14.3: Correlation versus Causation

We cover a great deal of material in introductory statistics, including how to interpret statistical information. Hopefully you've seen how the statistical results from quality studies can be used in your career, and day to day life, to help you make better decisions. We now come to what may be the most important lesson in introductory statistics: the difference between variables that are related (correlated) and a variable that causes a change in another variable (causation).

It is **very, very tempting** to look at variables that are correlated (have a significant Pearson's r) and assume that this means they are *causally* related; that is, it gives the impression that what we call the IV is *causing* changes in what we call the DV.

However, in reality, Pearson's correlational analysis doe not – and cannot – do this. Correlations DO NOT show causation. No matter how logical or how obvious or how convenient it may seem, no correlational analysis can demonstrate causality. The ONLY way to demonstrate a causal relation is with a properly designed and controlled experiment that rules out all of the other things that could have affected the DV.

Many times, we have prior information that suggests that one variable causes changes in the other. Thus, when we run our Pearson's r and find a strong, statistically significant results, it is very tempting to say that we found the causal relation that we are looking for. The reason we cannot do this is that, without an experimental design that includes random assignment and control variables, is that the relation we observe between the two variables may be caused by something else that we failed to measure. These “third variables” are lurking variables or confound variables, and they are impossible to detect and control for without an experiment.

Note: TW for drowning

A common example of this is the strong, positive correlation between ice cream sales and drowning; as ice cream sales increase, so does death by drowning. Does eating ice cream cause drowning? Probably not. Does drowning cause people to have eaten ice cream? Definitely not. Could there be a third variable? Yes! There is also a strong, positive correlation between outside temperatures and both ice cream sales and drowning. When it gets hotter outside, more people buy ice cream. And when it's hotter, more people go swimming (which just, statistically, leads to more drowning). Remember this example when you start thinking that a significant correlation means that one variable *caused* the change in the other!

Confound variables, which we will represent with C, can cause two variables (X and Y) to appear related when in fact they are not. They do this by being the hidden cause of each variable independently. That is, if C causes X and Z causes Y, the X and Y will appear to be related . However, if we control for the effect of C (the method for doing this is beyond the scope of this text), then the relation between X and Y will disappear. As another example, consider shoe size and spelling ability in elementary school children. Although there should clearly be no causal relation here (why would bigger feet lead to better spelling? Or why would better spelling lead to bigger feet?), the variables are nonetheless consistently correlated. The confound in this case? Age. Older children spell better than younger children and are also bigger, so they have larger shoes.

When there is the possibility of confounding variables being the hidden cause of our observed correlation, we will often collect data on C as well and control for it in our analysis. This is good practice and a wise thing for researchers to do. Thus, it would seem that it is easy to demonstrate causation with a correlation that controls for C. However, the number of variables that could potentially cause a correlation between X and Y is functionally limitless, so it would be impossible to control for everything. That is why we use experimental designs; by randomly assigning people to groups and manipulating variables in those groups, we can balance out individual differences in any variable that may be our cause.

It is not always possible to do an experiment, however, so there are certain situations in which we will have to be satisfied with our observed relation and do the best we can to control for known confounds. However, in these situations, even if we do an excellent job of controlling for many extraneous (a statistical and research term for “outside”) variables, we must be very careful not to use causal language. That is because, even after controls, sometimes variables are related just by chance.

Sometimes, variables will end up being related simply due to random chance, and we call these correlation *spurious*. Spurious just means random, so what we are seeing is random correlations because, given enough time, enough variables, and enough data, sampling error will eventually cause some variables to be related when they should not. Sometimes, this even results in incredibly strong, but completely nonsensical, correlations. This becomes more and more of a problem as our ability to collect massive

datasets and dig through them improves, so it is very important to think critically about any significant correlation that you encounter.

The next section talks more about spurious correlations and other issues with trying to say that one variable causes changes in the other.
