# Correlation

When analyzing data, there are many concepts to understand and consider. One such concept is correlation. The non-mathematical definition of this term from the Concise Oxford Dictionary is;

> *Correlation – mutual relation between two or more things*

Automatically thinking that a certain indicator has an impact on another indicator is a risky assumption. There are many times when variables and indicators have a mutual relationship that it is easily proven mathematically (correlation). Even so, that does not necessarily mean that one thing had an effect on the other.

Many of the indicators studied here are highly correlated, but one cannot automatically assume that one indicator causes another. This concept (causation) is much harder to prove. The data available on www.sumn.org cannot be used to prove causation.

An example of this is the argument that heavy drinking of alcohol tends to have a high correlation with more crime. This may not be true but it is potentially true depending upon which type of crime alcohol users are committing.

There is a perfect correlation when addressing instances when people are driving while intoxicated (DWI). This is to say consuming enough alcohol to have a 0.08 blood alcohol concentration and driving a vehicle causes DWI's. Another form of crime that shows high correlation with heavy alcohol use is violent crime. It has been demonstrated that alcohol use is associated with aggressive behavior. However, aggressive behavior also occurs in the absence of alcohol consumption.

Another example is that it has been clinically proven that an individuals smoking behavior can cause them to develop lung cancer. But just by looking at our data one cannot attribute all the lung cancer deaths to smoking. These two indicators are very highly correlated, but people are susceptible to lung cancer other ways (e.g. genetically, 2nd hand smoke exposure).

Another issue to consider is what if there is a third indicator that isn't being looked at that actually better describes what is really taking place. It could be that the third indicator "causes" the first or the second indicator. Potentially this third indicator can actually be responsible for both indicators. For example, the number of storks per year nesting in small villages of a given county and the number of newborns in these villages are clearly associated – the more storks there are the more newborns per year (this example is attributed to Yule according to Neyman (1952); see also Hofer et al., 2004). Where does the association come from? A closer look reveals that the number of storks as well as the number of newborns, reflect the size of a village: a larger village has more families producing more newborns and has more roofs allowing more storks to nest.

All that being said, one thing to remember: *causation* causes *correlation*. The reverse is not necessarily true (correlation does not prove causation). There is lot more to proving causation than a simple correlation formula. This is why you must to be very careful reading data, news stories, or even medical journals that state indicator A causes indicator B. Is it possible that indicator B causes indicator A or is it even possible that there is a completely different indicator C that is responsible for one or even both indicators A and B?

SUBSTANCE USE IN MN