Causality

Daniel Kaplan

2019-04-22

Activities

- Intervention and prediction
- Experiment and causality

Learning objectives

Additional sections

-

- Instructor orientation
- Role in statistical practice
- Classroom discussion
- Tips for an active classroom
- Student pre-requisites
- Pitfalls

Orientation for instructors

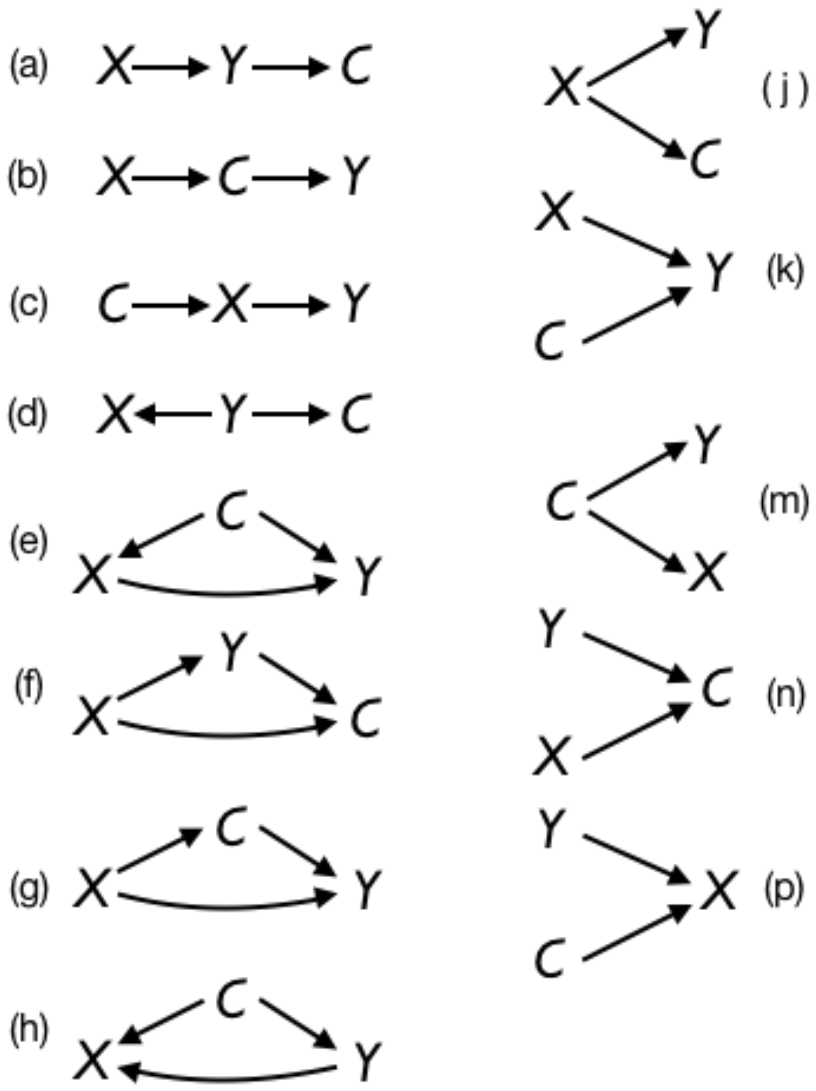
There is an epigram that is familiar to *all* statistics instructors:

> *Correlation is not causation.*
>
> Put on a more explicit logical footing, this is equivalent to:
>
> *Correlation does not imply causation.*
>
> But as familiar as this epigram is, it's wrong. We legitimately can say this:
>
> *No causation implies no correlation.*

The "no correlation" part of this is pretty easy. We have many statistical techniques for working with two variables X and Y – t-tests, regression, etc – that can establish correlation. So by "no correlation" we mean that any of these tests have failed to demonstrate a correlation.

But "no causation" is more subtle. Knowing what this means requires a bit of formalism. The formalism we will use is that of diagrams: directed acyclic graphs.

Here are several:



We can define “no causation” in terms of these graphs. Here’s a useful approximation: The procedure is to imagine putting ants into one of the nodes of the graph. And these imaginary ants can (and will!) move only along the arrows in the direction of the arrowhead. The ants can’t walk against the flow of the arrow.

By “no causation” between X and Y we mean that there is no node from which ants can get to both X and Y.

This definition works well enough so long as the variables X, Y, and C are left to their own devices, being caused according to the arrows in the diagram. But if we intervene, say holding some node at a constant value, then we might close down an otherwise legitimate along-the-arrows pathway involving X and Y, or we might even open up a pathway, like that through C in diagram (n), that ordinary wouldn’t let ants get to X and Y from C. Such a situation is called a

"collider."

It's starting to get complicated. I think this is part of the motivation for the traditional canon of introductory statistics, namely:

1. Correlation is not causation.
2. Experimental intervention is the only way to establish causation.
3. In an experiment we impose variability on X and look for correlated variability in Y.

This is a nice story, but it's incomplete. For example, not all experiments involve imposing variability on X. An experimental intervention could be holding some other variable *constant*. And there are three ways to hold a variable constant:

a. Physically pin the variable to a constant value, e.g. by regulating the temperature in an experimental chamber.
b. Select a subset of cases from the data for which the variable happens to be at a constant value.
c. In a model, include the variable to be held constant among the explanatory variables. Consequently, when using the model to generate simulated data, we can intervene to hold that variable constant. Effectively, we build a model of the world and do experiments on that model.

(a) and (b) are the most compelling, because we can't really know if the model in (c) actually represents the real-world system.

What we miss out entirely in the canonical intro stats curriculum is the kind of experiment that involves holding variables constant. Such experiments can actually be used to reject hypotheses about causal networks. If we observe a correlation when the hypothetical network (augmented, perhaps, by holding variables constant) says we shouldn't, we can reject the hypothesis. Or if we see no correlation when the network says we should see one, we can reject the hypothesis. This process of elimination, involving experiments where variables are held constant, can sometimes resolve a dispute between two rival hypotheses.

Going back to the statement that "no causation implies no correlation," we can take the contrapositive to arrive at this, more meaningful statement:

Correlation implies causation.

It's just that the correlation doesn't by itself distringuish between rival hypotheses, telling us, for example, whether X causes Y or the other way around, or whether there is a common cause shaping both X and Y. But by doing hold-a-variable-constant experiments, we can distinguish between different possible arrangements of X, Y, and C.

Role in statistical practice

Causation is an important issue in describing how systems work or deciding on what kind of intervention will be effective.

There are many situations where an impose-variation experiment is impossible for practical or ethical reasons. (We'll force this group to smoke and that group to abstain. An idea not likely to get far.) And so, it's important to be able to make reasoned judgements about causality.

Consider what happens in practice when we adhere to the "correlation is not causation" motto. At the end of some newspaper article about recent research showing the health benefits of an herbal preparation there will be a disclaimer: "This was an observational study, so no conclusion about causation can be drawn." Good. So why did the newspaper decide it's worthwhile to publish the story? And what is a reader supposed to do with the information. The reader is hardly in a position to decide on causality when the scientific experts doing the study couldn't do so.

Instead, the newspaper should list the variables that were "held constant," or "adjusted for," or "controlled for." And they should point out whether there are likely variables that play an important role in the system that have not been adjusted for.

Learning objectives

1. Understand that an observed correlation, on its own, doesn't distinguish among specific causal arrangements.
2. Understand that doing an intervention experiment (one where variability is imposed on X that is independent of all other sources of variability) can reveal a causal connection from X to Y, if a correlation is observed between X and Y in the experimental data.
3. Be able to make sense of the common practice in studies of "holding constant" or "adjusting for."

Student tasks and activities

See the activities listed at the top of this document.

Assessment items

See the activities listed at the top of this document.

Pushing the envelope/advancing the field

Just by talking about causation in a way that reflects informed professional practice in many fields, students will be able to see that what they learn in statistics is not in conflict with that practice.

Danny Kaplan, version 0.3, Wed May 29 16:03:54 2019