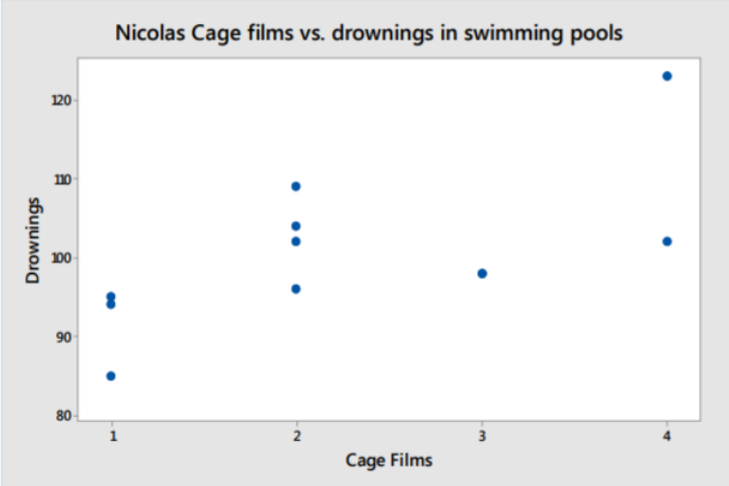
3.2.3: Correlation vs. Causation

One of the greatest mistakes people make in Statistics is in confusing correlation with causation.

Example: Nicolas Cage movies and drownings

A study done by law student Tyler Vigan showed a moderate to strong correlation between the number of movies Nicolas Cage releases in a year and the number of drownings in swimming pools in the same year.[36]

The scatterplot shows moderate positive correlation, supported by a correlation coefficient of 0.66.

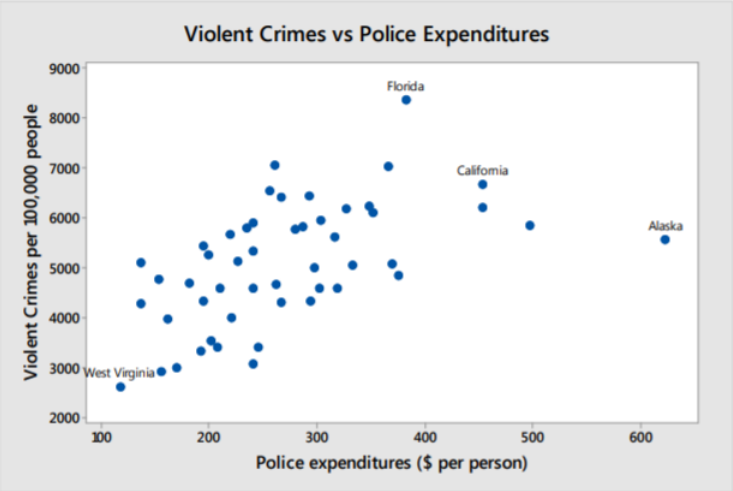


What does this mean? When Nicolas Cage releases a movie, people get excited and go jump in the pool? Or maybe in a year when there are many drownings, Nicolas Cage gets inspired to release a new movie?

This is an example of a **spurious** correlation, a correlation that just happens by chance.

Example: Crime and police expenditures

The scatterplot shows data from all 50 states adjusted for population differences. The horizontal axis is annual police expenditures per person. The vertical axis represents reported violent crimes per 100,000 people per year.



There is a moderate positive correlation present, with a correlation coefficient of 0.547.

What does this mean? Here are possible explanations.

1. **Police cost causes crime**. The more money spent on police, the more crime there is. Eliminate the police to reduce crime.
2. **Crime causes police cost**. The more crime there is, more police get hired. High crime states need to spend more money on the police.
3. **More police means more reported crimes**. The data shows reported crimes, but many crimes go unreported. Having more police means more reported crimes.
4. **Crime and police costs are higher in cities**. States like California, Texas and Florida have major cities where all expenses are higher and there is more crime. So in this example, urbanization is the cause of both variables increasing. (This is an example of a **confounding** variable).

The truth is we can't say why there is a correlation between police expenditures and violent crime. As statisticians, we can only say the variables are correlated, and we cannot support a cause and effect relationship.

In observational studies such as this, **correlation does not equal causation**.

Confounding (lurking) variables

A confounding or lurking variable is a variable that is not known to the researcher, but affects the results of the study.

Research has shown there is a strong, positive correlation between shark attacks and ice cream sales. There is actually a store in New York called Shark's Ice Cream, possibly inspired by this correlation.[37]

A possible confounding variable might be temperature. On hot days people are more likely to swim in the ocean and are also more likely to buy ice cream.

This graph from the BBC seems to support this claim. [38]Both shark attacks and ice cream sales are highest in the summer months.

In the next section, we will discuss how to design experiments that control for confounding variables.

Hopefully taking this Statistics class will help you avoid making the mistake of confusing correlation and causation. Or, maybe you already knew that, as inspired by this XKCD comic "Correlation."[39]