Study design

9/11/24

Housekeeping

- Please bring your laptops tomorrow! We will be installing R and RStudio!

Key terminology

Population and samples

Population and samples (cont.)

- Typically, the size of the sample is way smaller than the population. Why?
  - In the lucky event that we are able to collect data for every individual in the population, the sample is referred to as a **census**
- Example: the U.S. Census Bureau is responsible for producing data about the American people and economy. They collect data with different schemes and frequency:
  - Decennial census
  - American community survey

Parameters and statistics

- Often times, answering the research question simplifies to understanding a numerical summary.
  - A numerical summary calculated from (or considered for calculation from) the entire population is called a (population) **parameter**
  - In contrast, a numerical summary calculated from the sample is called a (sample) **statistic**
- Why do we differentiate? It's always good to remember that we are trying to answer questions about the population!

Sampling from a population

A “good” sample

- The way we sample data from a population can directly influence the quality of that sample.
- What are desirable characteristics of a sample?
  - **Representative**: the sample roughly “looks like” the population, i.e. the individuals in the sample offer a good representation of the population at large
  - **Generalizable**: any results based on the sample can generalize to the population, i.e. we can make “good guesses” about a population parameter using a sample statistic
  - **Unbiased**: every individual in the sample had an equal chance of being sampled

Bias in a sample

Bias in a sample can arise due to many causes. Here are a few:

- Selection bias
  - Often arises when convenience sample is taken
- Exclusion/undercoverage bias
- Non-response bias
- Response bias

Example: *Literary Digest* poll

- 1936 was an election year in the United States. Franklin D. Roosevelt (a Democrat) was completing his first term in office as president.
- Republican candidate Alfred Landon of Kansas was his competitor
- *Literary Digest* magazine conducted a polling survey, which received 2.4 million respondents (largest number of people every replying to a poll at that time)
  - Prediction: overwhelming victory for Landon (predicted FDR would only get **43%** of popular vote)
- Actual result: FDR won by a landslide! (**62%** to 38%)
- What happened? Selection and non-response bias

Probability sampling

- Any sampling method where the selection from the target population is based on random selection/chance
  - All subjects in the target population have equal chances of being selected
  - No one has discretion about who is included in the sample
- Randomly sampling from the population can help reduce bias in our sample
  - If we don’t randomly sample, results obtained from the sample will most likely not be representative and will not generalize to the target population
- Examples include: simple random, stratified, cluster, systematic

Simple random sampling (SRS)

- In a **simple random sample**, each individual is chosen entirely by chance from the population, and each member of the population has an equal chance of being sampled
  - Typically sampling *without* replacement
  - Knowing that an individual was sampled does not provide useful information about which other cases are included
  - Any given fixed-size subset of the population is equally likely to be chosen
- Consider again the research question: What proportion of current Middlebury professors attended a liberal arts college?

  How might I obtain a sample random sample of 25 professors?

Stratified sampling

- Depending on the population, we might need to use different random sampling methods to ensure the sample is representative.
- Assume that the population is/can be broken up into several different, distinct sub-populations or **strata**
  - The division should "make sense"
- Rather than randomly sampling from the population as whole, we take a random sample from *each stratum*
  - How many from each stratum? Typically use a sampling fraction that is proportional to entire population!
  - E.g. if population of trees on Middlebury campus are 80% deciduous and 20% coniferous and we want to sample $n = 10$ trees total, we should *randomly* sample ___ deciduous and ___ coniferous trees
- What are some pros/cons?

Cluster sampling

- Divide total population into M distinct groups or **clusters** of roughly equal size
- Perform a simple random sample on the M clusters, than sample all individuals within each of the randomly selected clusters
- Discuss the following:
  - Would you prefer the individuals within a cluster to be homogeneous (similar) or heterogeneous (varied)? Why?
  - Would you prefer that cluster A and cluster B be relatively similar or different in terms of their sub-populations?
  - What is the difference between stratified and cluster sampling?

Multistage cluster sampling

- Builds on the cluster sampling method, but rather than sampling all individuals within the selected clusters, only collect a simple random sample within each selected cluster
  - Can make more stages/layers if appropriate!
- Though seemingly more complicated, why might we prefer multistage sampling over cluster sampling?
- How might we devise a multistage cluster sample for Literary Digest?

Experiments vs. Observational studies

We now know *how* to collect data, but now we turn to examining *what kind* of study we'd like to perform in order to answer the research question. We begin with experiments. But first…

What is a variable?

Explanatory vs. Response

- Lots of scientific questions revolve around asking how x relates to y
- If y is the primary variable of interest, i.e. the variable whose behavior we want to understand, it is called the **response variable**
- If we try to understand how changing x affects y, then x is called the **explanatory variable**
    - Explanatory variables can often be manipulated/controlled/observed by the researcher ahead of time

Experimental design

- **Experiments** are studies where the researcher *assigns* treatments to cases
  - Note: experiments are often conducted in medical settings, hence the word "treatment"
  - Are treatments considered explanatory or response variables?
- When the researcher *randomly* assigns the treatments, we have a **randomized experiment**
  - Randomized experiments are critical when trying to assess the causal effect of the explanatory variable on the response variable
  - Note: random assignment $\neq$ random sampling

Confounding variables

- Understanding a *causal* relationship is made difficult by **confounding variables**: variables that are associated with both the explanatory and response variable of interest
  - Confounders are bad!! Why?
- Example: consider a study that seeks to examine the effect of coffee consumption on heart disease.
  - From each person, we only collect information on the average amount of coffee they consume per day and whether or not they have heart disease.
  - We find a positive association: more coffee → higher risk of heart disease
  - Possible confounder: smoker status. Smokers tend to drink more coffee and tend to have higher rates of heart disease than non-smokers.
  - So the increase in heart disease may be due to smoker status rather than caffeine intake

Principles of experimental design

1. **Randomization**: randomly assign patients to treatments
    - Helps account for variables that cannot be controlled and possible confounding variables
2. **Controlling** for differences in the treatment: ensure that everyone follows the same protocol exactly
3. **Replication**: the more cases we observe, the more confidence we have in the effect of the explanatory on the response
    - Achieved by collecting a sufficiently large sample in a single study, or repeating the entire study more than once

These first three principles are crucial! The following is principle is desirable, but more complicated.

Principles of experimental design (cont.)

4. **Blocking**: suppose we know ahead of time that there is/are variable(s) that could influence the response besides just the explanatory. We assign patients to their respective blocks, and than randomly assign treatments within blocks.
    - Helps to decrease unexplained variability by accounting for **nuisance variables**: variables that affect the response variable but are not of interest for answering the scientific question
    - Can lead to greater interpretation of results

Treatment vs. control

- Randomized experiments are the gold standard for data collection, but biases can still occur!
- When we want to learn if the explanatory variable causes some effect in the response variable.
  - We have a **control group** which establishes a baseline, and typically receives "zero amount" of the explanatory variable.
  - We also have a **treatment group** which receives some "non-zero amount" of the explanatory variable
- Example: suppose we want to test the effect of a drug that is developed to help people fall asleep.
  - Treatment group: receives 50mg of the drug in pill form
  - Control group: does not receive the drug at all
  - What is a potential issue?

Reducing bias in human experiments

- We should make the experiment a **blind experiment** by not allowing participants to know which group they've been assigned to
  - Give a fake treatment known as a **placebo** to those in the control group (e.g. a sugar pill that looks exactly like the actual treatment pill)
  - Oftentimes, a placebo results in a slight but real improvement in patients. This is known as the **placebo effect**
- Doctors and researchers involved in the study should also be blinded so they do not give preferential treatment or care to patients in certain groups.
  - **Double-blind** experiments: both the patients and the doctors/researchers who interact with patients are unaware of who is or is not receiving the treatment
- Question of ethics

Observational studies

- Studies where no treatment is explicitly applied
- Nothing is manipulated; researchers simply record/observe without intervening
- Typically cannot obtain causal conclusions using data from observational studies
  - There are too many confounding variables at play in observational studies
- But we can use these studies to identify associations or form hypotheses for future experiments!