Print

What Is Correlation?

Recall that a scatterplot is used to show the relationship between two variables. The direction of the relationship might be positive, where the values of both variables generally increase together.

Or the relationship might be negative, where the values of one variable tend to increase while the values of the other variable decrease. The relationship might be linear. Or it could be curvilinear. Or there might be no linear relationship between the two variables. A scatterplot can also be used to detect unusual patterns in the data. For example, there might be a cyclical relationship between the two variables. Or there might be an extreme outlier or an unusual observation. These patterns are very clear and are easy to detect when the data are plotted using scatterplots.

Correlation is a measure of the strength of the linear relationship between two variables. The sample correlation coefficient, r, is used to quantify the strength of the linear association between two variables. Correlation is a unit-free measure, ranging from -1 to 1. The closer the correlation is to +1, the stronger the positive linear relationship. The closer the correlation is to -1, the stronger the negative linear relationship. And the closer the correlation is to zero, the weaker the linear relationship, or association. A perfect positive correlation has a value of 1, and a perfect negative correlation has a value of -1. But, in practice, a perfect correlation doesn't occur unless one variable is derived from the other variable.

We can add shaded density ellipses to the scatterplots. A density ellipse encompasses the densest region of the points in a scatterplot and provides a graphical representation of the strength and direction of the correlation. For example, a 95% density ellipse captures approximately the densest 95% of the observations. Using a density ellipse, we can see why correlation is an inappropriate measure when the relationship between the variables is curvilinear.

The tighter, or narrower, the density ellipse, the stronger the correlation. The rounder the ellipse, the weaker the correlation. Density ellipses are particularly useful when studying the correlations of many pairs of variables. In this example, we can see that the variables OD and ID are positively correlated with Removal, and that OD and ID are positively correlated with one another. We can also see that Width is not strongly correlated with any of the other variables. We'll return to this example, which involves cleaning of metal parts, in the Simple Linear Regression lesson.

Statistical Thinking for Industrial Problem Solving

Close