Study design

2/12/25

Housekeeping

- Please bring your laptops tomorrow installed with R and RStudio!
- Optional reading: [Chapter 2 Study Design] from textbook

Sampling from a population

Variables in statistics

What is a variable?

- Lots of research questions revolve around asking how variable x affects variable 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**
    - In scientific studies, explanatory variables can often be manipulated/controlled/observed by the researcher ahead of time

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 characteristics of participants in the sample are similar to those of the population
    - **Generalizable**: any results based on the sample can generalize to the population
        - i.e. we can use results from a sample to draw conclusions about a specific population

Bias in a sample

- **Biased** samples occur when the methods used to obtain data result in inaccurate/skewed depictions of the population. This is bad!!
  - Can occur if a sample is not representative
- Bias in a sample can arise due to many causes. Here are just a few:
  - **Selection bias**: systematic tendency in procedure that causes some members of population to be more likely to be included than others
  - **Non-response bias**: the values of the response variable of non-respondents differ systematically from those that do respond
  - **Response bias**: systematic favoring of certain response variable values that occurs when people don't answer truthfully (e.g. lying)
- Any type of bias could lead to our sample being non representative or not generalizable

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

Sampling methods

Convenience sampling

The worst kind of sampling (but often the easiest)!

- **Convenience sampling** takes place when cases that are easily accessible are more likely to be included in the sample
- Example:
  - Population: students enrolled in statistics courses at Middlebury
  - Sample: students in STAT 201 AZ

Probability sampling

- In general, sampling methods that include a bit of randomness can help reduce the chance of bias
- **Probability/random sampling**: any sampling method where the selection from the population is based on random selection/chance
  - No one has full discretion about who is included in the sample
- **Random sampling usually yields a representative and generalizable sample**
- Examples include: simple random, stratified, cluster, systematic

1. Simple random sampling (SRS)

- In a **simple random sample**, each case is chosen entirely by chance from the population, and each member of the population has an equal chance of being sampled
  - 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 simple random sample of 25 professors?

2. Stratified sampling

- Assume that the population is/can be broken up into several different, distinct sub-populations or **strata**
  - Cases grouped into a strata should be similar to each other
- Then take a (simple) random sample from ***each*** stratum ("divide and conquer")
  - 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?

3. 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?

Types of 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.

Experiments vs Observational studies

- **Observational studies** occur when a research *observes* cases without manipulating any variables
- **Experiments** are studies where the researcher *assigns* specific treatments to cases
  - Note: experiments are often conducted in medical settings, hence the word “treatment”
- Example: I want to design a study to learn if students who take quizzes throughout the semester end up performing better on the final exam.
  - Observational study: students optionally take quizzes
  - Experiment: I choose half of the students to take quizzes and the other half to not take quizzes.
- Are treatments in experiments considered explanatory or response variables?

Treatment vs control

- Treatments are typically divided into two categories:
   1. **Control group**: establishes a baseline, and typically receives “zero amount” of the explanatory variable
   2. **Treatment group(s):** receive some “non-zero amount” of the explanatory variable
- Quiz example continued:
   - Control group: no quizzes
   - Treatment group 1: takes one quiz
   - Treatment group 2: takes two quizzes
   - How to decide which case gets which treatment?

Randomized experiments

- 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 in experiments $\neq$ random sampling for participation in the sample
- Continuing example:
  - Randomized experiment is achieved if I use SRS to determine who received which treatment
  - But the students who "participate" in the experiment were not obtained via SRS

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 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

Reducing bias in human experiments

- Biases can still unintentionally arise in experiments, even if we follow these three principles.
- 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)
    - **Placebo effect:** a placebo results in a slight but real improvement in control patients
- 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 receive which treatment

Reducing bias in human experiments (cont.)

Caveats:

- Blinding not always possible! It would be hard to give a placebo in the quiz experiment
- Question of ethics

Observational studies

- Causal conclusions ***cannot*** be obtained using data from observational studies
  - There are too many confounding variables at play!
- But they are much cheaper, and can be used to identify associations or form hypotheses for future experiments!