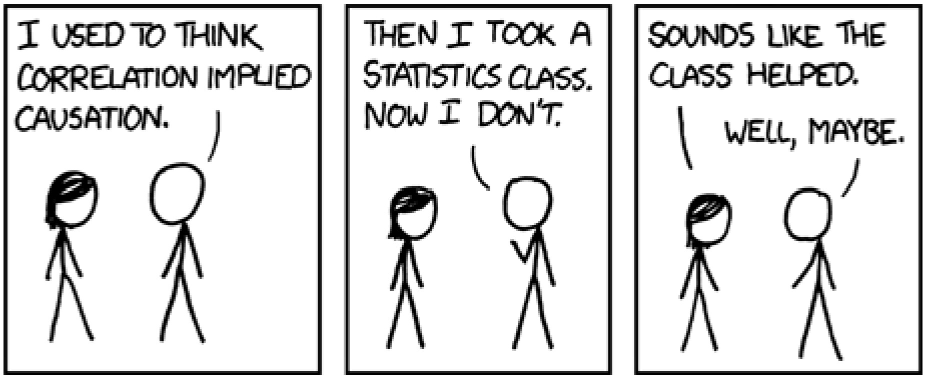
Causation versus Correlation


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

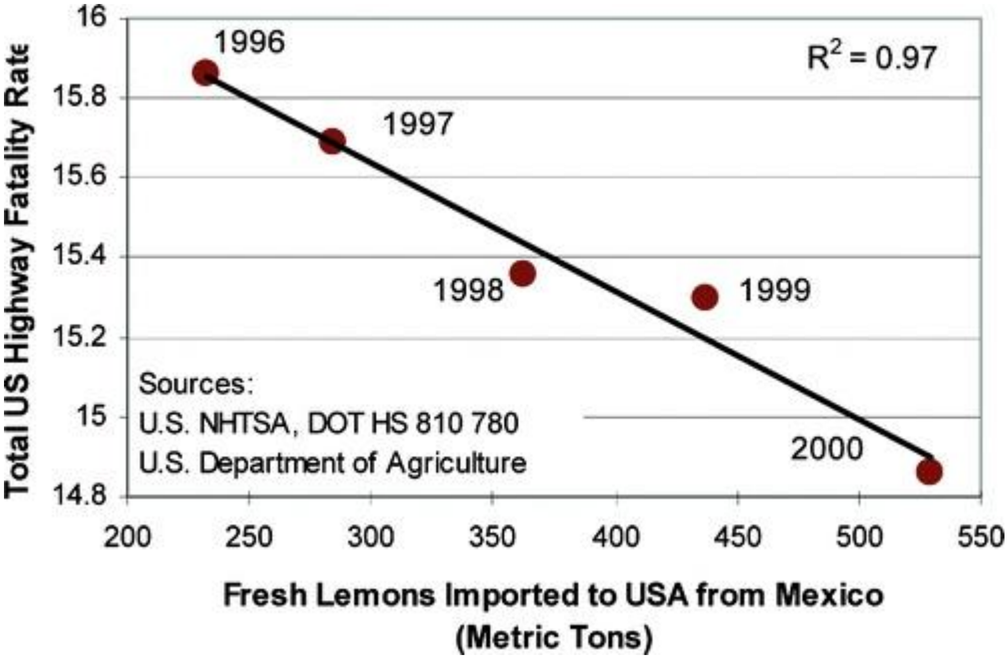
http://xkcd.com

The distinction between causation and correlation is very important in scientific thought. Oftentimes the two concepts get mixed up, sometimes out of a misunderstanding and other times due to a desire to provide a plausible explanation for a scientific observation. Therefore, it is very important to be able to understand the difference between the two ideas. In this document we will distinguish the two concepts giving examples of each. We will also describe how they can be confused and how scientists often strive to transform a correlation into a causal relationship.

Correlation is defined by Merriam-Webster's online dictionary to be: "the state or relation of being correlated; *specifically* **:** a relation existing between phenomena or things or between mathematical or statistical variables which tend to vary, be associated, or occur together in a way not expected on the basis of chance alone "

In other words, a correlation is a relationship between two or more things which change (variables) that can be described mathematically. Correlation refers to how closely two sets of information or data are related.

Correlation is often described statistically, utilizing what is known as the square of the correlation coefficient or R^2. It is not important for you to understand the mathematical background of this number but only to know that the closer R^2 is to one, the better the correlation. Let's take a look at a specific example of a correlation. If we examine the total US highway fatality rate and compare this to the metric tons of fresh lemons imported from Mexico, we observe a very strong correlation ($R^2 = 0.97$). This correlation is depicted in the graph below (source: United States Federal Government and Wikipedia entry: "Correlation does not imply causation").



What we have in this example is an excellent correlation between two variables. That is, as the amount of imported lemons increases, so do the traffic fatalities. However, it is fairly obvious just from logical thought that there is more than likely no causal relationship between the two. That is, the importing of lemons does not cause traffic fatalities. Conversely, if we stopped importing lemons, we would not expect the number of traffic fatalities to decline.

This leads us to the definition of causation. According to Merriam-Webster's online dictionary, causation is "the act or process of causing; the act or agency which produces and effect". This is often referred to as "cause and effect". That is, a causal relationship between two things or events exists if one occurs because of the other. For example, I can say that when I work fewer hours, I earn less money. There is a direct cause and effect relationship between working a certain number of hours and earning an amount of money. Likewise, when I hit the golf ball, the ball moves. Again, a direct cause-effect exists between hitting the golf ball and the ball moving.

Another important aspect to consider regarding causation is the direction of the cause and effect. Logically, for one event to cause another, it must occur first in time. In many cases, the reverse of a causal relationship is not true. Take for example the cause and effect of earning money. When I work fewer hours, I earn less money. However, if I earn less money it does not necessarily infer that I will be working fewer hours. It may be due to the fact that I got a job that pays less and therefore I have to work more hours to make the same money. Another example is diabetes and obesity. There has been an established causal relationship between obesity and diabetes. However, the reverse is not a true causal relationship. Thus, one cannot say that if I have diabetes, I will be obese.

Now sometimes, correlation and causation can be confused. In many cases, one can think that correlation implies causality. However, this is a logical fallacy. In other words, just because two events or things occur together does not imply that one is the cause of the other. Or, that without one thing or event occurring, the other would not happen. For example, I may say that when I awake at 10:00 AM, the sun is up. I can accumulate data for these events over a period of time. I would expect to find a very strong correlation. That is, every time I awake at 10:00AM, the sun will be up. I can even gather data to support the correlation in the negative; that is, when I awake at 4:00 AM, the sun is NOT up. However, if I were to try to make the relationship of my awakening at 10:00AM and the sun being up a causal one, I would be in error. That is, it is a logical fallacy to say that my awakening at 10:00 AM causes the sun to be up.

Oftentimes in science, an attempt is made at understanding by studying events, looking for correlations, and then trying to extend the correlation into a cause-effect relationship. This is often difficult to do and is fraught with potential pitfalls. Such attempts at extending a correlation into a causal relationship are quite common in the medical research arena. Let's take an example. Several epidemiological studies showed a correlation between post menopausal women who were taking hormone replacement therapy and the fact that they experienced a lowered risk of coronary heart disease (Lawlor DA, Davey Smith G, Ebrahim S (June 2004). "Commentary: the hormone replacement-coronary heart disease conundrum: is this the death of observational epidemiology?". *Int J Epidemiol* **33** (3): 464–7; and Wikipedia entry: "Correlation does not imply causation"). These studies led some to believe that hormone replacement therapy <u>caused</u> a lower risk of coronary heart disease. However, a closer examination of the data showed that a majority of the women on the trial tended to exercise regularly and also have better nutritional habits. For this reason, it was not possible to conclude which factors- the hormone therapy or lifestyle-led to the improvement in cardiac health of this patient group. It is also possible that other factors or a combination of factors were responsible for the outcome.

So how does a correlation come to be considered a causal relationship? Proving causation is a major challenge. There are no set "rules" or criteria for saying that a correlation is causation. In general, however, the more robust the correlations, the more likely they are to imply causation. An example of this is the link between smoking and cancer. Over the years, many studies have been conducted and the correlation between the incidence of cancer and smoking is strong enough that most today consider this to be a causal relationship. That is, smoking causes cancer. (Although as stated earlier, the reverse is not valid: cancer leads to smoking).