Causality and experimentation

Data Science 101 Team

Association versus causality

- Association (roughly speaking, correlation) describes situations where phenomena occur more often together (or not together) than would be expected under independence.
- Causality (also referred to as causation or cause and effect) is the process by which one state, a cause, contributes to the production of another state, an effect.
- Association is much easier to establish, but does not imply causality! This is a point often missed by the press and the public.
- Some of the material in this section is adapted from Vanessa Didelez's article "Statistical Causality'' (https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.422.7911&rep=rep1&type=pdf).

Defining causality

- There is no single agreed-on definition of causality in the statistical and philosophical literature.
- Most of these approaches deal, more or less explicitly, with causation as the effect of an intervention in one (or more) variable(s) on some response variable. Typically, scientists are interested in causal relations because they want to intervene in some sense, to prevent diseases or to make life easier etc.

Example

- The number of storks per year nesting in small villages of a given country and the number of newborns in these villages are clearly associated.
- The more storks there are the more newborns per year.
- Obviously there is no causal relation, so where does the association come from?

Example continued

- A closer look reveals that the number of storks as well as the number of newborns reflect the size of a village: a larger village has more families producing more newborns and has more roofs allowing more storks to nest (Figure 1).

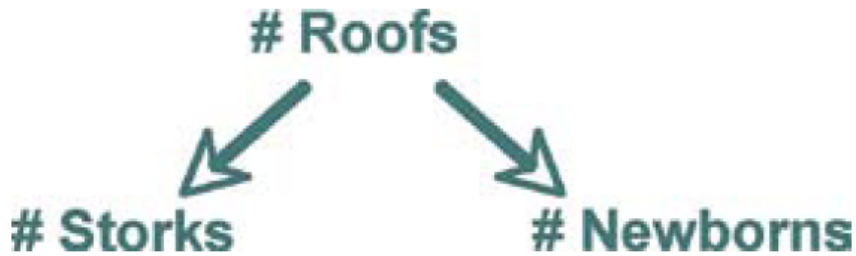


Figure 1: The number of newborns and the number of storks are associated.

Example continued

- We can be pretty sure that manipulating the number of storks in a village, e.g. setting it to zero by killing them, will not change the number of newborns in that same village – this association is not causal. We now turn to the question of why associations can be observed without an underlying causal relation.

Another example

- The bread price in Britain and the sea level in Venice over the past two centuries are positively correlated.
- Most people would agree that neither is a cause of the other, so where does the positive correlation come from?

Another example

- The bread price in Britain and the sea level in Venice over the past two centuries are positively correlated.
- Most people would agree that neither is a cause of the other, so where does the positive correlation come from?
- The explanation is that both quantities have steadily increased over time due to their respective local conditions which are not further related to each other (Figure 2). Hence it is two unrelated time trends that induce an association.

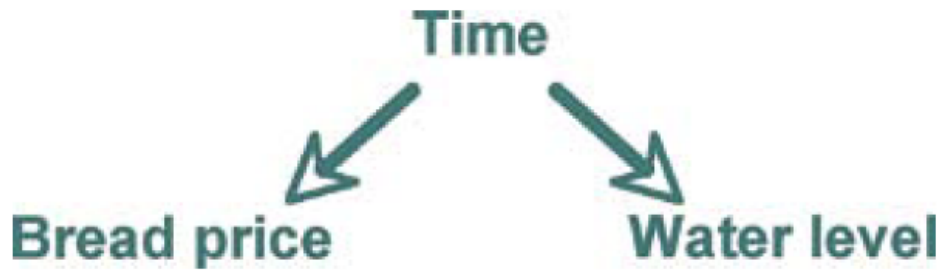


Figure 2: Bread price in Britain and water level in Venice both exhibit a time trend.

Distinguishing between Causality and Association

- A cause X and a response Y will be associated if X is indeed causal for Y but not necessarily vice versa, as demonstrated with the following examples.

Alternative explanations

- **Common Cause – Confounding**. If X and Y have a common cause, as in the storks/newborns example, they can be associated without being causally related at all. The presence of a common cause is often called confounding.
- **Reverse Causation**. In reality, Y might be the cause of X and not, as we think, vice versa. An example (?): plaque in the brain and Alzheimer's disease.
- **Time Trends**. X and Y may only be associated because they are the results of two processes with time trends without these time trends being related to each other, as for example the bread price and water level in Venice.
- **Feedback**. X and Y may be associated because they instigate each other. As an example consider alcohol abuse and social problems: does a person drink due to social problems, like problems in his job, or are such problems the consequence of alcohol abuse or both?

Observational studies and randomized experiments

Observational study: Suppose that we collect data on eating habits and health events for a group of people.

We observe that people who eat red meat have higher rates of certain cancers than people who don't eat red meat.

- This means that there is an association between red meat consumption and cancer: there is a link between these two.
- But this does not mean that eating red meat causes cancer: people who don't eat red meat are known to exercise more and drink less alcohol, and it could be the latter two issues that cause the difference in cancer rates.
- This is an observational study: It measures outcomes of interest and this can be used to establish association.
- But association is not causation, because there may be confounding factors such as exercise that are associated both with red meat consumption and cancer.

Randomized studies

If we want to investigate what happens when we manipulate a variable, then the best method is to actually carry out such manipulations and observe the result. This is what is done in experimental studies.

Design of a randomized study

A treatment (e.g. eating red meat) is assigned to people in the treatment group but not to people in the control group. Then the outcomes in the two groups are compared. To rule out confounders, both groups should be similar, apart from the treatment. To this end:

- The subjects are assigned into treatment and control groups at random.
- When possible, subjects in the control group get a placebo: it resembles the treatment but is neutral. Assigning a placebo makes sure that both groups are equally affected by the placebo effect: the idea of being treated may have an effect by itself.
- Ideally, the experiment should be double-blind: i.e., neither the subjects nor the evaluators know the assignments to treatment and control.

What problems do you see in carrying out such a study?

Challenges in carrying out a randomized study, especially in medicine

- You can't give people a treatment unless you are sure that it is safe.
- You can't give people a treatment that you know is inferior to another available treatment.

The placebo effect

- The placebo effect is a very interesting phenomenon.
- The placebo effect is a beneficial effect produced by a placebo drug or treatment, which cannot be attributed to the properties of the placebo itself, and must therefore be due to the patient's belief in the treatment or the study
- As a result, it is common to give the control arm of a study an inert (inactive) substance, typically a tablet like a sugar pill.

The logic of randomized controlled experiments

Randomization serves two purposes:

- It makes the treatment group similar to the control group. Therefore influences other than the treatment operate equally on both groups, apart from differences due to chance.
- It allows to measure the treatment effect, by calculating the size of chance effects when comparing the outcomes in the two groups.

When randomized experiments are not feasible

- In many subjects, in particular in epidemiology, it is impossible to carry out experiments.
- For instance if the 'cause' is smoking behaviour, alcohol consumption or education, we cannot randomly allocate subjects to different groups.
- Instead we have to make do with data on the behaviour as it is, but this will typically mean having to deal with confounding as, for example, smokers are likely to exhibit a life style that is different also in other respects from that of non-smokers (Figure 3).

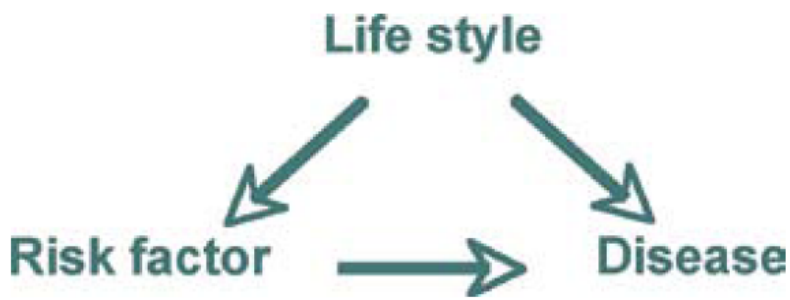


Figure 3: The problem of confounding in Epidemiology.

Inferring causality from non-experimental data

- In some circumstances when a thorough knowledge of the subject matter is available one can identify the confounders and measure them in addition to the X and Y variable of interest.
- The causal effect can then be assessed within every level of the confounders, i.e. based on stratification. This yields valid causal inference if a sufficient set of confounders is used.
- Alternatively, we can include the confounding variables on the right hand side of a regression model:

$$Y \sim \text{Treatment} + \text{Confounding variables}$$

and then use the coefficient of Treatment to estimate the Treatment effect

There are problems with these approaches

- One can never be sure about what the relevant confounders are and even then, there may be many different ways of measuring them; in addition, typical confounders are prone to errors, e.g. self-reported alcohol consumption is known to be unreliable.
- How was confounding dealt with in the cell phone study?

The role of time

- Recall that confounding is only one of the reasons why we may observe associations without the desired causation. Reverse causation, time trends and feedback all involve time but it is even more difficult to 'adjust for time' than the above 'adjustment for confounding'.
- First of all we obviously need to have data over time – usually from so-called longitudinal studies. Secondly, we must be careful not to adjust for so-called mediating (or intermediate) variables.

Smoking and lung cancer

Consider the simple example in Figure 4.

Figure 4: Example for an intermediate variable.

The effect of smoking on developing lung cancer can plausibly be assumed to be mediated by the ensuing amount of tar deposit in the lungs – note that passive smoking may also result in tar in the lungs.

The above graph even suggests that once the amount of tar is known, cancer risk and smoking are independent. If we mistakenly think of 'tar deposit' as a confounder and adjust for it, we may therefore wrongly find that there is no effect of smoking on lung cancer.

Different kinds of observational studies

- Case-control study: Find people who have a disease (e.g. lung cancer) and a matched control patient, and look back in time at their exposures (e.g. smoking).
- Retrospective Cohort: Compares groups of individuals who are alike in many ways but differ by a certain characteristic (for example, female nurses who smoke and ones who do not smoke) in terms of a particular outcome (such as lung cancer). They are sampled historically, e.g. sample nurses starting in Jan 1, 2010 and followed forward in time.
- Prospective Cohort: Compares groups of individuals who are alike in many ways but differ by a certain characteristic (for example, female nurses who smoke and ones who do not smoke) in terms of a particular outcome (such as lung cancer). They are followed forward in time.
- Case-Crossover Design - each person acts as their own matched control (e.g. the cell phone study).

Instrumental variables

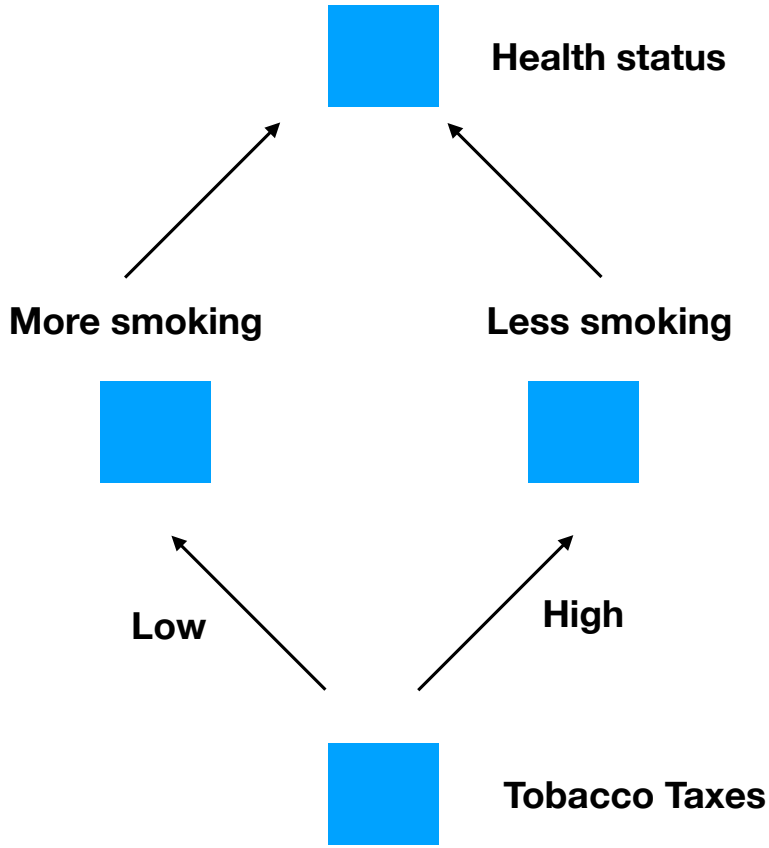
- This is a clever method for trying to infer causality in situations where carrying out a randomized experiment is not feasible. Often used by economists.
- The idea is to look for a variable - the "instrument" which has done the randomization for you, in nature, or in society.

Example

Smoking and Health:

- We can attempt to estimate the causal effect of smoking on health from observational data by using the tax rate in different states or countries for tobacco products as an instrument for smoking.
- The tax rate for tobacco products is a reasonable choice for an instrument because the researcher assumes that it can only be correlated with health through its effect on smoking.
- If the researcher then finds tobacco taxes and state of health to be correlated, this may be viewed as evidence that smoking causes changes in health.

Tobacco tax as an instrumental variable



Examples instrumental variables

- Angrist and Krueger (1991): Effect of educational attainment on weekly earnings of males born in the 1930s and 1940s. Idea: Use instrumental variables: Use birthdays as an instrumental variable (month of birth is random).
- Angrist and Lavy (1999): Causal effect of class size on scholastic achievement. Idea for instrument: Discontinuities in class size due to some institutional maximum class-size rules.

The effect of military service on future earnings

- Can use Vietnam-era draft lottery numbers as an instrument to estimate the effect of military service on earnings later in life.
- The draft lottery numbers randomly assigned to young men in the early 1970s were highly correlated with the probability of being drafted into the military, but not correlated with other factors that might change earnings later.
- Researchers found a correlation between high draft numbers and lower earnings, suggesting that military service caused a decrease in future earnings, see here.

Genotype as an instrumental variables

- Researchers use Mendelian randomization to try to pinpoint the effects of individual genotypes on pheotypes (disease states). One looks for genetic variations that are correlated with modifiable risk factors.

Mendelian randomization: an example

- Observing association between alcohol and blood pressure in an observational study may be a poor indicator of the causal effects (confounders, etc.), see here.
- In the case of alcohol and blood pressure, a variant in the ALDH2 gene slows the metabolism of acetaldehyde, which causes adverse responses to alcohol consumption.
- In a study of 4057 people selected from the general population, 170 of 1919 men carried two copies of the A allele and drank an average of 1.1 g of alcohol a day, whereas those with no copies drank 23.7 g.
- Idea: Use genetic variation as an instrument. This assumes that the gene variant has no direct effect on blood pressure.

Summary

- Establishing causality is the "holy grail'' in many areas of science.
- Causal inference is an active and important are of research in data science – especially in Statistics and Econometrics.

Searching for Clarity: A Primer on Medical Studies (Gina Kolata, NyTimes 2008)

- Link to article.
- Everyone, it seemed, from the general public to many scientists, was enthralled by the idea that beta carotene would protect against cancer. In the early 1990s, the evidence seemed compelling that this chemical, an antioxidant found in fruit and vegetables and converted by the body to vitamin A, was a key to good health.
- There were laboratory studies showing how beta carotene would work. There were animal studies confirming that it was protective against cancer. There were observational studies showing that the more fruit and vegetables people ate, the lower their cancer risk.
- So convinced were some scientists that they themselves were taking beta carotene supplements.

But then

- Then came three large, rigorous clinical trials that randomly assigned people to take beta carotene pills or a placebo. And the beta carotene hypothesis crumbled. The trials concluded that not only did beta carotene fail to protect against cancer and heart disease, but it might increase the risk of developing cancer.
- It was "the biggest disappointment of my career," said one of the study researchers, Dr. Charles Hennekens, then at Brigham and Women's Hospital.
- But Frankie Avalon, a '50s singer and actor turned supplement marketer, had another view. When the bad news was released, he appeared in an infomercial. On one side of him was a huge stack of papers. At his other side were a few lonely pages. What are you going to believe, he asked, all these studies saying beta carotene works or these saying it doesn't?

What to believe?

- The beta carotene case is unusual because much of the time when laboratory studies, animal studies and observational studies point in the same direction, clinical trials confirm these results.
- Another case: Women's Health Initiative, a huge study begun in 1991 by the National Institutes of Health. It asked, among other things, if estrogen or estrogen and progestin could protect postmenopausal women against heart disease.
- As with beta carotene, the evidence said the drugs would work. But the clinical trial showed that women who took the drugs had slightly more heart disease and an increased risk of breast cancer.

Confounding variables

- Cynthia Pearson, executive director of the National Women's Health Network, has a favorite example of how easy it is to be fooled.
- Study after study found that women taking estrogen had less heart disease than women who did not. But, Ms. Pearson says, it turns out that women who faithfully take any medication for years — even a sugar pill — are different from women who don't.
- The compliant pill-takers tend to be healthier, perhaps because they follow doctor's orders. So when scientists said they were comparing two equal populations, the estrogen users and the nonestrogen users, they may have actually been comparing the health of the sort of women who conscientiously take pills with that of the sort of women who don't or who do so less rigorously.

Wrapping up

- "The major message," Dr. Richard Klausner (then the director of the National Cancer Institute), said, "is that no matter how compelling and exciting a hypothesis is, we don't know whether it works without clinical trials."