Data Science Career Track

The Art of Statistics, Chapter 4: What causes what? Take-Away Notes

---

The chapter introduces one of the most fascinating challenges in statistics: ***inferring causation from our observations***.

The **statistical concept of causation** is as follows: a type of event X ***causes*** a type of event Y just if Y-events happen more often when X-events occur than when X-events don't occur.

- *e.g:* Smoking causes cancer because cancer-events happen more often when smoking-events happen than when smoking-events don't happen.

Two consequences of this concept are:

1. We can infer causation with confidence only by performing experiments
2. The more we intervene, the more evidence for causation we accumulate.

Clinical trials try to establish causation. Proper clinical trials should be:

- **Controlled**: i.e., have an intervention group (= a group given some exposure of interest), and a control group (= a group not given that exposure).
- **Allocated properly:** i.e., the intervention and control groups are made as similar (excepting the exposure of interest) as possible by random selection.
- **Allocated rigidly:** i.e., once an individual is assigned to a group, this doesn't change
- **Blinded:**
    - Single-blinding means the individuals in the groups don't know what groups they're in

    - Double-blinding means the people monitoring those individuals don't know this either
    - Triple-blinding means the statisticians don't know this, too.
- **Equal treatment:** i.e., the groups are treated as similarly as possible (excepting the exposure of interest).
- **Everyone is measured:** i.e., effects of exposure (or not) should be followed up
- **Multiple studies:** i.e., more than one study should be carried out.
- **Systematic review:** i.e., if more than one study has been done, collate all the evidence.

A variable is a **confounder** exactly if it's associated with both a response variable and a predictor variable, and may explain some of the relationships between them.

- e.g: Age is a confounder with respect to the response variable of weight and the predictor variable of height.

The simplest method for handling confounders is to look at the apparent relationship within each level of the confounder. This is known as **adjustment/stratification**.

Simpson's paradox occurs when an apparent relationship reverses its sign when a confounding variable is taken into account.