**Correlation Types**

Correlation is a measure of association between two variables. The variables are not designated as dependent or independent. The two most popular correlation coefficients are: Spearman's correlation coefficient rho and Pearson's product-moment correlation coefficient.

When calculating a correlation coefficient for ordinal data, select Spearman's technique. For interval or ratio-type data, use Pearson's technique.

The value of a correlation coefficient can vary from minus one to plus one. A minus one indicates a perfect negative correlation, while a plus one indicates a perfect positive correlation. A correlation of zero means there is no relationship between the two variables. When there is a negative correlation between two variables, as the value of one variable increases, the value of the other variable decreases, and vise versa. In other words, for a negative correlation, the variables work opposite each other. When there is a positive correlation between two variables, as the value of one variable increases, the value of the other variable also increases. The variables move together.

The standard error of a correlation coefficient is used to determine the confidence intervals around a true correlation of zero. If your correlation coefficient falls outside of this range, then it is significantly different than zero. The standard error can be calculated for interval or ratio-type data (i.e., only for Pearson's product-moment correlation).

The significance (probability) of the correlation coefficient is determined from the t-statistic. The probability of the t-statistic indicates whether the observed correlation coefficient occurred by chance if the true correlation is zero. In other words, it asks if the correlation is significantly different than zero. When the t-statistic is calculated for Spearman's rank-difference correlation coefficient, there must be at least 30 cases before the t-distribution can be used to determine the probability. If there are fewer than 30 cases, you must refer to a special table to find the probability of the correlation coefficient.

**Example**

A company wanted to know if there is a significant relationship between the total number of salespeople and the total number of sales. They collect data for five months.

|  |  |
| --- | --- |
| Variable 1 | Variable 2 |
| 207 | 6907 |
| 180 | 5991 |
| 220 | 6810 |
| 205 | 6553 |
| 190 | 6190 |

--------------------------------

Correlation coefficient = .921
Standard error of the coefficient = ..068
t-test for the significance of the coefficient = 4.100
Degrees of freedom = 3
Two-tailed probability = .0263

**Another Example**

Respondents to a survey were asked to judge the quality of a product on a four-point Likert scale (excellent, good, fair, poor). They were also asked to judge the reputation of the company that made the product on a three-point scale (good, fair, poor). Is there a significant relationship between respondents perceptions of the company and their perceptions of quality of the product?

Since both variables are ordinal, Spearman's method is chosen. The first variable is the rating for the quality the product. Responses are coded as 4=excellent, 3=good, 2=fair, and 1=poor. The second variable is the perceived reputation of the company and is coded 3=good, 2=fair, and 1=poor.

|  |  |
| --- | --- |
| Variable 1 | Variable 2 |
| 4 | 3 |
| 2 | 2 |
| 1 | 2 |
| 3 | 3 |
| 4 | 3 |
| 1 | 1 |
| 2 | 1 |

-------------------------------------------

Correlation coefficient rho = .830
t-test for the significance of the coefficient = 3.332
Number of data pairs = 7

Probability must be determined from a table because of the small sample size.

**Regression**

Simple regression is used to examine the relationship between one dependent and one independent variable. After performing an analysis, the regression statistics can be used to predict the dependent variable when the independent variable is known. Regression goes beyond correlation by adding prediction capabilities.

People use regression on an intuitive level every day. In business, a well-dressed man is thought to be financially successful. A mother knows that more sugar in her children's diet results in higher energy levels. The ease of waking up in the morning often depends on how late you went to bed the night before. Quantitative regression adds precision by developing a mathematical formula that can be used for predictive purposes.

For example, a medical researcher might want to use body weight (independent variable) to predict the most appropriate dose for a new drug (dependent variable). The purpose of running the regression is to find a formula that fits the relationship between the two variables. Then you can use that formula to predict values for the dependent variable when only the independent variable is known. A doctor could prescribe the proper dose based on a person's body weight.

The regression line (known as the *least squares line*) is a plot of the expected value of the dependent variable for all values of the independent variable. Technically, it is the line that "minimizes the squared residuals". The regression line is the one that best fits the data on a scatterplot.

Using the regression equation, the dependent variable may be predicted from the independent variable. The slope of the regression line (b) is defined as the rise divided by the run. The y intercept (a) is the point on the y axis where the regression line would intercept the y axis. The slope and y intercept are incorporated into the regression equation. The intercept is usually called the constant, and the slope is referred to as the coefficient. Since the regression model is usually not a perfect predictor, there is also an error term in the equation.

In the regression equation, y is always the dependent variable and x is always the independent variable. Here are three equivalent ways to mathematically describe a linear regression model.

y = intercept + (slope x) + error

y = constant + (coefficientx) + error

y = a + bx + e

The significance of the slope of the regression line is determined from the t-statistic. It is the probability that the observed correlation coefficient occurred by chance if the true correlation is zero. Some researchers prefer to report the F-ratio instead of the t-statistic. The F-ratio is equal to the t-statistic squared.

The t-statistic for the significance of the slope is essentially a test to determine if the regression model (equation) is usable. If the slope is significantly different than zero, then we can use the regression model to predict the dependent variable for any value of the independent variable.

On the other hand, take an example where the slope is zero. It has no prediction ability because for every value of the independent variable, the prediction for the dependent variable would be the same. Knowing the value of the independent variable would not improve our ability to predict the dependent variable. Thus, if the slope is not significantly different than zero, don't use the model to make predictions.

The coefficient of determination (r-squared) is the square of the correlation coefficient. Its value may vary from zero to one. It has the advantage over the correlation coefficient in that it may be interpreted directly as the proportion of variance in the dependent variable that can be accounted for by the regression equation. For example, an r-squared value of .49 means that 49% of the variance in the dependent variable can be explained by the regression equation. The other 51% is unexplained.

The standard error of the estimate for regression measures the amount of variability in the points around the regression line. It is the standard deviation of the data points as they are distributed around the regression line. The standard error of the estimate can be used to develop confidence intervals around a prediction.

**Example**

A company wants to know if there is a significant relationship between its advertising expenditures and its sales volume. The independent variable is advertising budget and the dependent variable is sales volume. A lag time of one month will be used because sales are expected to lag behind actual advertising expenditures. Data was collected for a six month period. All figures are in thousands of dollars. Is there a significant relationship between advertising budget and sales volume?

|  |  |
| --- | --- |
| Indep. Var. | Depen. Var |
| 4.2 | 27.1 |
| 6.1 | 30.4 |
| 3.9 | 25.0 |
| 5.7 | 29.7 |
| 7.3 | 40.1 |
| 5.9 | 28.8 |

--------------------------------------------------

Model: y = 9.873 + (3.682x) + error
Standard error of the estimate = 2.637
t-test for the significance of the slope = 3.961
Degrees of freedom = 4
Two-tailed probability = .0149
r-squared = .807

You might make a statement in a report like this: A simple linear regression was performed on six months of data to determine if there was a significant relationship between advertising expenditures and sales volume. The t-statistic for the slope was significant at the .05 critical alpha level, t(4)=3.96, p=.015. Thus, we reject the null hypothesis and conclude that there was a positive significant relationship between advertising expenditures and sales volume. Furthermore, 80.7% of the variability in sales volume could be explained by advertising expenditures.