Twitter Thread by ■Kareem Carr■



■Kareem Carr■
@kareem_carr

Statisticians like me say CORRELATION ISN'T CAUSATION but that's not the whole story.

There are at least FOUR different scenarios!

A thread. ■

1. CORRELATED BY CHANCE. There's always a possibility that variables will correlate by chance. If you have a lot of data, you're almost certain to get a few high correlations. You will know you're in this situation if the same variables are much less correlated in new data.

2. CORRELATED DUE TO STRUCTURE. Clocks are correlated with each other but there's nothing about Clock A that can be changed in order to cause a change in Clock B or vice versa. There is no third thing you can change that will cause both clocks to change. There is no causation.

You might be tempted to say that the clocks have the common cause of being created by humans. Imagine two random stars that have a cyclical change in brightness every 24 hours. They will be correlated as well. It's not about who created them. It's about their similar structure.

3. MURKY CAUSATION. In the simplest case, if A and B are correlated and there is some causation then this could mean that A causes B, B causes A or some third thing C causes both A and B. In the most complex case, there could be complicated feedback loops between A and B.

In these cases, when we say "correlation isn't causation", what we mean is that we can't identify exactly what kind of causation there is but there is some.

4. EVEN MURKIER CAUSATION. A and B might not be related at all in the real world but something about your data collection may have caused data about A to be related to data about B. Technically, you could say you or your data collection are the cause of the correlation.

However, in the context of the original variables themselves and the real world, A is not causally related to B.

Hope this was educational! ■