ST 516: Foundations of Data Analytics

Scope of Inference

Review: Observational Studies vs. Randomized Experiments

Causal Inference

- Causality Definition
- Confounding
- Examples of Confounding

Why Randomized Experiments?

Review: Observational Study vs. Randomized Experiment

- Recall: Important difference in types of study design
  - Observational Study: Experimental units are collected and their membership in different groups is observed
    - Example: A sample of 100 undergraduates is collected, and asked whether or not they belong to a fraternity/sorority. The students' end-of-year GPAs are recorded. Since students are not assigned to fraternity/sorority membership, this is an observational study.
  - Randomized Experiment: Experimental units are randomly assigned to different treatment groups (treatment levels).
    - Example: A sample of 100 undergraduates is collected, and 50 of them are randomly assigned to take a course on study skills. The students' end-of-year GPAs are recorded. Since it is randomly assigned whether a student takes the study skills course, this is a randomized experiment.

Causal Inference

- Which type of study you perform has implications for the inference you can make at the end of the study
- Causal Inference: Can we say that an explanatory variable E (predictor variable) *causes* a change in outcome of interest O (response variable)?
    - Example: Suppose we are interested in whether a particular fertilizer increases tomato yield. We would like to be able answer whether using the fertilizer (E) *causes* an increase in tomato yield (O).

Causality

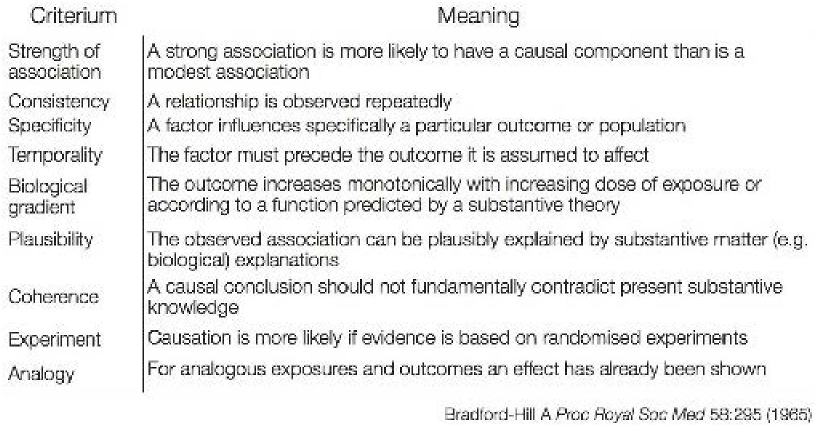
- What do we mean by ‘causal relationship’ or ‘cause-and-effect’?
    - What does it mean to say variable E causes a change in the values of variable O?

Causality

- What do we mean by 'causal relationship' or 'cause-and-effect'?
  - What does it mean to say variable E causes a change in the values of variable O?
- If we intervene and manipulate E, leaving *all* other factors the same, there should be a change in the values of O.
- Example: If we raise many potted tomato plants under the exact same conditions (light, water, temperature, etc.) but fertilize half of the plants and leave the other half unfertilized, does the tomato yield from the fertilized plants differ from the yield from the unfertilized plants?

Causality

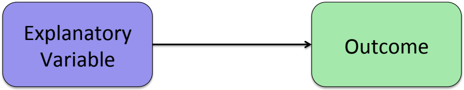
Bradford-Hill Causality Criteria:



Criterium	Meaning
Strength of association	A strong association is more likely to have a causal component than is a modest association
Consistency	A relationship is observed repeatedly
Specificity	A factor influences specifically a particular outcome or population
Temporality	The factor must precede the outcome it is assumed to affect
Biological gradient	The outcome increases monotonically with increasing dose of exposure or according to a function predicted by a substantive theory
Plausibility	The observed association can be plausibly explained by substantive matter (e.g. biological) explanations
Coherence	A causal conclusion should not fundamentally contradict present substantive knowledge
Experiment	Causation is more likely if evidence is based on randomised experiments
Analogy	For analogous exposures and outcomes an effect has already been shown

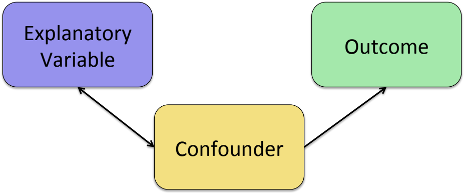
Bradford-Hill A *Proc Royal Soc Med* 58:295 (1965)

Confounding



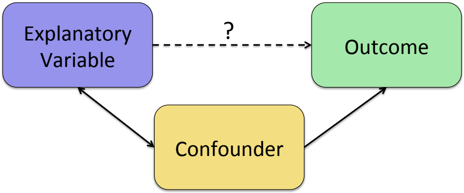
- A relationship between an explanatory variable E and an outcome of interest O

Confounding



- A relationship between an explanatory variable E and an outcome of interest O is affected by a confounder C if
  - C is related to which value of the explanatory variable E a subject has, AND
  - C affects the outcome variable value O

Confounding



- A relationship between an explanatory variable E and an outcome of interest O is affected by a confounder C if
  - C is related to which value of the explanatory variable E a subject has, AND
  - C affects the outcome variable value O
- Confounding variables (confounders) can be the reason for an observed association between E and O.

Example of Confounding

- (Explanatory Variable) Smoking
- (Outcome) Heart Disease
- (Potential Confounders) Exercise, Diet, Income

For instance, people who smoke may be less likely to exercise, and less exercise may be associated with heart disease. Therefore 'Exercise' is a confounding variable: it is related to both the explanatory variable (smoking status) and the outcome (heart disease).

It is therefore difficult to determine whether *smoking* causes heart disease, or whether the association is due to the fact that people who smoke exercise less, and exercising less causes heart disease.

Example of Confounding

- (Explanatory Variable) Alpine Lake Elevation
- (Outcome) Trout Weight
- (Potential Confounders) Water Temperature, Water Quality, Fishing Intensity

For instance, lakes at higher elevations may have lower water temperature, and lower water temperature may result in smaller trout. Therefore 'Water Temperature' is a confounding variable: it is related to both the explanatory variable (lake elevation) and the outcome (trout weight).

It is therefore difficult to determine whether *elevation* causes smaller trout, or whether the association is due to the lakes at higher elevations are colder, and colder lakes tend to have smaller trout.

Why Randomized Experiments?

- Randomized Experiments make it so that it is unlikely that any external variable affects group membership/treatment assignment.
- If we randomly assign subject to groups, it is unlikely that we have severe imbalance in any other variable between groups.
- We have therefore reduced the potential for a confounding relationship: by the randomization process, we have made it so that other variables are unlikely to be associated with our explanatory variable (group membership/treatment).

Example: Observational Study vs. Randomized Experiment

- Example: Suppose we are interested in the effect of a supplement that claims to build muscle.
  - Observational Study: Sample 50 gym members; ask whether they use the supplement; measure the change in muscle mass for each gym member over the next month.
  - Randomized Experiment: Sample 50 gym members; randomly assign one portion of the sample to use the supplement, while the other portion takes a **placebo** (this is called a **placebo-controlled trial**); measure the change in muscle mass for each gym member over the next month.
- Can you think of any potential confounding factors in the observational study? How would the randomized experiment help eliminate the effect of the confounders?

Example: Observational Study vs. Randomized Experiment

- Example (continued):
  - One potential confounding factor in the observational study is whether or not the gym member is a body-builder (that is, do they tend to lift weights more than they do aerobic exercise?).
    1. Body-builders would be more likely to use a supplement that is supposed to help increase muscle mass, AND
    2. Body-builders perform more muscle-building exercises, so are more likely to experience an increase in muscle mass over the month.
  - Thus being a body-builder would be associated both with the explanatory variable (using the supplement) and the outcome of interest (change in muscle mass), and would therefore be a confounder.

Example: Observational Study vs. Randomized Experiment

- Example (continued):
    - The randomized experiment would help eliminate this effect by randomly assigning the gym members to the supplement.
    - It is unlikely that by chance all of the body-builders would be assigned to receive the supplement, while all of the non-body-builders would be assigned to receive placebo.
    - Therefore, the randomization breaks the association between supplement use (the explanatory variable) and being a body-builder (the confounder in the observational study).