12.7: Correlation versus Causation

We cover a great deal of material in introductory statistics and, as mentioned chapter 1, many of the principles underlying what we do in statistics can be used in your day to day life to help you interpret information objectively and make better decisions. We now come to what may be the most important lesson in introductory statistics: the difference between correlation and causation.

It is very, very tempting to look at variables that are correlated and assume that this means they are causally related; that is, it gives the impression that X is causing Y.

However, in reality, correlation do not – and cannot – do this. Correlations DO NOT prove 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.

Many times, we have good reason for assessing the correlation between two variables, and often that reason will be that we suspect that one causes the other. Thus, when we run our analyses and find 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, 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.

Confound variables, which we will represent with Z, can cause two variables X and Y to appear related when in fact they are not. They do this by being the hidden – or lurking – cause of each variable independently. That is, if Z causes X and Z causes Y, the X and Y will appear to be related . However, if we control for the effect of Z (the method for doing this is beyond the scope of this text), then the relation between X and Y will disappear.

A popular example for this effect is the correlation between ice cream sales and deaths by drowning. These variables are known to correlate very strongly over time. However, this does not prove that one causes the other. The lurking variable in this case is the weather – people enjoy swimming and enjoy eating ice cream more during hot weather as a way to cool off. As another example, consider shoe size and spelling ability in elementary school children. Although there should clearly be no causal relation here, the variables and 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 Z 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 Z. 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 relation you encounter.