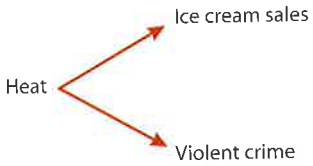
Correlation does not require causation

Science is aimed at understanding how the world works. Identifying the causes of events is a key part of the process of science. A first step in that process is the discovery of patterns, which usually involves noticing associations between events. When two events are associated, the possibility is raised that one is the cause of the other.

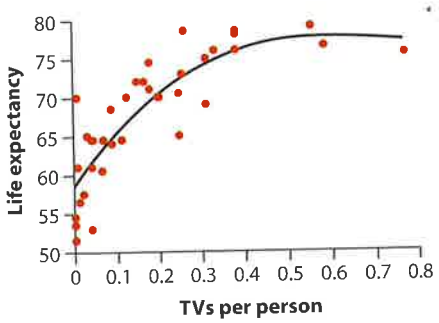
For example, our ancestors noticed that chewing on willow bark made sufferers of headaches and fevers feel better. We now know that the association between bark-chewing and pain relief is explained by the fact that salicylic acid in the bark blocks the release of prostaglandins in the body, which mediate pain and inflammation. This association led to the invention of aspirin.

An association (or correlation) between variables is a powerful clue that there may be a causal relationship between those variables. Smoking is associated with lung cancer; drinking is associated with fatal automobile accidents; taking streptomycin is associated with reduced bacterial infection. These associations exist because one thing causes the other, and the causal relationship was discovered because someone noticed their correlation.

The problem is that two variables can be correlated without one being the cause of the other. A correlation between two variables can result instead from a common cause. For example, the number of violent crimes tends to increase when ice cream sales increase. Does this mean that violence instills a deep need for ice cream? Or does it mean that squabbles over who ate the last of the Chunky Monkey® escalate to violence? Perhaps, but the more likely explanation for this association is that they share a common cause: hot weather encourages both ice cream consumption and bad behavior. Ice cream sales and violence are correlated, but this doesn't mean that one causes the other.



If we plot the mean life expectancy of the people of a country against the number of televisions per person in that country, we see quite a strong relationship (Rossman 1994):



But the magical healing powers of the TV have yet to be demonstrated. Instead, it is likely that both televisions per capita and life expectancy have a common cause in the overall wealth of the citizens of the country.

These examples demonstrate the problems posed by **confounding variables**. A confound-

ing variable is an unmeasured variable that changes in tandem with one or more of the measured variables, which gives the false appearance of a causal relationship between the measured variables. The apparent relationship between violence and ice cream sales is probably the result of the confounding variable *temperature. Overall wealth*, or something related to it, is probably the confounding variable in the correlation between the number of televisions and life expectancy.

Even more confusingly, a correlation between two variables may actually result from reverse causation. That is, the variable identified as the effect by the researcher may actually be the cause. For example, studies have repeatedly shown that babies that are breast-fed grow slightly slower than babies that are fed by formula. This has been interpreted as evidence that breast-feeding causes slow growth, but the truth turns out to be the reverse. Babies that grow rapidly are more likely to feed more, to be more demanding, and to be moved off the breast onto formula to give the poor mothers a break. Experimental studies confirm that exclusive breast-feeding has no measurable effect on infant growth (Kramer and Kakuma 2002).

These examples demonstrate the limitations of **observational studies,** which illuminate patterns but are unable to fully disentangle the effects of measured explanatory variables and unmeasured confounding variables. The main purpose of **experimental studies** is to disentangle such effects. In an experiment, the researcher is able to assign subjects randomly to different treatment groups. Random assignment breaks the association between the confounding variable and the explanatory variable, allowing the causal relationship between treatment and response variables to be assessed.

Sir Ronald Fisher, our hero in many other respects, never believed that smoking caused lung cancer; instead, he thought that smoking may be caused by a genetic predisposition and that this genetic effect might also predispose one to cancer.[1] In other words, he thought that the genotype of an individual was a confounding factor. Fisher himself invented experimental design, which would make it possible to test his claim. In theory, one could assign subjects randomly to smoking and nonsmoking treatments and any underlying correlation with genetics would be broken, because on average, just as many subjects genetically predisposed to cancer would be assigned to both treatments. If Fisher's hypothesis were correct, the smokers would not have an increased cancer rate. Such an experiment is not ethically possible with humans, but it has been done with other species.

Finding correlations and associations between variables is the first step in developing a scientific view of the world. The next step is determining whether these relationships are causal or coincidental. This requires careful experimentation.

[1] Fisher was a lifelong smoker who consulted with the Tobacco Manufacturers' Standing Committee at the time he made this claim.