ST 516: Foundations of Data Analytics

Scope of Inference

Why Randomized Experiments?

- Causal Inference
- Examples

Population Inference

- When is Population Inference Possible?

Causal and Population Inference: Summary

Recall: 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 confounding variable between groups.

When is Causal Inference Possible?

Statistical inferences of cause-and-effect relationships can be drawn from randomized experiments, but not from observational studies.

Or, in other (possibly familiar) words:

Correlation does not imply causation.

Examples: Reversal of Observational Studies

- The potential for substantial confounding in observational studies means that we should not try to infer causality.
- There are numerous examples of cases where a relationship observed in an observational study is found *not* to be causal in later randomized experiments.
- The following two slides present two well-known examples of cases where an association between some explanatory variable and an outcome of interest that was found in observational studies does not hold in randomized experiments.
  - This means that the association between the explanatory variable and the outcome was not *causal*. Instead, it was likely due to some confounding factor.

Beta Carotene trial (CARET)

- Observational Studies:
  - Higher dietary or serum levels of beta carotene or vitamin A *associated with* **lower** rates of lung cancer and overall mortality
- Randomized Trial (CARET):
  - Beta-carotene supplementation group had 28 percent **higher** incidence of lung cancer than placebo group, and 9 percent **higher** mortality.
- Possible confounders in the observational studies?
  - Maybe people with higher dietary beta carotene eat a healthier diet in general, and that is responsible for the lower lung cancer and mortality rates?
  - Maybe there is some synergistic or independent factor found in foods with high beta carotene that reduces cancer and mortality rates?

Women's Health Initiative (WHI)

- Observational Studies:
  - Hormone therapy (HT) for post-menopausal women had been **associated**, in observational epidemiological studies, **with a reduction** in cardiovascular disease
- Randomized Trial (WHI):
  - Study was stopped early due to **increased risk** of coronary heart disease, stroke, pulmonary embolism, and venous thromboembolism in patients receiving hormone therapy.
- Possible confounders in the observational studies?
  - Maybe women who self-selected to receive hormone therapy were more likely to be proactive about their health: visiting doctors, exercising, watching diet, etc., and therefore were less likely to experience the effects of cardiovascular disease?

Population Inference

- Inference to Populations: Can we extend the results of our analysis beyond the sample we analyzed?
- Example:
  - Suppose we are interested in whether Fuji apples grown using organic farming methods have higher Vitamin C content than Fuji apples grown using conventional farming methods.
  - We cannot test all Fuji apples grown organically and conventionally, so we would like to use a **sample** of organic apples and a **sample** of conventional apples to try to learn about the populations (all organic Fuji apples and all conventional Fuji apples).
- When is it reasonable to expect that the results in a sample might be representative of the population values?

Inference to Population(s)

- To extend inference from your sample to some larger population(s), you must have sampled from those *exact population(s)* in some kind of *'representative'* way:
  - If we only sample employees at McDonald's, it would be wrong to make inference to the population of fast-food employees (McDonald's employees may not be representative of all fast-food employees).
  - If we only sample fish from a single alpine lake, it would be incorrect to make inference to the population of all fish in all alpine lakes (Fish in this particular lake may not be representative of fish in other lakes).

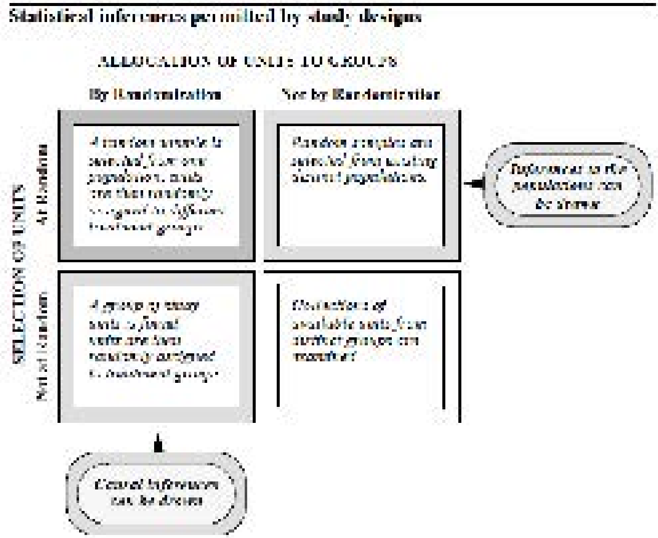
When is Population Inference Possible?

Inferences to populations can be drawn from representative random sampling studies, but not otherwise.

As discussed previously, ‘representative random sampling’ could include

- Simple random sampling
- Systematic sampling
- Cluster sampling
- Stratified random sampling

Causal and Population Inference: Summary



Statistical inferences permitted by study designs

ALLOCATION OF UNITS TO GROUPS

SELECTION OF UNITS	By Randomization	Not by Randomization
At Random	*A random sample is selected from one population; units are then randomly assigned to different treatment groups.*	*Random samples are selected from existing distinct populations.*
Not at Random	*A group of study units is found; units are then randomly assigned to treatment groups.*	*Collections of available units from distinct groups are examined.*

Inferences to the populations can be drawn

Causal inferences can be drawn