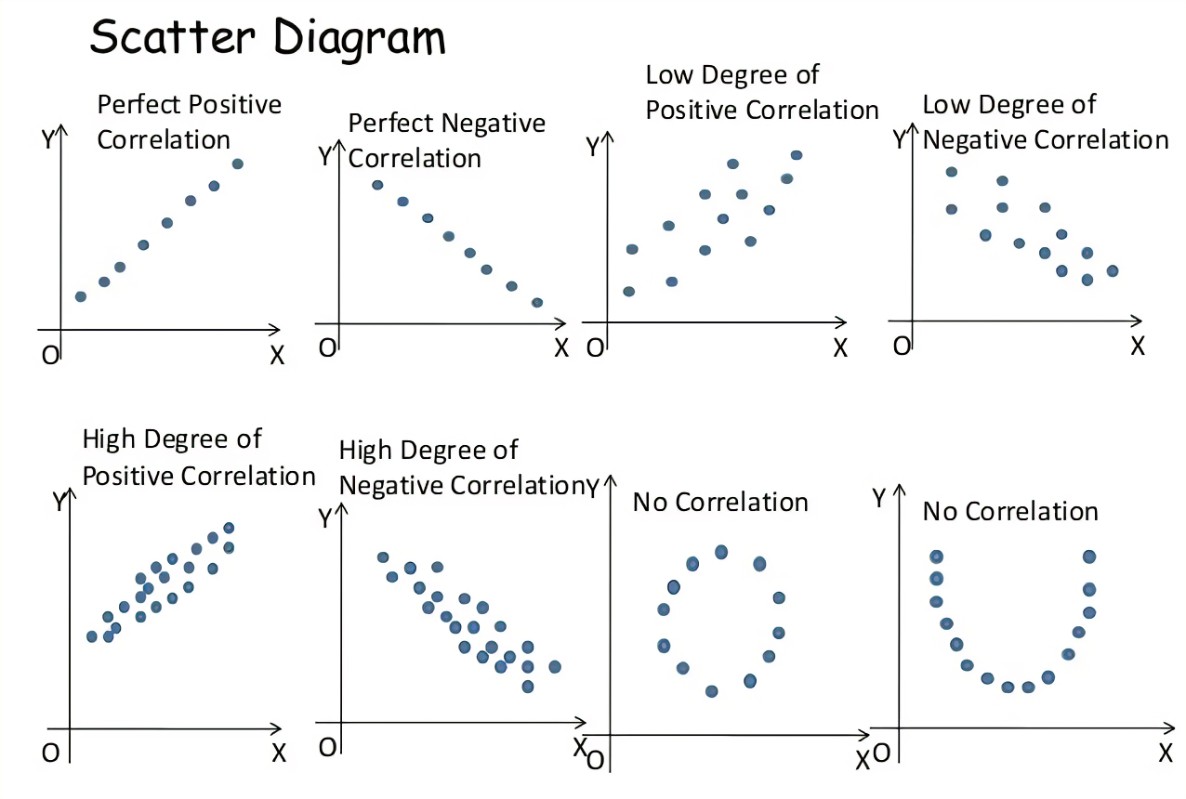
**SCATTER DIAGRAM**

A scatter diagram is a diagram that shows the values of two variables X and Y, along with the way in which these two variables relate to each other.

The values of variable X are given along the horizontal axis, with the values of the variable Y given on the vertical axis.

Later, when the regression model is used, one of the variables is defined as an independent variable, and the other is defined as a dependent variable. In regression, the independent variable X is considered to have some effect or influence on the dependent variable Y.

Correlation methods are symmetric with respect to the two variables, with no indication of causation or direction of influence being part of the statistical consideration.





**Types of Correlation**

The scatter plot explains the correlation between the two attributes or variables. It represents how closely the two variables are connected. There can be three such situations to see the relation between the two variables –

* **Positive Correlation** – when the values of the two variables move in the same direction so that an increase/decrease in the value of one variable is followed by an increase/decrease in the value of the other variable.
* **Negative Correlation** – when the values of the two variables move in the opposite direction so that an increase/decrease in the value of one variable is followed by decrease/increase in the value of the other variable.
* **No Correlation** – when there is no linear dependence or no relation between the two variables.

**METHODS OF DETERMINING CORRELATION**

1. **PEARSON’S CORRELATION 2. SPEARMAN CORRELATION**

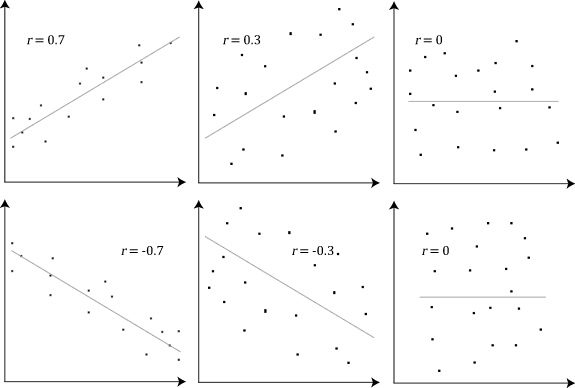
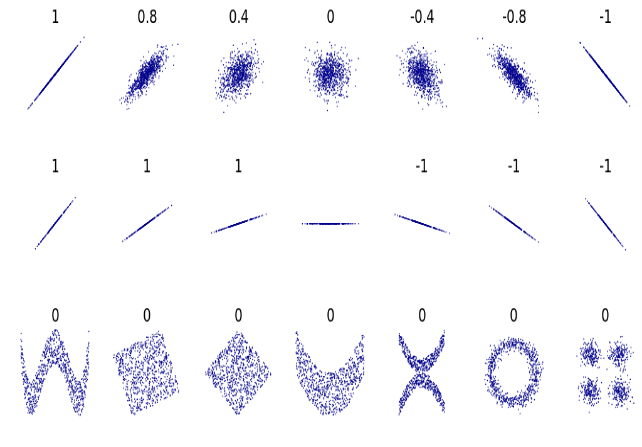
The most common formula is the Pearson Correlation coefficient used for **linear dependency** between the data sets**. The value of the coefficient lies between -1 to +1 (it is the scale).**

**When the coefficient comes down to zero, then the data is considered as not related. While, if we get the value of +1, then the data are positively correlated, and -1 has a negative correlation.**

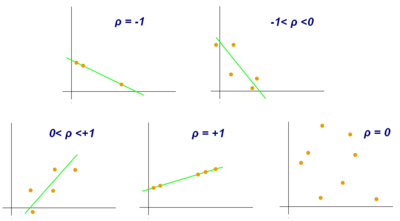
**"Pearson's correlation coefficient", commonly called simply "the correlation coefficient".**

*It is obtained by taking the ratio of the covariance of the two variables in question of our numerical dataset, normalized to the square root of their variances.*

**A Pearson product-moment correlation coefficient attempts to establish *a line of best fit* through a dataset of two variables by essentially laying out the expected values and the resulting Pearson's correlation coefficient indicates how far away the actual dataset is from the expected values.** Depending on the sign of our Pearson's correlation coefficient, we can end up with either a negative or positive correlation if there is any sort of relationship between the variables of our data set.



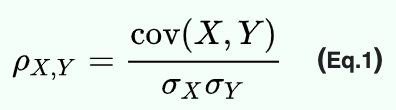
|  |
| --- |
| Pearson correlation coefficient of *x* and *y* for each set. The correlation reflects the noisiness and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the figure in the centre has a slope of 0 but in that case the correlation coefficient is undefined because the variance of *Y* is zero. |



The above diagrams, represents the relationship between two variables and telling the strength and direction in all the diagrams.

* 1. If correlation (r) is +1 then it means the positive direction and all the data are exactly on that straight line called as perfect correlation.
  2. If correlation (r) is -1 then the data represents perfect negative relationship and the best fit lines has all the data on it.
  3. If correlation (r) is between -1<r<0, it means that that data is around that downward falling line and not all the points failing exactly on the line.
  4. If correlation (r) is 0<r<1 then it means the data is around the best fit positive sloping line and not all the points failing exactly on the respective line.

**FORMULA TO CALCULATE PEARSON’S CORRELATION:**



QUESTION RELATED TO ABOVE FORMULA:

1. Have you ever wondered why the formula of correlation includes division of standard division?
2. Why not covariance is enough to tells about the relation?

This standard deviation of x and y variables let us know about the strength that how much the variables are correlated.

Whereas, covariance only talks about the direction and nothing about strength. So, including standard deviation, the correlations give the clear picture about the x any y variables relationship.

**SPEARMAN CORRELATION OR RANK CORRELATION COEFFECIENT**

In statistics, **Spearman's rank correlation coefficient** or **Spearman's *ρ***, named after Charles Spearman and often denoted by the Greek letter ***ρ*** (rho).

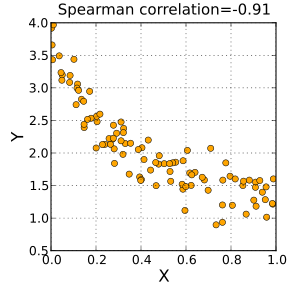
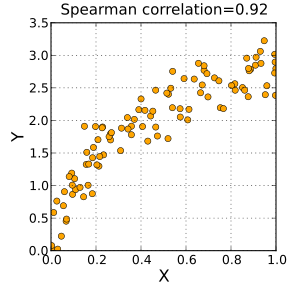
It **assesses** how well the **relationship** between two variables can be described using a **monotonic function.**

*Spearman's correlation assesses monotonic relationships (whether linear or not). If there are no repeated data values, a perfect Spearman correlation of +1 or −1 occurs when each of the variables is a perfect monotone function of the other.*

*Spearman correlation has also the same scale as of Pearson’s correlation (-1<r<1).*

**Intuitively, the Spearman correlation between two variables will be high** when observations have a **similar** (or identical for a correlation of 1) **rank** (i.e., relative position label of the observations within the variable: 1st, 2nd, 3rd, etc.) between the two variables, and **low** when observations have a **dissimilar** (or fully opposed for a correlation of −1) **rank** between the two variables.

Spearman's coefficient is appropriate for both continuous and discrete ordinal variables.

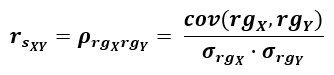
  
 positive spearman correlation negative spearman correlation

**Notice and interpretation:**

1. The graph is non-linear
2. The sign of the Spearman correlation indicates the direction of association between *X* (the independent variable) and *Y* (the dependent variable).
3. If *Y* tends to increase when *X* increases, the Spearman correlation coefficient is positive.
4. If *Y* tends to decrease when *X* increases, the Spearman correlation coefficient is negative.
5. A Spearman correlation of zero indicates that there is no tendency for *Y* to either increase or decrease when *X* increases.
6. The Spearman correlation increases in magnitude as *X* and *Y* become closer to being perfectly monotone functions of each other.
7. When *X* and *Y* are perfectly monotonically related, the Spearman correlation coefficient becomes 1.

The only difference in formula is that spearman calculate relationship on the basis of ranks.

Formula:



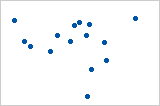
The above rgx and rgy are ranks of x and y respectively.

**How ranks are given:**

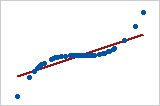
1. The lowest one data in the whole data series gets the rank as one.
2. The highest one data in the data series gets the rank as highest (where the total n count should stop there).
3. The counting goes in ascending order.
4. If somehow, any number is repeated then take the average of the two and give that same average answer to both that numbers.

**Difference between Pearson’s and Spearman’s correlation coefficient.**

1. The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables; while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not).
2. One more difference is that Pearson works with raw data values of the variables whereas Spearman works with rank-ordered variables.
3. Pearson = −0.093, Spearman = −0.093



1. Pearson = +0.851, Spearman = +1 (This is a monotonically increasing relationship, thus Spearman is exactly 1)



DIFFERENCE BETWEEN COVARIANCE AND CORRELATION

The (**sample**) **covariance** between two samples {x1, …, xn} and {y1, …, yn} is a measure of the linear association between two variables x and y based on the corresponding samples, and is defined by the formula

[Covariance](https://www.real-statistics.com/wp-content/uploads/2012/12/covariance.png)

**Observation**:

1. The covariance is similar to the variance, except that the covariance is defined for two variables (x and y above) whereas the variance is defined for only one variable. In fact, cov(x, x) = var(x).
2. The covariance can be thought of as the sum of matches and mismatches among the pairs of data elements for x and y:
3. a match occurs when both elements in the pair are on the same side of their mean;
4. a mismatch occurs when one element in the pair is above its mean and the other is below its mean.
5. The covariance is positive when the matches outweigh the mismatches and is negative when the mismatches outweigh the matches.
6. The size of the covariance in absolute value indicates the intensity of the linear relationship between x and y: the stronger the linear relationship the larger the value of the covariance will be.

## ****Correlation Properties****

**Properties:**

1. [image1514](https://www.real-statistics.com/wp-content/uploads/2013/02/image1514.png)

If r is close to 1 then x and y are positively correlated. A **positive linear correlation** means that high values of x are associated with high values of y and low values of x are associated with low values of y.

If r is close to -1 then x and y are negatively correlated. A **negative linear correlation**means that high values of x are associated with low values of y, and low values of x are associated with high values of y.

When r is close to 0 there is little linear relationship between x and y.

2. [image1518](https://www.real-statistics.com/wp-content/uploads/2013/02/image1518.png)

**DIFFERENCES:**

1. The function of covariance is correlation.
2. The values of correlation are standardized but covariance values are not. The correlation coefficient can be obtained by dividing the covariance of the variables by the product of their standard deviation values. Standard deviation measures the variability of datasets absolutely. When it is divided by the standard deviation it falls in the range of -1 to +1, which is the range of correlation values.
3. The normalized form of covariance is correlation.
4. In the formula of covariance, the units are assumed from the product of the units of the variables. Correlation is non-dimensional. It is a measure of the relationship between the variables.
5. The covariance value is affected by the change of scale in the variables. If all the values of one variable are multiplied to a constant and all the values of the other variable are multiplied by a similar or a different constant, the covariance value changes. On doing the same, the correlation value is not affected by the change in scale of the variables.
6. size of the covariance is influenced by the scale of the data elements, and so in order to eliminate the scale factor, the correlation coefficient is used as a scale-free metric of the linear relationship.
7. Covariance only talks about the direction and tells the relationship between the two variables in this way only, whereas, correlation provides STRENGTH AND DIRECTION so that er can have the magnitude (how much) and direction of the two variables movement.

**SUMMARY:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Basis** |  | **Covariance** |  | **Correlation** |
| **Meaning** |  | Covariance indicates the extent of the variable being dependent on each other. Higher value denotes higher dependency. |  | Correlation signifies the strength of association between the variables when the other things are constant. |
| **Relationship** |  | Correlation can be gathered from covariance. |  | Correlation gives the value of covariance on a standard scale. |
| **Values** |  | Lie between -∞ and +∞ |  | Correlation has limited values in the range of -1 and +1. |
| **Scalability** |  | Affects covariance |  | Correlation isn’t affected by a change in scale. |
| **Units** |  | Covariance will have a definite unit as it is concluded from the multiplication of numbers and their units. |  | Correlation is a number without units but includes decimal values. |