This class is being conducted over Zoom. As the instructor, I will be recording this session. I have disabled the recording feature for others so that no one else will be able to record this session. I will be posting this session to the course's website.

If you have privacy concerns and do not wish to appear in the recording, you may turn video off (click "stop video") so that Zoom does not record you.

The chat box is always open for discussion and questions to the entire class. You may also send messages privately to the instructor or the TAs. Please note that Zoom saves all chat transcripts.

I create a live transcription of each session using Otter.ai. This means that Otter.ai will transcribe anything spoken over the Zoom audio. The transcript will be posted with the session video on the course website.

Sampling & Bias

Stats 7

Mary Ryan


Course website:

https://canvas.eee.uci.edu/courses/28451


Slides can be found at:

https://maryryan.github.io/stats7-SS2-2020-slides/stats7-SS2-2020-samplingBias/stats7-SS2-2020-samplingBias

Learning Objectives

By the end of today's lecture you should:

- understand the difference between a sample and a population
- be able to identify different data sampling techniques
- be able to explain how bias can be introduced into data collection and what we can do to prevent it
- be able to differentiate between experiments and observational studies

Populations & Samples

- A population is all of the members of a specific group
  - All the animals at the Irvine animal shelter
  - All the books in the Orange County Public Library system
  - All the people living in California

Populations & Samples

- A **population** is all of the members of a specific group
  - All the animals at the Irvine animal shelter
  - All the books in the Orange County Public Library system
  - All the people living in California
- A larger population can be made up of several smaller sub-populations
  - All the dogs at the Irvine animal shelter
  - All the novels in the Orange County Public Library system
  - All the people living in Orange County

Populations & Samples

- A **population** is all of the members of a specific group
  - All the animals at the Irvine animal shelter
  - All the books in the Orange County Public Library system
  - All the people living in California
- A larger population can be made up of several smaller sub-populations
  - All the dogs at the Irvine animal shelter
  - All the novels in the Orange County Public Library system
  - All the people living in Orange County
- A **sample** is a subset of a population
  - Five animals at the Irvine animal shelter
  - Twenty-five books in the Orange County Public Library system
  - Fifty-seven people living in California

Why Sample?

- The goal of a **census** is to collect information about an entire population
  - We want to know about a population's **parameters**
    - i.e., population's true mean age, true proportion of homeless, true range of yearly incomes, etc.

Why Sample?

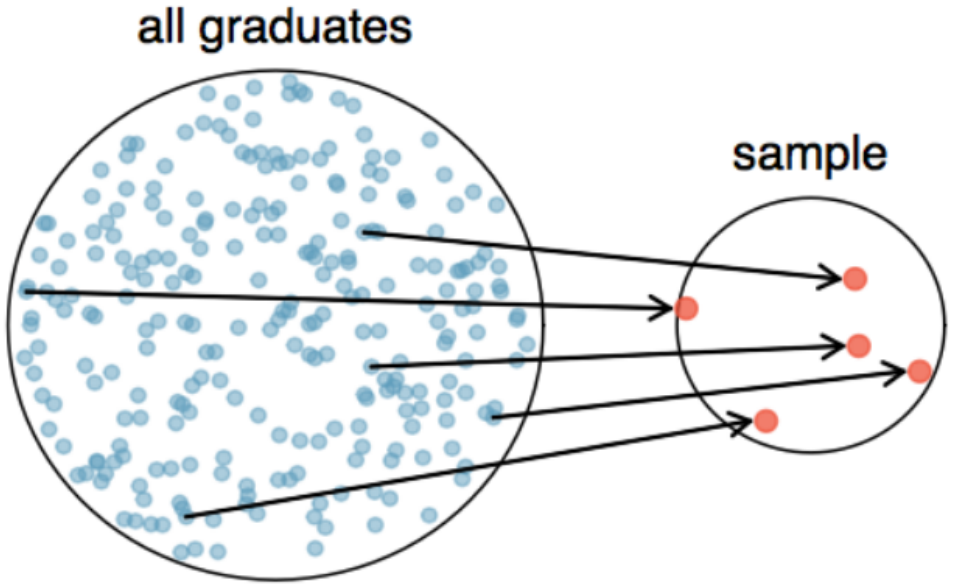
- The goal of a **census** is to collect information about an entire population
  - We want to know about a population's **parameters**
    - i.e., population's true mean age, true proportion of homeless, true range of yearly incomes, etc.
- We want to learn about a population, but populations are often very large
  - Difficult to observe all populations members
  - Observing every member may be very expensive or take a very long time
- By selecting a sample and learning about its members, we assume their characteristics or opinions are representative of the whole population
  - The act of selecting a sample is known as **sampling**
  - An estimate of a population parameter (using sample data) is known as a **statistic**

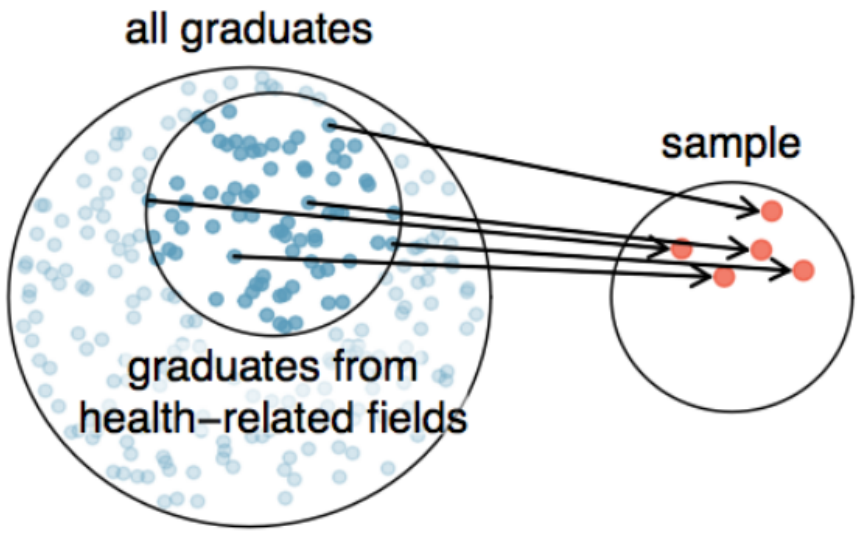
Why Sample?

Parameters come from **p**opulations

Statistics come from **s**amples

Population of Interest





- What do we want our **population of interest** to be?
- How widely do we want to **generalize** our results?

Sampling

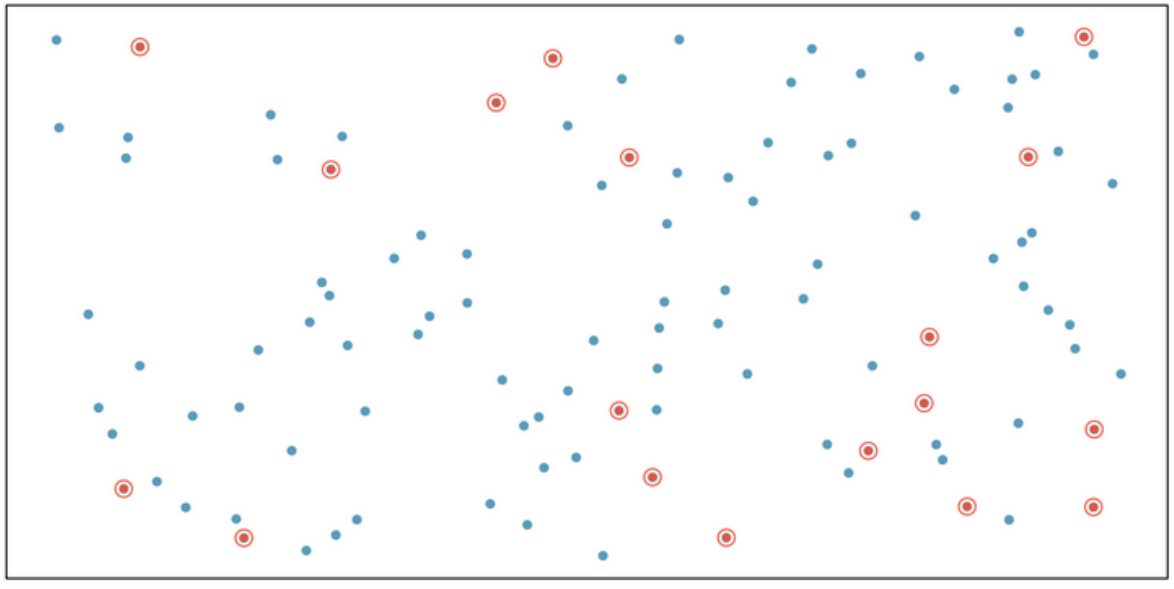
- We want to make sure our sample is **unbiased**, or not systematically different from the population
- Whenever possible, we want to ensure that sampling units are taken from the population at **random**
    - This reduces the chance of introducing (sampling) bias, like **self-selection bias**

Sampling

- We want to make sure our sample is **unbiased**, or not systematically different from the population
- Whenever possible, we want to ensure that sampling units are taken from the population at **random**
  - This reduces the chance of introducing (sampling) bias, like **self-selection bias**
- Why might there be problems with not selecting sampling units at random?
  - Conducting a survey at Disneyland about guest preferences but only selecting people in high heels
  - Wanting to learn about shopping habits of all people in U.S., and we only talk to our friends
- Want to avoid **convenience samples** when possible
  - Easily accessible individuals have a higher probability of being selected

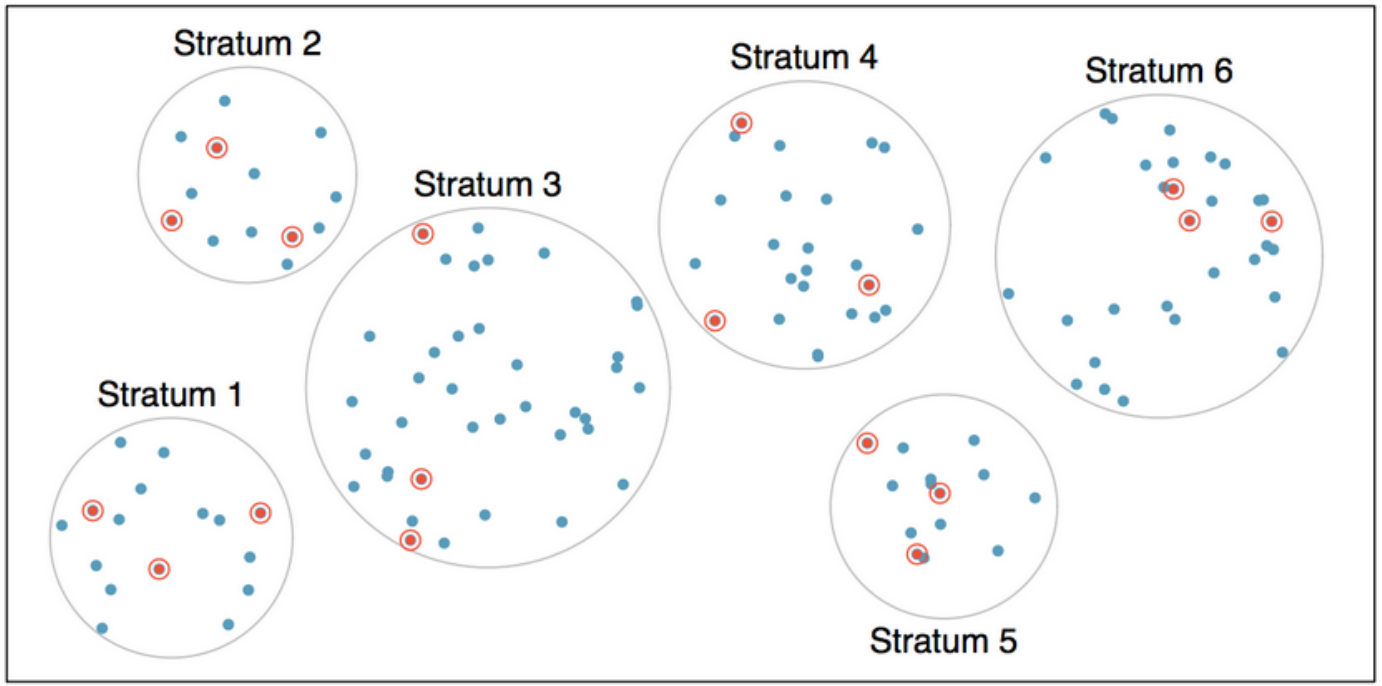
Types of Random Sampling

- Many types of sampling techniques, each with pros and cons depending on your goal
- Simple random sample: sampling technique in which each case in a population has an equal chance of being included in the final sample and knowing that a particular case is in the sample gives us no knowledge about the other cases in the sample
    - This is the most common type of sampling you will encounter
    - When we say "randomly assign", we generally mean simple random sample



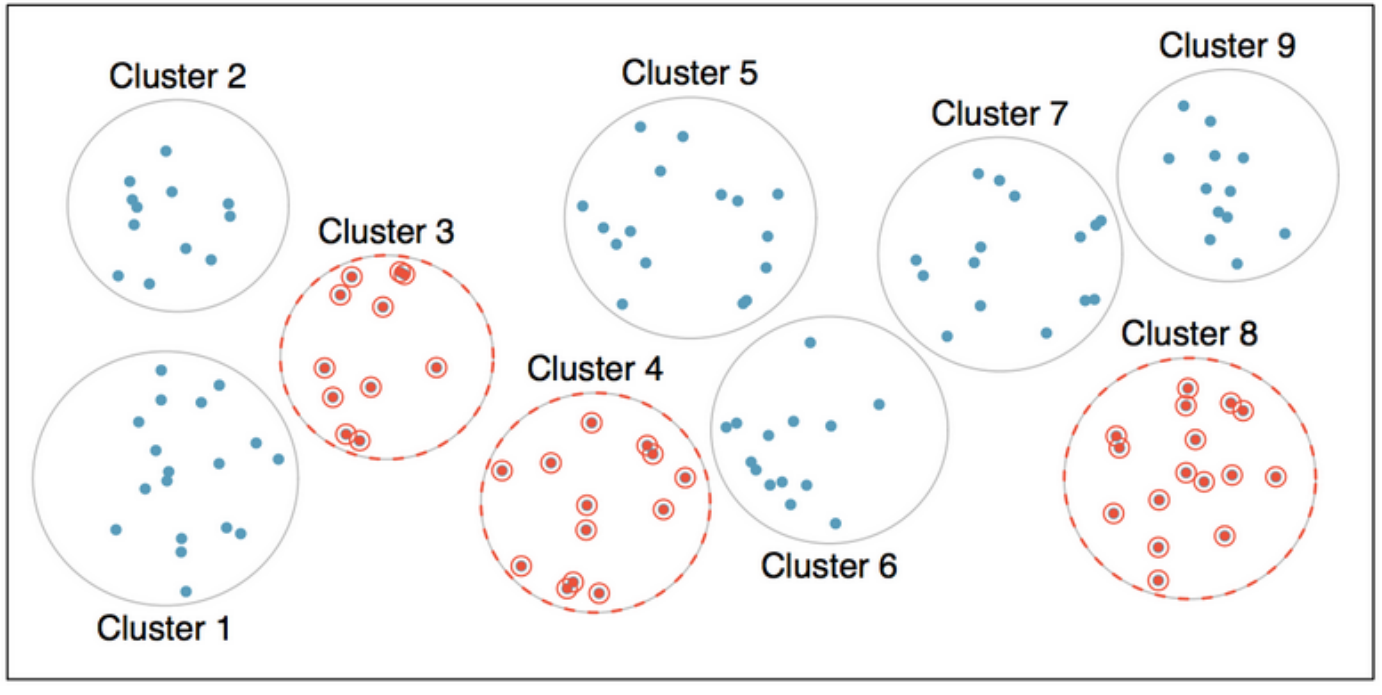
Types of Sampling

- **Stratified sampling**: the population are divided into mutually exclusive groups called **strata** and then a simple random sample is taken from each strata
    - Cases within a strata will have something in common like gender, age, or race
    - Two strata will look different, while the cases within a strata will look similar



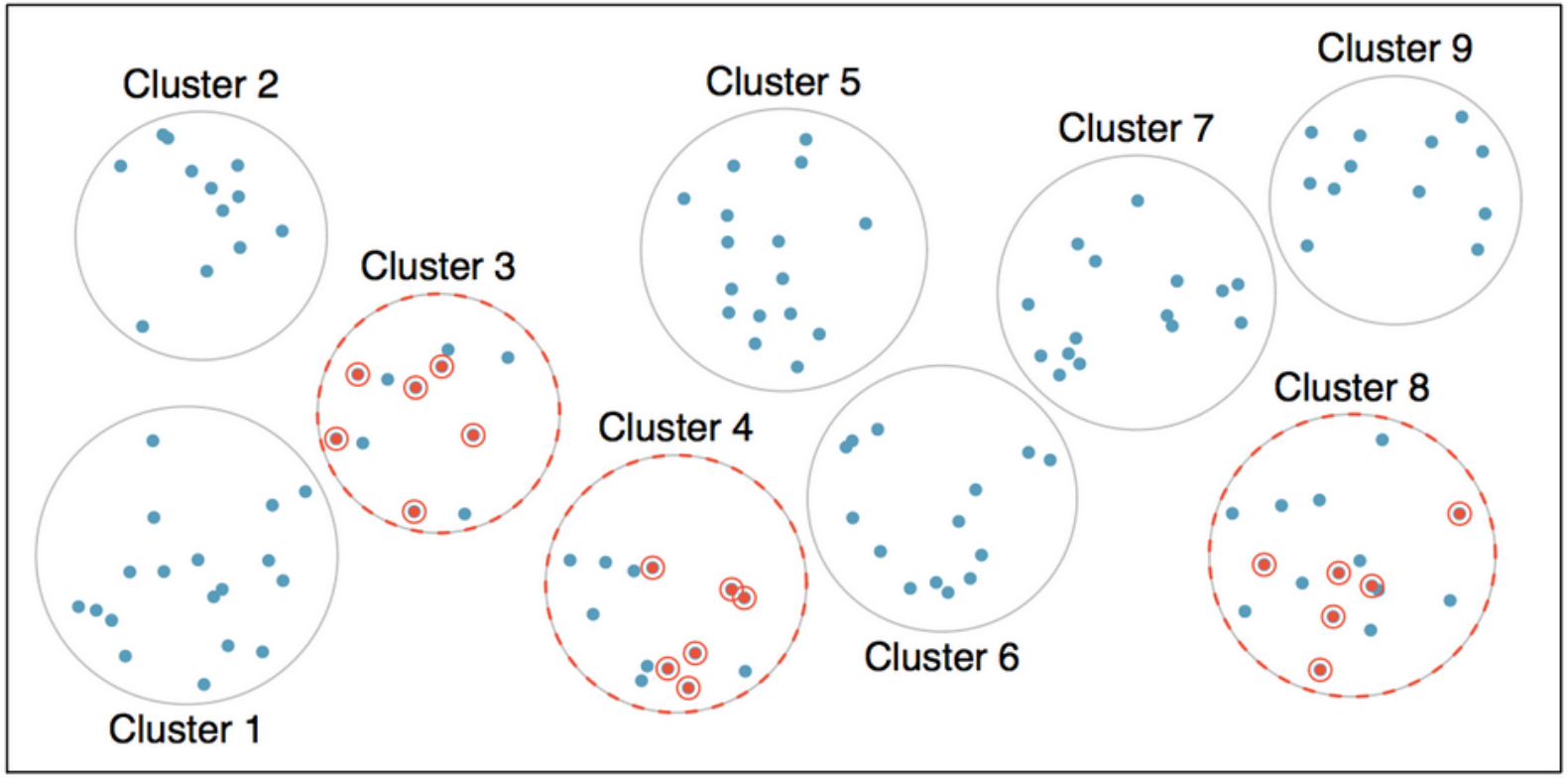
Types of Sampling

- Cluster sampling: the population is divided into naturally occurring groups called clusters, and then entire clusters are randomly chosen
    - A common clustering is geographical location
    - Often, there will be differences within a cluster, but two clusters considered as a whole will look similar



Types of Sampling

- **Multi-stage cluster sampling**: similar to cluster sampling, entire clusters are chosen at random, then a simple random sample is taken of the individual cases within those selected clusters



Sampling Example 1

Suppose we are interested in estimating the malaria rate in a densely tropical portion of rural Indonesia. We learn that there are 30 villages in that part of the Indonesian jungle, each more or less similar to the next. To sample 150 individuals to test for malaria, we randomly select half of the villages, then randomly select 10 people from each village.

- What type of sampling procedure does this describe?

Sampling Example 2

Consider the salaries of Major League Baseball (MLB) players, where each player is a member of one of the league's 30 teams. To take a sample of 120 baseball players and their salaries, we could write the names of that season's several hundreds of players onto slips of paper, drop the slips into a bucket, shake the bucket around until we are sure the names are all mixed up, then draw out slips until we have the sample of 120 players.

- What type of sampling procedure does this describe?

Sampling Example 3

A university wants to determine what fraction of its undergraduate student body support a new $25 annual fee to improve the student union. To obtain a sample of students, administrator break the students by their field of study, then sample 10% of students from each field.

- What type of sampling procedure does this describe?

Types of Studies: Experiments

- Two types of studies: experimental and observational
- **Experimental studies** are those where researchers explicitly administer a treatment
- Researchers **randomly** assign volunteers to treatment and control groups
- We can also reduce bias in experiments by blinding
  - Studies that are **blinded** mean that the volunteer doesn't know whether they're in a treatment or control group
  - Studies that are **double-blind** mean that neither the volunteer nor the treatment administrator know whether they're in the treatment or placebo groups
    - Usually the treatment administrator is a "middle man" who is given a treatment from the researchers to give to the volunteer

Types of Studies: Experiments

- Two types of studies: experimental and observational
- **Experimental studies** are those where researchers explicitly administer a treatment
- Researchers **randomly** assign volunteers to treatment and control groups
- We can also reduce bias in experiments by blinding
    - Studies that are **blinded** mean that the volunteer doesn't know whether they're in a treatment or control group
    - Studies that are **double-blind** mean that neither the volunteer nor the treatment administrator know whether they're in the treatment or placebo groups
        - Usually the treatment administrator is a "middle man" who is given a treatment from the researchers to give to the volunteer
- **Why might we do this?**

Types of Studies: Experiments

- Two types of studies: experimental and observational
- **Experimental studies** are those where researchers explicitly administer a treatment
- Researchers **randomly** assign volunteers to treatment and control groups
- We can also reduce bias in experiments by blinding
    - Studies that are **blinded** mean that the volunteer doesn't know whether they're in a treatment or control group
    - Studies that are **double-blind** mean that neither the volunteer nor the treatment administrator know whether they're in the treatment or placebo groups
        - Usually the treatment administrator is a "middle man" who is given a treatment from the researchers to give to the volunteer
- **Why might we do this?**
    - Blinding helps ensure that subjects or researchers aren't behaving consciously (or subconsciously) to **influence the outcome**
    - AKA: **placebo effect** or **placebo bias**

Principles of Experimental Design

There are several principles we need to keep in mind when we're designing experiments

- Controlling for Influencing Factors
  - Researchers do their best to reduce the differences between groups to isolate effect of treatment
  - Done by developing a thorough study design
    - Noting factors that may affect response (sex, other health issues, etc.)
    - Writing into the experiment design ways to reduce their influence on the response (making sure equal numbers of men and women are in each group, excluding people who have had a heart attack in the last 5 years, etc.)
  - If done well, we can therefore make causal conclusions: "Treatment X causes Y effect in the response"

Principles of Experimental Design

There are several principles we need to keep in mind when we're designing experiments

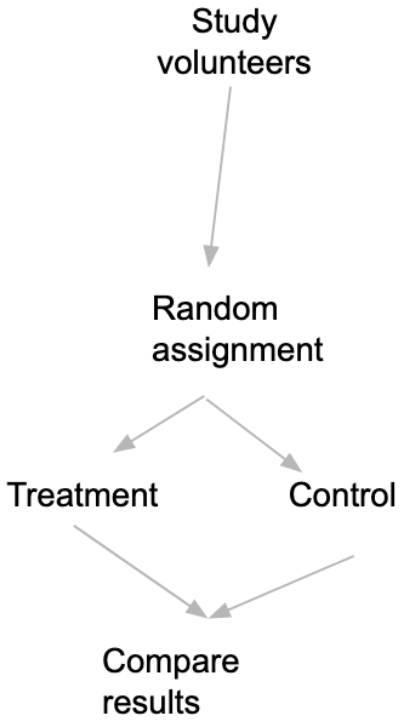
- Controlling for Influencing Factors
  - Researchers do their best to reduce the differences between groups to isolate effect of treatment
  - Done by developing a thorough study design
    - Noting factors that may affect response (sex, other health issues, etc.)
    - Writing into the experiment design ways to reduce their influence on the response (making sure equal numbers of men and women are in each group, excluding people who have had a heart attack in the last 5 years, etc.)
  - If done well, we can therefore make causal conclusions: "Treatment X causes Y effect in the response"
- Randomization
  - Researchers randomize volunteers into groups to account for factors that can't be controlled (often because we can't think of every little thing to control for)
  - If factors are seen in equal amounts in both treatment and control groups, they won't unevenly influence the results of one group more than the other

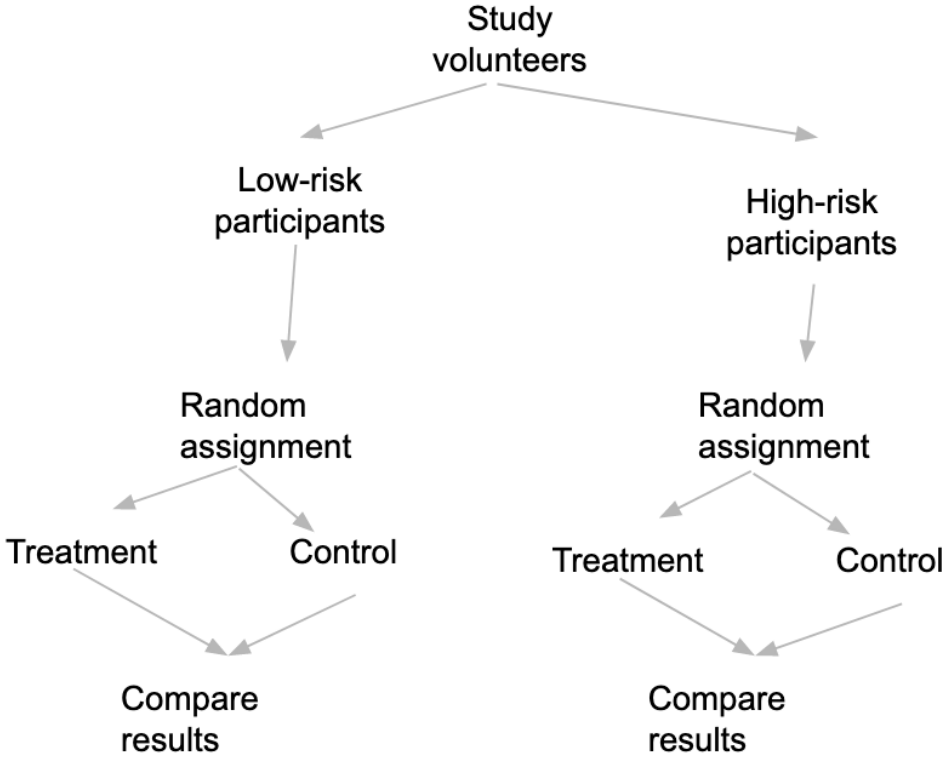
Principles of Experimental Design

- Replication of Results
  - Entire experiments can be repeated to double-check results (though this doesn't happen often)
  - Writing and publishing clear experimental protocol will help other researchers reproduce your results
  - The smaller a sample you take, the more likely the effects you're seeing are specific to that group of people
    - To make an experiment replicable, a .alert2[**large**] enough sample must be taken
- Blocking *
  - Sometimes necessary to divide your sampling pool into blocks based on a similar trait, and run the experiment on both blocks
  - After dividing into blocks, each block gets divided into control and treatment groups
  - Helps make conclusions and results more specific

* Not always necessary

Completely Randomized vs. Block Design





Other Types of Experiments: Matched-Pairs & Cross-Over

- Matched-Pairs experiments are experimental studies where similar study participants are placed into pairs, and one half of the pair will receive the treatment and other other half will receive the control
    - Which pair half receives treatment is randomized
    - Making within pair comparisons
    - i.e., Twin studies
- Cross-over experiments are experimental studies where all study participants will end up receiving both the treatment and control, but the order in which they receive them is randomized
    - Participant becomes their own control
    - Making within individual comparisons
    - i.e., Take a drug for 2 weeks, wait 4 weeks for effects to "wash out", take placebo for 2 weeks

One Minute Reflection

- Take 60 seconds to catch up on notes and reflect what we just went over
- Did you have an "aha!" moment?
- Are there points that are still unclear?

Types of Studies: Observational

- Sometimes it's not possible or ethical to conduct an experiment
- **Observational studies** are those where researchers passively observe subjects and record certain variables as time goes on
- Because researchers cannot control outside factors, though, **we cannot draw causal conclusions**

Common Types of Observational Studies

- Cross-sectional studies are when we take one random sample of the population, at one point in time
    - i.e., Administering a survey on attitudes about the Irvine Company to a random sample of Irvine residents
- Longitudinal or repeated-measures studies are when we take a random sample of the population, and observe them over multiple points in time
    - i.e., Taking a random sample of students at UCI, but recording their GPAs for each quarter
- Case-control studies are when we take separate random samples from the "case" population and the "control" population
    - Often used to study rare diseases
    - Think of "case" as those with a condition, and "control" as those without the condition

Types of Studies vs. Types of Conclusions

Study Type	How Sample is Collected: Sample Representative of Population	How Sample is Collected: Sample Not Representative of Population
Experiment	**Can** conclude causal relationship **Can** generalize results to entire population	**Can** conclude causal relationship **Cannot** generalize results to entire population
Observational Study	**Cannot** conclude causal relationship **Can** generalize results to entire population	**Cannot** conclude causal relationship **Cannot** generalize results to entire population

Experiments vs. Observational Studies: Pros and Cons

Pros	Cons	Pros	Cons

Other Common Types of Study Bias

- Undercoverage: parts of the population will be left out depending on the sampling method (i.e., the Census and homeless people)
- Nonresponse: certain types of people choose not to respond or participate
- Wording Effects: questions may be leading, biased, or confusing
- Response Bias: people can lie about their answers to make themselves sound like better people

Bias Example 1

A statistics student who is curious about the relationship between the amount of time students spend on social networking sites and their performance at school decides to conduct a survey.

They post a link to an online survey on Facebook and asks their friends to fill out the survey.

- What bias (if any) might you expect from this method?

Bias Example 2

A statistics student who is curious about the relationship between the amount of time students spend on social networking sites and their performance at school decides to conduct a survey.

They randomly sample 40 students from the study's population, gives them the survey, asks them to fill it out and bring it back the next day.

- What bias (if any) might you expect from this method?

Bias Example 3

A statistics student who is curious about the relationship between the amount of time students spend on social networking sites and their performance at school decides to conduct a survey.

They randomly sample 5 classes and asks a random sample of students from those classes to fill out the survey.

- What bias (if any) might you expect from this method?

A Caveat

But yesterday you sent us an email about a UCI professor asking UCI students to volunteer for brain studies - that's not a random sample!

A Caveat

But yesterday you sent us an email about a UCI professor asking UCI students to volunteer for brain studies - that's not a random sample!

- Well... yes
- Important to try to get a random sample whenever possible, but people still need to consent to participating (which also introduces bias!)
- Also sometimes not possible to logistically get a random sample

A Caveat

But yesterday you sent us an email about a UCI professor asking UCI students to volunteer for brain studies - that's not a random sample!

- Well... yes
- Important to try to get a random sample **whenever possible**, but people still need to **consent** to participating (which also introduces bias!)
- Also sometimes not possible to logistically get a random sample
- Sometimes best thing you can do is recognize that this will potentially **bias** your results and **reduce generalizability**
- Excerpts from papers I've helped write:
  - "Participants were recruited from a local recruitment registry and may not be representative of the general population."
  - "The survey was sent to participants of a local recruitment registry. This population may have more favorable attitudes toward research (and potentially genetic testing) than the general public."
  - "The generalizability of these data is limited by the small sample size and possible sample bias resulting from recruiting from an ongoing research study."

Be Your Own Researcher Activity (20 minutes)

Split up into breakout room groups & make up your own study

Steps

1. Come up with a team name
2. Identify & describe:
    - Research question
    - Population of interest
    - Sampling strategy
    - Response/dependent variable
    - Variable of interest/independent variable
    - Variables to be recorded
3. Create a short presentation (< 5 slides) on your proposed study

Roles to be assigned:

- Spokesperson (will present to rest of class if called on)
- Presentation designer
- Note taker
- Time keeper

As you present your study, we will try to identify:

- Study type
- Types of data being collected
- Types of bias your study may suffer from