

SECTION 4.3 – CORRELATION AND CAUSATION

We have previously discussed the meaning of correlation between 2 variables. The notion in the context of 2 factors is simpler. To say 2 factors are positively correlated means that when the first factor is present, the second factor occurs a larger percentage of the time than when the first factor is not present. For example, smoking and getting cancer are positively correlated because a larger percentage of smokers get cancer than nonsmokers. Being female and having heart attacks are negatively correlated because a smaller percentage of females have heart attacks than males.

On the other hand, to say smoking is a causal factor for cancer does not mean that all smokers get cancer or that only smokers get cancer. But it does mean that *if one takes up smoking, then that person is more likely to get cancer BECAUSE of smoking than he would have been if he did not take up smoking*. In other words, *if everyone smoked, more people would get cancer than if no one smoked*. To see the difference, consider the following possibilities.

Suppose some people are born with a gene that both makes them more likely to get cancer and more disposed to smoke than those born without it. Then a larger percentage of the smokers will get cancer than for the nonsmokers, i.e. the factors will be positively correlated. But smoking would not be a cause of the cancer (or cancer be a cause of smoking). If the people with the gene refrained from smoking, their chances of getting cancer would not change. In fact, the number of cancers would be the same whether everybody smoked or nobody smoked. The gene, a confounding factor, would be a cause both of smoking and of cancer.

Or suppose instead that some people who get cancer take up smoking (perhaps to relieve the anxiety). Then the 2 factors would again be positively correlated. In this case, cancer would be a cause of smoking; but smoking would not be a cause of cancer. Note that correlation is symmetric, but causation is not.

Please note that I am not claiming that there is any evidence that either of these suppositions is true. But that does not preclude the possibility of them being true.

Correlations based on observational studies can have 3 types of explanations:

- (1) variability, i. e. coincidence;
- (2) confounding factors;
- (3) causation (but which causes which?)

Since various explanations are possible, one can never deduce causation from correlation with absolute certainty. As we saw with sampling, variability decreases with sample size. So if our comparison groups are large, then the coincidence explanation is less likely. One can also eliminate a confounding factor by ensuring that the same percentage of each group has that factor. For example, if you think that smokers are more likely to get cancer, not because of smoking, but because they drink coffee, then you would include the same percentage of coffee drinkers in the smokers and nonsmokers groups. The problem is that you can never anticipate all possible confounding factors.

Nevertheless, while the high incidence of cancer among smokers does not prove causation, it does suggest causation as a plausible reason and should not be ignored. As with

any inductive argument, you cannot arrive at the conclusion with certainty. However, some arguments are stronger than others.

Secondly, you could seek a physical model. This was done in the case of Kepler's law of elliptic orbits. First he induced it from observations of existing orbits. But the law can also be deduced from Newton's laws. Keep in mind that the conclusion from a valid deduction follow with certainty from the premises. But you also need to be concerned with the truth of the premises. In this case, the premises are Newton's laws. They have also been arrived at inductively, but the evidence for them is so large that most people have a high level of confidence in them. Likewise, biologists have produced a model of cancerous mutations in lung cells from which the causation conclusion could be deduced. Of course, the strength of that argument will depend on the evidence for the model.

Thirdly, we could test our hypothesis by an experiment. The definitive experiment would be to have everybody smoke and everybody not smoke and see how many people get cancer in the two cases. But that experiment is impossible to conduct. In fact, since causation is an assertion about hypotheticals (if everybody smoked and if everybody did not smoke), we can not test it directly. But we can simulate the hypotheticals in the following way. Randomly choose 2 disjoint samples from the population of new born babies. Require the first sample (the treatment group) to start smoking at an early age. Require the second sample (the control group) to refrain from smoking. If possible, make them think they are smoking by giving them something they think are cigarettes, but do not have an effect (a placebo). Suppose that after several years, the percentage of the treatment group that has cancer is significantly larger than for the control group. What does that show? It might be a coincidence. The chances of that depend on how much larger the percentage is and on the sample sizes. By the design of our experiment we have eliminated the kind of confounding factors discussed for observational studies. Since the subjects did not choose whether or not to smoke, we can conclude that the only factor that could have caused the difference was the smoking.

In theory, the third method is very good. In practice, it has some logistical and ethical problems. Sufficiently large samples may not be available and if they are, the experiment could be prohibitively expensive. Human beings (unlike rats) have minds of their own. They may not follow your dictates, or report their actions to you honestly. Even if you have it in your power to enforce your requirements, it may not be ethical to do so. Since the time of the Tuskegee experiments there are protocols which must be observed in order not to violate the rights of the subjects. In particular, you can only choose the samples from volunteers. In some cases, that may introduce some bias into the selection of samples. Some ethicists now assert that it is not proper to even include volunteers in a control group.

In conclusion, each of these methods can produce some evidence of causation, but not absolute certainty. However, in some cases (such as smoking and cancer) the combination of these 3 methods justifies the conclusion of causation with a high level of confidence.

Exercises

1. A study found that of 100,000 people who took aspirin regularly, 10,000 had heart attacks; of 40,000 who did not, 8,000 had heart attacks. Does that mean that taking

aspirin and having heart attacks are correlated? Explain.

2. Do the statistics above prove that by taking aspirin regularly, one increases or decreases his chances of having a heart attack? Explain.
3. Design an experiment which would establish the causal relationship suggested by the statistics above.
4. Statistics show that a larger percentage of children who are spanked grow up to be maladjusted than for children who are not spanked. Give 2 plausible causal explanations for these statistics.
5. Suggest some other confounding factors which could account for the correlation between smoking and cancer.
6. A larger percentage of students who take foreign language do well in English courses than those who don't. Does that mean that if everybody took foreign language, performance in English courses would improve? Explain.
7. College graduates on average earn more money than nongraduates. Does that mean that getting a degree increases one's earning potential? Explain.
8. Studies have found correlations between racial groups and societal problems. Can you suggest confounding factors that could account for this?