International Journal of Multidisciplinary Research and Growth Evaluation
ISSN: 2582-7138
Received: 17-11-2021; Accepted: 04-12-2021
www.allmultidisciplinaryjournal.com
Volume 2; Issue 6; November-December 2021; Page No. 317-318

Significance of correlation in statistics

Sayeeda Walizada
Department of Mathematics, Education Faculty, Jawzjan University, Afghanistan

Corresponding Author: **Sayeeda Walizada**

DOI: https://doi.org/10.54660/.IJMRGE.2021.2.6.317-318

Abstract
Correlation quantifies the degree and direction to which two variables are related. Correlation does not fit a line through the data points. The sign (+, -) of the correlation coefficient indicates the direction of the association. The magnitude of the correlation coefficient indicates the strength of the association, e.g. A correlation of r = - 0.8 suggests a strong, negative association (reverse trend) between two variables, whereas a correlation of r = 0.4 suggest a weak, positive association. A correlation close to zero suggests no linear association between two continuous variables. Linear regression finds the best line that predicts dependent variable from independent variable. The decision of which variable calls dependent and which calls independent is an important matter in regression, as it'll get a different best-fit line if you swap the two. The line that best predicts independent variable from dependent variable is not the same as the line that predicts dependent variable from independent variable in spite of both those lines have the same value for R^2. Linear regression quantifies goodness of fit with R^2, if the same data put into correlation matrix the square of r degree from correlation will equal R2 degree from regression. The sign (+, -) of the regression coefficient indicates the direction of the effect of independent variable(s) into dependent variable, where the degree of the regression coefficient indicates the effect of the each independent variable into dependent variable.

Keywords: 'r' value, correlation coefficient, p-value, significance

Introduction
In statistical terms, correlation is a method of assessing a possible two-way linear association between two continuous variables (Altman, 1990). Correlation is measured by a statistic called the correlation coefficient, which represents the strength of the putative linear association between the variables in question. It is a dimensionless quantity that takes a value in the range −1 to +1 (Swinscow, 1997) [5]. A correlation coefficient of zero indicates that no linear relationship exists between two continuous variables, and a correlation coefficient of −1 or +1 indicates a perfect linear relationship. The strength of relationship can be anywhere between −1 and +1. The stronger the correlation, the closer the correlation coefficient comes to ±1. If the coefficient is a positive number, the variables are directly related (i.e., as the value of one variable goes up, the value of the other also tends to do so). If, on the other hand, the coefficient is a negative number, the variables are inversely related (i.e., as the value of one variable goes up, the value of the other tends to go down). Any other form of relationship between two continuous variables that is not linear is not correlation in statistical terms. To emphasize this point, a mathematical relationship does not necessarily mean that there is correlation. For example, consider the equation y=2×2. In statistical terms, it is inappropriate to say that there is correlation between x and y. This is so because, although there is a relationship, the relationship is not linear over this range of the specified values of x. It is possible to predict y exactly for each value of x in the given range, but correlation is neither −1 nor +1. Hence, it would be inconsistent with the definition of correlation and it cannot therefore be said that x is correlated with y.

Types of correlation coefficients
There are two main types of correlation coefficients: Pearson's product moment correlation coefficient and Spearman's rank correlation coefficient. The correct usage of correlation coefficient type depends on the types of variables being studied (Hinkle et al., 2003) [3].
The correlation coefficient, r, tells us about the strength and direction of the linear relationship between x and y. However, the reliability of the linear model also depends on how many observed data points are in the sample. We need to look at both the value of the correlation coefficient r and the sample size n, together.

We perform a hypothesis test of the "significance of the correlation coefficient" to decide whether the linear relationship in the sample data is strong enough to use to model the relationship in the population.

The sample data are used to compute r, the correlation coefficient for the sample. If we had data for the entire population, we could find the population correlation coefficient. But because we have only had sample data, we cannot calculate the population correlation coefficient. The sample correlation coefficient, r, is our estimate of the unknown population correlation coefficient.

The symbol for the population correlation coefficient is ρ, the Greek letter "rho."

ρ = population correlation coefficient (unknown)

r = sample correlation coefficient (known; calculated from sample data)

The hypothesis test lets us decide whether the value of the population correlation coefficient ρ is "close to zero" or "significantly different from zero". We decide this based on the sample correlation coefficient r and the sample size n.

If the test concludes that the correlation coefficient is significantly different from zero, we say that the correlation coefficient is "significant."

Conclusion: There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from zero.

What the conclusion means: There is a significant linear relationship between x and y. We can use the regression line to model the linear relationship between x and y in the population.

If the test concludes that the correlation coefficient is not significantly different from zero (it is close to zero), we say that correlation coefficient is "not significant."

Conclusion: "There is insufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation is not significantly different from zero."

What the conclusion means: There is not a significant linear relationship between x and y. Therefore, we CANNOT use the regression line to model a linear relationship between x and y in the population.

Note

- If r is significant and the scatter plot shows a linear trend, the line can be used to predict the value of y for values of x that are within the domain of observed x values.
- If r is not significant OR if the scatter plot does not show a linear trend, the line should not be used for prediction.
- If r is significant and if the scatter plot shows a linear trend, the line may NOT be appropriate or reliable for prediction OUTSIDE the domain of observed x values in the data.

The correlation coefficient, denoted by r, is a measure of the strength of the straight-line or linear relationship between two variables. The well-known correlation coefficient is often misused, because its linearity assumption is not tested. The correlation coefficient can – by definition, that is, theoretically – assume any value in the interval between +1 and −1, including the end values +1 or −1.

Accepted guidelines for interpreting the correlation coefficient

- 0 indicates no linear relationship.
- +1 indicates a perfect positive linear relationship – as one variable increases in its values, the other variable also increases in its values through an exact linear rule.
- −1 indicates a perfect negative linear relationship – as one variable increases in its values, the other variable decreases in its values through an exact linear rule.
- Values between 0 and 0.3 (0 and −0.3) indicate a weak positive (negative) linear relationship through a shaky linear rule.
- Values between 0.3 and 0.7 (0.3 and −0.7) indicate a moderate positive (negative) linear relationship through a fuzzy-firm linear rule.
- Values between 0.7 and 1.0 (−0.7 and −1.0) indicate a strong positive (negative) linear relationship through a firm linear rule.
- The value of r^2, called the coefficient of determination, and denoted R^2 is typically interpreted as 'the percent of variation in one variable explained by the other variable,' or 'the percent of variation shared between the two variables.' Good things to know about R^2
- Linearity Assumption: the correlation coefficient requires that the underlying relationship between the two variables under consideration is linear. If the relationship is known to be linear, or the observed pattern between the two variables appears to be linear, then the correlation coefficient provides a reliable measure of the strength of the linear relationship. If the relationship is known to be non-linear, or the observed pattern appears to be non-linear, then the correlation coefficient is not useful, or at least questionable.

Rematching Process

The length of the realized correlation coefficient closed interval is determined by the process of 'rematching'. Rematching takes the original (X, Y) paired data to create new (X, Y) 'rematched-paired' data such that all the rematched-paired data produce the strongest positive and strongest negative relationships. The correlation coefficients of the strongest positive and strongest negative relationships yield the length of the realized correlation coefficient closed interval. The rematching process is as follows:

The strongest positive relationship comes about when the highest X-value is paired with the highest Y-value; the second highest X-value is paired with the second highest Y-value, and so on until the lowest X-value is paired with the lowest Y-value.

The strongest negative relationship comes about when the highest, say, X-value is paired with the lowest Y-value; the second highest X-value is paired with the second lowest Y-value, and so on until the highest X-value is paired with the lowest Y-value.

Implication of Re-matching

The correlation coefficient is restricted by the observed shapes of the individual X- and Y-values (Ratner, 2009). The shape of the data has the following effects:

1. Regardless of the shape of either variable, symmetric or otherwise, if one variable's shape is different than the

other variable's shape, the correlation coefficient is restricted.
2. The restriction is indicated by the rematch.
3. It is not possible to obtain perfect correlation unless the variables have the same shape, symmetric or otherwise.
4. A condition that is necessary for a perfect correlation is that the shapes must be the same, but it does not guarantee a perfect correlation.

Test of Significance level

Significance levels show how likely a pattern in a data is due to chance. The most common level, used to mean something is good enough to be believed, is "0.95". This means that the finding has a 95% chance of being true which also means that the finding has a confidence degree 95% of being true. No statistical package will show you "95%" or ".95" to indicate this level. Instead, it will show you ".05," meaning that the finding has a five percent (.05) chance of not being true "error", which is the converse of a 95% chance of being true. To find the significance level, subtract the number shown from one. For example, a value of ".01" means that there is a confidence degree 99% (1-.01=.99) chance of it being true.

Correlation Coefficients

Correlation is a bivariate analysis that measures the strengths of association between two variables. In statistics, the value of the correlation coefficient varies between +1 and -1. When the value of the correlation coefficient lies around ± 1, then it is said to be a perfect degree of association between the two variables. As the correlation coefficient value goes towards 0, the relationship between the two variables will be weaker. Usually, in statistics, we measure three types of correlations: Pearson correlation, Kendall rank correlation and Spearman correlation.

Pearson r correlation

Pearson correlation is widely used in statistics to measure the degree of the relationship between linear related variables. For example, in the stock market, if we want to measure how two commodities are related to each other, Pearson correlation is used to measure the degree of relationship between the two commodities.

Kendall's Tau rank correlation

Kendall rank correlation is a non-parametric test that measures the strength of dependence between two variables. If we consider two samples, x and y, where each sample size is n, we know that the total number of pairings with x y is n (n1)/2.

Spearman rank correlation

Spearman rank correlation is a non-parametric test that is used to measure the degree of association between two variables. It was developed by Spearman, thus it is called the Spearman rank correlation. Spearman rank correlation test does not assume any assumptions about the distribution of the data and is the appropriate correlation analysis when the variables are measured on a scale that is at least ordinal.

Practical use of correlation coefficient

We can expect a positive linear relationship between maternal age in years and parity because parity cannot decrease with age, but we cannot predict the strength of this relationship. The task is one of quantifying the strength of the association. That is, we are interested in the strength of relationship between the two variables rather than direction since direction is obvious in this case. Maternal age is continuous and usually skewed while parity is ordinal and skewed. With these scales of measurement for the data, the appropriate correlation coefficient to use is Spearman's.

Conclusion

In summary, correlation coefficients are used to assess the strength and direction of the linear relationships between pairs of variables. When both variables are normally distributed use Pearson's correlation coefficient, otherwise use Spearman's correlation coefficient. Spearman's correlation coefficient is more robust to outliers than is Pearson's correlation coefficient. Correlation coefficients do not communicate information about whether one variable moves in response to another. There is no attempt to establish one variable as dependent and the other as independent. Thus, relationships identified using correlation coefficients should be interpreted for what they are: associations, not causal relationships (Clarke and Cooke, 1978) [2].

References

1. Altman DG. Practical Statistics for Medical Research. Chapman & Hall/CRC
2. Clarke GM, Cooke D. A basic course in Statistics. 3rd ed. 1978.
3. Hinkle DE, Wiersma W, Jurs SG. Applied Statistics for the Behavioral Sciences. 5th ed. Boston: Houghton Mifflin, 2003.
4. Ratner B. The correlation coefficient: Its values range between +1/−1, or do they?. J Target Meas Anal. 2009; 17:139-142.
5. Swinscow TDV. In: Statistics at square one. 9th ed. Campbell M J, editor. University of Southampton; Copyright BMJ Publishing Group, 1997.