# Avoiding Common Errors When Interpreting Correlations

Along with the problem of large sample sizes and their effect on p-values, it's also important to avoid other mistakes when interpreting the correlation between variables.

The first common error is that correlation does not imply causation. That is, a strong correlation between two variables doesn't mean that a change in one variable causes a change in the other variable, or vice versa. It's possible that other reasons account for a strong correlation between two variables. First, the variables might be related, but not causally. For example, people's heights and weights might be strongly correlated, because tall individuals typically weigh more, but weight doesn't determine height. For example, eating lots of French fries won't make you taller.

Second, sample correlation coefficients can be large, because both variables are affected by other variables, such as time of year. For example, both ice cream sales and drownings are high in the summer. However, the increase in drowning deaths is obviously caused by more exposure to activities in the water, not by ice cream.

Here's another example that illustrates improperly concluding a cause-and-effect relationship. This scatter plot analyzes data from the Scholastic Aptitude Test (SAT) from 1997. It shows each state's average SAT score on the Y axis and the state's spending per public school student on the X axis. The correlation between the two variables is -0.233. Based on the scatter plot, the two variables have a negative linear relationship. One might incorrectly conclude that spending money on education reduces average student performance. Although educational spending and student performance are negatively correlated, this isn't the complete story.

It turns out that there's a stronger negative correlation between state mean SAT scores and the proportion of students who take the test (r=-0.903). This might be because many states do not require the SAT and use a competing standardized test. Perhaps in low participating states, only the high-achieving students tend to take the test. After the proportion of students taking the SAT in each state is adjusted, the correlation between educational spending and student performance becomes positive (r= 0.32).

The second common error is misinterpreting the type of relationship between variables. Pearson correlations measure only the linear association between variables. That is, variables can have a near-zero correlation, but can be strongly related in a nonlinear fashion.

For example, in the 1940s, babies' birth weights had a quadratic relationship with survival. Low birth weight babies had lower survival rates because they were premature, and high birth weight babies had low survival rates from complications during delivery. So the scatter plot would show a strong relationship between survival and birth weight, but the correlation, or linear association, was low.

In this scatter plot, a best-fit regression line is shown that is nearly horizontal. The correlation coefficient for this scatter plot is close to zero, because correlation coefficients measure the strength of the linear association between variables. However, the two variables are clearly related in some nonlinear way. A curvilinear or simply a quadratic relationship exists between the two variables. Consequently, strong relationships between variables can exist despite near-zero correlations.

Finally, let's look at the third common error: failing to recognize the influence of outliers on the correlation. Outliers highly affect correlation coefficients. These plots show how one data point can misrepresent the linear relationship between two variables, and make it seem stronger than it really is. The two sets of data are identical except for the extreme outlier in the top right corner of the plot for DataTwo. This outlier drastically changes the correlation coefficient. With the outlier, the correlation coefficient is close to 1, r=0.82. Without the outlier, it's close to 0, r=0.02. For this reason, it’s important to plot your data instead of looking at only the correlation coefficients. The plot informs you that the correlation for DataTwo is unreasonable for the data.

If you run into problems with outliers in your data, what should you do? First, try to understand why a given data point is an outlier. Verify that it's a valid measurement and not an error. If the point is valid and you can collect more data, try collecting more data to see whether a linear relationship unfolds, focusing on the area between the outlier and the group of data points. If you can, try to replicate the unusual data point by collecting data at a fixed value of X.

What should you do if you can neither verify that the data point is valid nor replicate it? In this case, you should compute two correlation coefficients, one with the outlier and one without it. Then report both correlation coefficients in order to report how influential the unusual data point is in your analysis.

Close