# Chapter 4: Gathering Data

Yu Yang

School of Statistics
University of Minnesota

September 16, 2022

# Prerequisites

**Variables:** Explanatory variables (Independent variables) - Response variable (Dependent Variable)

**Examples:** If we measured mother's heights and daughter's height.
Explanatory Variable: Mother's Height
Response Variable: Daughter's Height

# Types of Studies

- **Experiment:** A researcher conducts an experiment by **assigning** subjects to certain experimental conditions and then observing outcomes on the response variable. The experiment conditions are called *treatments*.
- **Observational Study:** The researcher **observes** values of the response variable and explanatory variables for the samples subjects without anything being done to the subjects.

# Examples

## Example of Experiments

- To examine whether a new medicine is effective in reducing blood pressure, researchers hire 100 patients who have similar blood pressure situation, and then randomly assign them either the medicine and placebo and measure their blood pressure after a period of time.

## Example of Observational Studies

- A sample survey selects a sample of people from a population and interviews them to collect data.
- A census survey is a survey that attempts to count the number of people in the population and to measure certain characteristics about them.

# Experiment or Observational Study?

You are asked to collect data that would address the following question:

> *Does smoking during pregnancy have a negative impact on the baby?*

Would you set up an observational study or an experimental study?

# Experiments vs Observational Studies

An experiment reduces the potential for **lurking variables** to affect the result. Thus an experiment gives the researcher more control over outside influences. Only experiments can establish cause and effect. Observational studies can not.

At the same time, experiments are not always possible due to ethical reasons, time considerations and other factors.

# Sampling

- **Sampling Frame:** List of subjects in the population from which the sample is taken, ideally it lists the entire population of interest.
- **Sampling Design:** Determines how the sample is selected.

# Good ways to Sample

## Simple Random Sampling (SRS)

Random Sampling is the best way of obtaining a sample that is representative of the population. A *simple random sample* of $n$ subjects from a population is one in which each possible sample of that size has **the same of chance of being selected**. This is often just called a random sample.

# Poor ways to Sample

- **Convenience Sample:** a type of survey sample that is easy to obtain If I give the Snickers out to only the 25 people nearest to me
  - unlikely to be representative
  - often severe biases result from such a sample
  - results apply ONLY to the observed subjects

- **Volunteer Sample:** most common form of convenience sample. If I ask "who wants Snickers?"
  - Subjects volunteer for the sample
  - Volunteer do not tend to be representative of the entire population

# Bias

**Bias:** When certain outcomes will occur more often in the sample than they do in the population.

- *Sampling bias* occurs from using nonrandom samples
- *Nonresponse bias* occurs when some sampled subjects cannot be reached or refuse to participate or fail to answer some questions.
- *Response bias* occurs when the subject gives an incorrect response (perhaps lying) or the way the interviewer asks the questions (or wording of a question in print) is confusing or misleading.

A large sample does not guarantee an unbiased sample!

# Summary

- Identify the population of all subjects of interest
- Construct a sampling frame which attempts to list all subjects in the population.
- Use a random sampling design to select $n$ subjects from the sampling frame.
- Be cautious of sampling bias due to nonrandom samples (such as volunteer samples) and sample undercoverage, response bias from the subjects not giving their true response.

We can make inferences about the population of interest when random sampling is used.