Welcome to STAT100A!

Instructor: Lauren Cappiello, M.S.

July 29, 2019

Welcome!

A little about me:

- Why statistics?
- What else do I do?

Administrative Business

Course Website: lgpcappiello.github.io/teaching/stat100a/su19

- Bookmark this page!
- Website includes course calendar, assignments, links, and syllabus. There is also a link to download our open access textbook.
- Course site is linked to on iLearn.
- Grades will be posted on iLearn.

Syllabus

The syllabus is your first and only required reading. (Reading the textbook is also highly recommended!)

- If you have any questions about the syllabus, we will take a few minutes tomorrow to talk about it.

Textbook

OpenIntro Statistics by David Diez, Mine Cetinkaya-Rundel, and Christopher Barr.

- This is an open source (aka free) textbook.
- Homework problems will come directly from the textbook.
- Please feel free to provide me with feedback on the book!
  - Is it easy to follow? Is it okay to read? How are the homework problems? Etc.

Office Hours

Instructor

Lauren Cappiello (OLMH 1417) XX XX-XX am

Teaching Assistants

Rajasekhar Anguluri (OLMH 1102), office hr
Deepak Bastola (OLMH 1415), XXXX
Po-Yao Niu (OLMH 1108), XXXX

We are also all available by appointment.

Discussions

Discussions will consist of

- Review of material
- Student questions (including homework help)
- A weekly quiz

Labs

Labs will consist of

- Review
- Computer-based lab activities

Lab activities are not designed to take the full 3 hours! After you have completed your lab activity, you may work on homework, ask questions, or leave early.

Answers to some FAQs

- Research suggests that handwritten notes are one of the best ways to learn and retain new information... but slides will also be posted after class on the course website.
- Homeworks will be posted approximately one week in advance. You do not need to turn these in, but quiz questions will come directly from the homework.
- Lab and discussion attendance is required.

Are there any other questions?

Is this a math class?

Sort of!

- We will certainly use some math and talk about math-related concepts.
- Homeworks and labs may also ask you to do some math
- However, exams will be based on your *conceptual* understanding of the course material.

My Classroom

I have only one formal classroom policy: you can do whatever you want *as long as it doesn't disrupt anyone else's learning*.

A Few Requests

- In-class computer use is fine, but if you are going to be doing anything other than taking notes, please sit in the back so that nobody behind you can get distracted by your screen.
- Use professional language when emailing myself and your TAs. This is a good habit to practice whenever you send an email!
- If you have any problems - with me, the class, other students, your TA - please let me know! Summer sessions cram a lot of material into a short amount of time and I know not all of you want to be here... but I still want this to be the best possible experience for you all.

Student Survey

There is a link to a brief survey on the course website. The survey is **due today at 4pm**. You are encouraged to fill it out during lab.

There are no wrong answers! Everyone who fills out the survey will receive full credit. Points will go toward your lab grade.

Why Study Statistics?

Just a few reasons (of many!)...

1. You can't do scientific research without statistics. Having a solid understanding of statistics will be a huge benefit if you want to apply to Masters or PhD programs.
2. Even if you don't want to get a PhD, many jobs require either some research or the ability to read technical reports.
3. Still not interested? A basic understanding of statistics will make you a more critical consumer of media. In a world of biased or "fake" news, this is a really important skill!

Case Study: Using Stents to Prevent Strokes

A classic challenge in statistics is evaluating the efficacy of medical treatments.

- Stents are medical devices used to assist patients after cardiac events like strokes.
- Suppose we want to know if stents are also beneficial in helping to *prevent* strokes.
- We start by writing our principal question:

  Does the use of stents reduce the risk of stroke?

- Now we can gather data to answer this question.

Case Study

Some researchers conducted a study with 451 at-risk patients. Each patient was randomly assigned to either treatment (preventative stent) or control (no stent).

Case Study

I do research in this area!

- Usually when we test medical treatments, we do a randomized control trial. Basically, we get a sample of the population and randomly assign them to treatment or placebo. Then, we examine the difference in outcomes.
- RCTs have a lot of constraints that exclude certain individuals or occasionally make this kind of experiment unethical.
- Because RCTs are sometimes not an options or certain groups of people are excluded, we might not be able to generalize our results.
- My research focuses on examining ways to use patient characteristics to estimate how these results might generalize.

The Data Matrix

	loan_amount	interest_rate	term	grade	state	total_income	homeownership
1	7500	7.34	36	A	MD	70000	rent
2	25000	9.43	60	B	OH	254000	mortgage
3	14500	6.08	36	A	MO	80000	mortgage
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	3000	7.96	36	A	CA	34000	rent

This **data matrix** shows rows 1, 2, 3, and 50 of a data set on loans.

- Each row represents one loan.
- We call each row a **case** or **observational unit**. We *observe* a number of different characteristics on each *unit*.
- Each column represents some measured characteristic.
- We call these characteristics **variables** because they can *vary* between observations.

Understanding Our Data

	loan_amount	interest_rate	term	grade	state	total_income	homeownership
1	7500	7.34	36	A	MD	70000	rent
2	25000	9.43	60	B	OH	254000	mortgage
3	14500	6.08	36	A	MO	80000	mortgage
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	3000	7.96	36	A	CA	34000	rent

Whenever, we have data, it's important to start by making sure that we understand it.

- What are some questions we might want to ask ourselves about this data set?

Understanding Our Data

Here are a few things I like to consider for all data sets:

- What does each variable represent?
- What are the units?
- Does the data make sense?
  - What if the data showed an interest rate of -999?
  - ...or a state labelled "42"?

Types of Variables

Let's return to our data set:

	loan_amount	interest_rate	term	grade	state	total_income	homeownership
1	7500	7.34	36	A	MD	70000	rent
2	25000	9.43	60	B	OH	254000	mortgage
3	14500	6.08	36	A	MO	80000	mortgage
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	3000	7.96	36	A	CA	34000	rent

Notice that we have some variables made up of letters and some of numbers. This is the basic concept behind variable types.

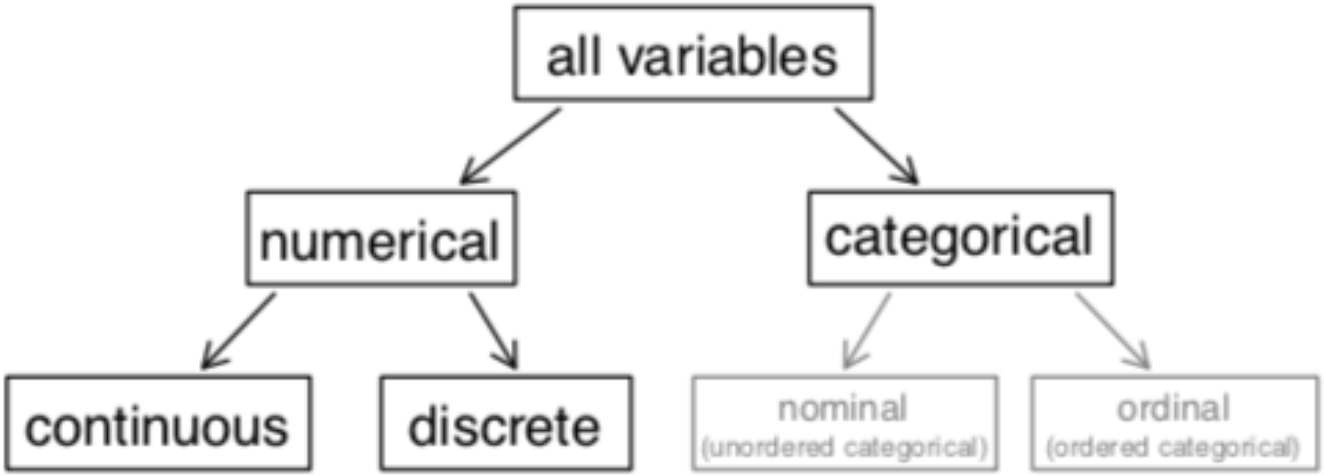
Types of Variables

- **Categorical**
  - The responses are *categories*.
  - The state variable in our data set can take one of 50 possible values.
- **Numeric**
  - The responses are *numeric*.
  - The numbers are meaningful (it makes sense to add, subtract, or take an average using those values).

Types of Numeric Variables

- **Discrete**
  - The responses can take on only whole number values.
  - Population count is a discrete variable.
- **Continuous**
  - The responses can take on values on a continuous scale - there is no jump from one value to the next.
  - Unemployment rate is a continuous variable.

Types of Variables



Note: there are also two types of categorical variables.

- Ordinal variables are ordered (e.g., "like", "neutral", "dislike").
- Nominal variables are unordered (e.g., US states).

Relationships Between Variables

Our brains are constantly working on relationships between variables!

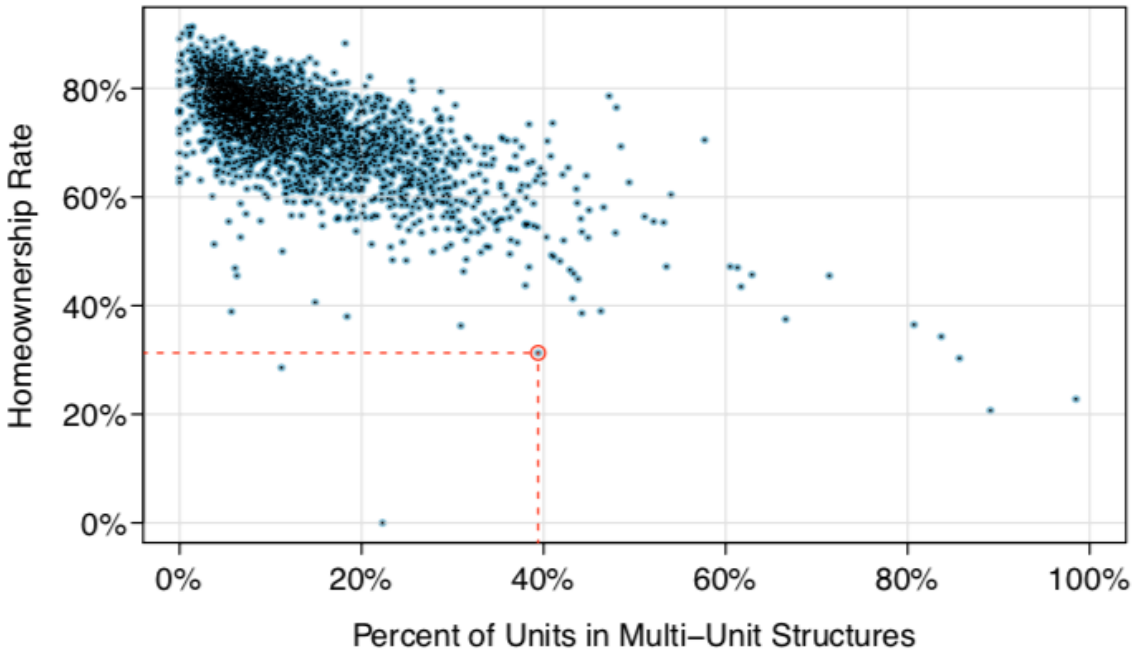
- Imagine if you walked down a flight of stairs outside your apartment 10 times and 9 of those times you fell down the stairs.
- You'd probably decide that something needs to change! Maybe you need to add some traction... or you need an extra cup of coffee before heading out in the morning.
- In statistical terms, you decided that walking down those particular stairs relates to your falling down. You then make adjustments based on that association.

Relationships Between Variables

Statistics takes these kinds of questions about how variables relate to one another (if I go out today, how likely am I to fall down the stairs?) and formalizes them so that we can make sound scientific claims.

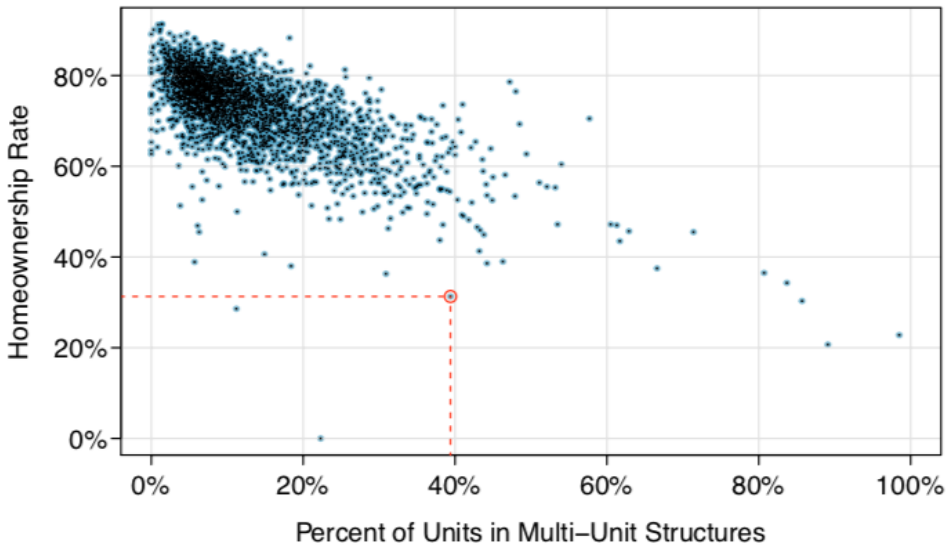
Relationships Between Variables

We can start thinking about how variables relate to one another through data visualization.



Consider the **scatterplot**. Do you think there's a relationship between a county's home ownership rate and its percent of units in multi-unit structures? Why might that be?

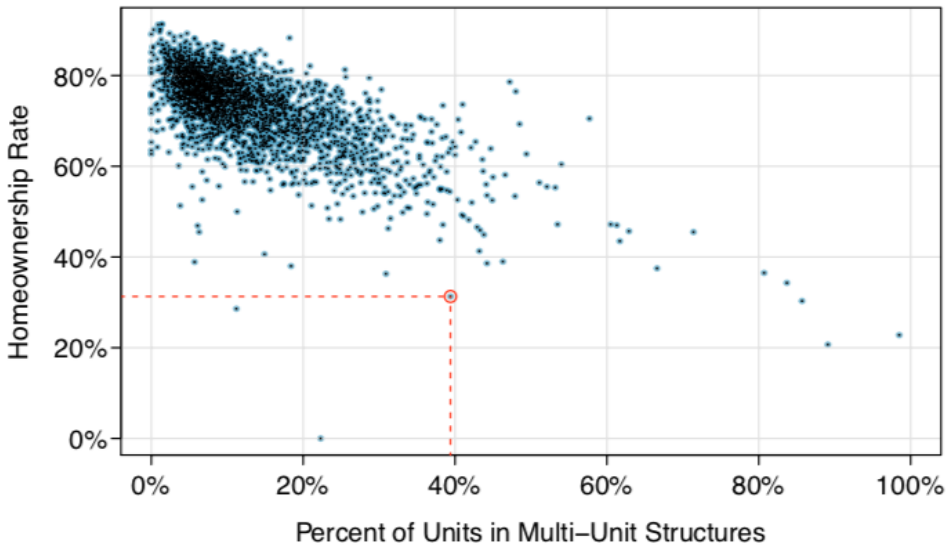
Relationships Between Variables



There is a clear pattern in the plot, so we say that these two variables are **associated**.

- Associated variables *depend* on each other, so we say that they are **dependent variables**.
- If two variables are not associated (no pattern), we say that they are **independent variables**.

Trend



When two variables are related, we can consider the **trend**.

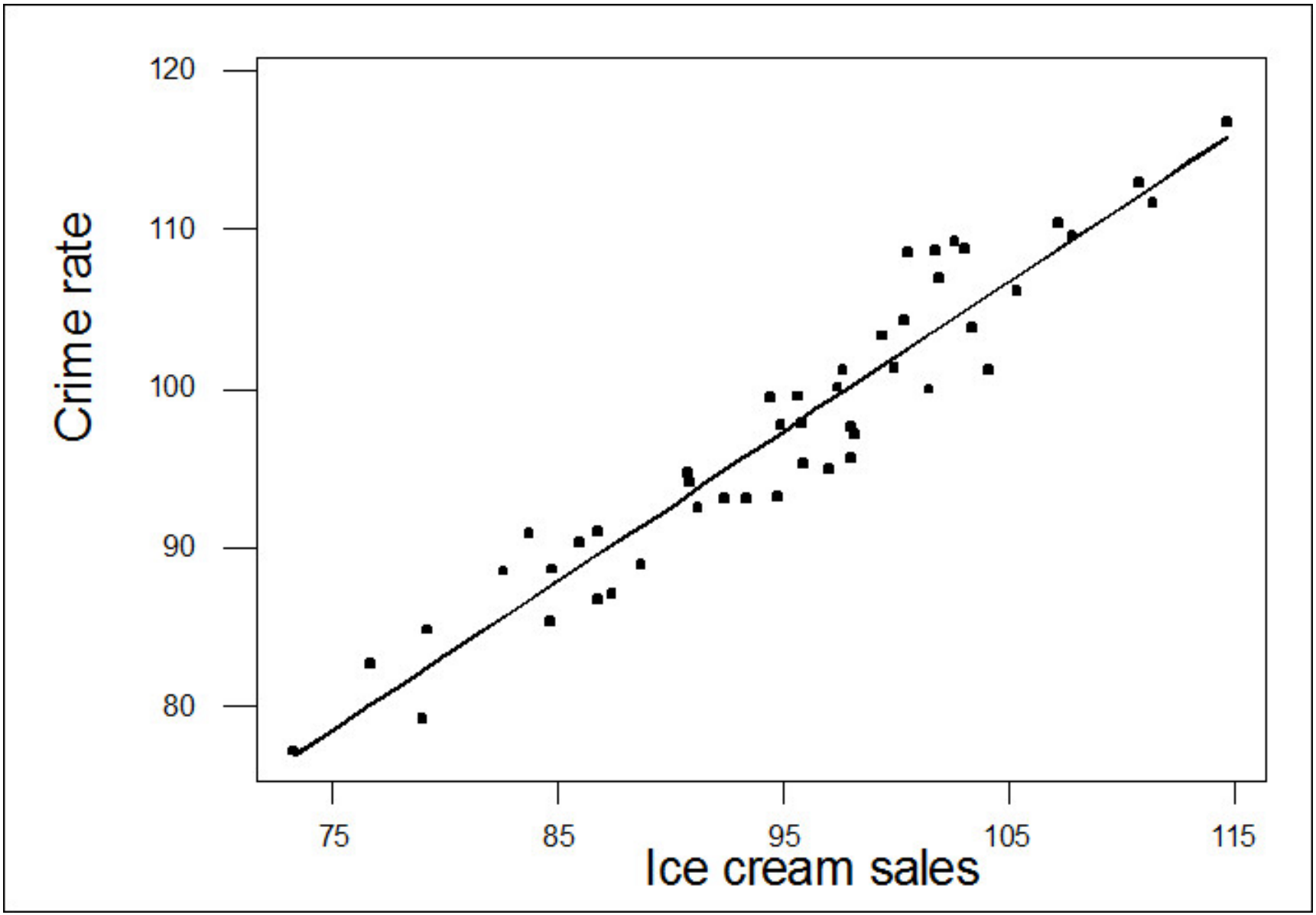
- Here, there is a downward trend, suggesting that these two variables are **negatively associated**.
- When we see an upward trend, we say that the variables are **positively associated**.

A Note on Correlation vs Causation

Who has heard someone say that "correlation is not causation"?

Can you think of an example of two things that correlate but neither one causes the other?

Correlation vs Causation



Explanatory and Response Variables

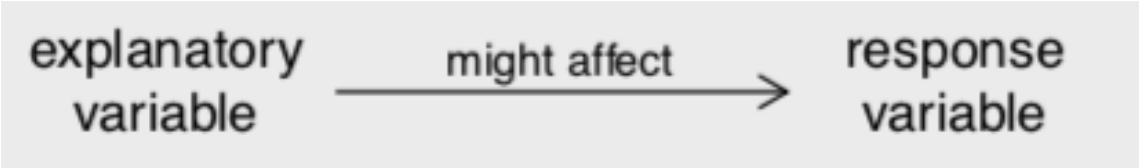
Sometimes we do have causal questions. For example, suppose we have the following question:

If there is an increase in the median household income in a county, does this drive an increase in its population?

- Median household income is an **explanatory variable**.
  - We want to know if increases in median household income *explain* population increases.
- Population increase is a **response variable**.
  - We want to know if the population increases *in response to* increased median household income.

Explanatory and Response Variables

When we predict some causal relationship, we can label our variables accordingly.



However, predicting causality and labeling variables as explanatory and response does *not* guarantee that a causal relationship actually exists.

(Remember our ice cream example - we can be wrong in our predictions but still find an association!)

Conducting Research

Statistics permeates research from start to finish!

- The first step in any experiment is to design it.
- We design an experiment by deciding
  1. What we want to know.
  2. Our target population (who or what we want to know about).
  3. The statistical methods we will use to analyze our data (this helps us decide what kind of data to collect).

Choosing How to Collect Data

A clear, specific research question can go a long way in helping to identify what subjects/cases are important and which variables we should measure.

But we also need to consider *how* these variables are measured.

Research Questions

- Research questions ask a question about some target **population**, which can be made up of anything we are interested in - people, dogs, bicycles, you name it.
- Typically we don't have access to every single person/dog/case in a population, so instead we look at a subset of the population.
- This subset is called a **sample**. It is often a small fraction of the total population.

Research Questions

Let's think about some possible research questions:

1. What is the average mercury content in swordfish in the Atlantic Ocean?
2. Over the last 5 years, what is the average time to complete a degree for UCR undergrads?
3. Does a new drug reduce the number of deaths in patients with severe heart disease?

What makes these research questions clear and specific?

Populations and Samples

What is the average mercury content in swordfish in the Atlantic Ocean?

- What is the target population in this research questions?
- What would represent an individual case?

Populations and Samples

What is the average mercury content in swordfish in the Atlantic Ocean?

- What is the target population in this research questions?
  - All swordfish in the Atlantic Ocean.
- What would represent an individual case?
  - Each individual swordfish in the Atlantic Ocean.

Populations and Samples

Discuss with a neighbor and jot down your thoughts on our other two research question examples. What is the target population in each research question? What represents an individual case?

1. Over the last 5 years, what is the average time to complete a degree for UCR undergrads?
2. Does a new drug reduce the number of deaths in patients with severe heart disease?

Populations and Samples

1. Over the last 5 years, what is the average time to complete a degree for UCR undergrads?
   - Population: all UCR undergrads
   - Individual case: each UCR undergrad

Populations and Samples

2. Does a new drug reduce the number of deaths in patients with severe heart disease?

- Population: all patients with severe heart disease
- Individual case: each patient with severe heart disease

Anecdotal Evidence

Consider the following:

1. I ate Atlantic swordfish and got mercury poisoning, so the mercury levels must really high.
2. I know of two UCR undergrads who took 8 years to graduate, so it must take an unusually long time to graduate from UCR.
3. My dog took a new heart disease drug and hasn't had a heart attack, so it must work.

Anecdotal Evidence

Each claim on the previous slide is based on data! But...

1. The sample sizes are very small!
2. Even if we manage to verify these claims (e.g., the mercury poisoning was actually caused by swordfish), we have no way of knowing if they represent the population well or are extreme cases.
3. We often remember only the extreme cases because they are striking (or possibly due to expectation bias).

Can you think of a time when you heard someone to use **anecdotal evidence** to demonstrate a point?

Sampling from a Population

One we've established our target population and what constitutes an individual case, it's time to think about collecting a sample.

For our question about UCR time-to-graduation, recall that

- our *population* is all UCR undergrads
- and our *sample* is made up of whatever graduated students we selected for review.

Sampling from a Population

What if we took a sample of everyone in an upper division physics course? Are these students likely to be representative of UCR students *on average*?

Sampling from a Population

If we select samples this way, we are likely to get a **biased** sample. In this case, we would *bias* the sample toward however long it takes physics students to graduate.

Sampling from a Population

In order to get a sample that represents UCR students overall, we want to *randomly* sample from our population of all graduated UCR undergrads.

Think of a random sample as a raffle. Each graduated UCR student from the past 5 years get a raffle ticket, and 100 of them are selected randomly. We then ask these 100 students how long they took to graduate.

The Simple Random Sample

The most straightforward way to collect a random sample is the **simple random sample**. This is essentially equivalent to the raffle example:

- Each case in the population has an equal probability of being included in the sample.
- There is no connection between cases in the sample.

Sources of Bias

Bias can occur in simple random samples due to individuals not responding.

If students who took 6 or more years to graduate also happen to be less likely to respond, we may end up with data to suggest that our average time-to-graduation is quicker than it actually is.

Sources of Bias

Usually bias occurs in a sample when we do something out of convenience instead of going through all of the steps to get a truly random sample.

A **convenience sample** is where individuals are sampled because it's easy. E.g., you sample everyone in this class instead of trying to randomly sample from all UCR students.

Bias in the Wild: Amazon Reviews

Suppose you're looking for a new outfit for your pet lizard. You find what appears to be the perfect outfit on Amazon, but the reviews are pretty mixed.

Bias in the Wild: Amazon Reviews

Do you think the reviews give an accurate representation of how buyers feel about this item? When do you think people are more likely to leave reviews?

Observational Studies

Data where no treatment is explicitly applied (or withheld) is **observational data**.

- These kinds of studies are *non-experimental*.
- It is difficult to make causal claims based on observational data.
- Instead, we can use observational studies to
  - form hypotheses for future experiments.
  - demonstrate associations (that may or may not be causal).

Examples of Observational Studies

- Relationship between smoking and lung health.
  - Here, smoking is the "treatment".
  - Ethical constraints prevent us from randomly assigning people to smoking or nonsmoking.
- Relationship between gender and number of pets.
  - Here, gender is the "treatment".
  - We can't impose a gender on someone, so we are unable to randomly assign gender.

Observational Studies

Can you think of something you might be interested in studying where "treatment" can't be explicitly applied?

How do you think we deal with some of the ethical constraints like those in the smoking study?

Why No Causal Claims?

Consider an observational study that tracked sunscreen use versus skin cancer. The study found that the more sunscreen someone used, the more likely they were to develop skin cancer.

- Does sunscreen use cause skin cancer?
- What information could we be missing?

Sunscreen and Skin Cancer

- This missing information consists of variables that we didn't measure. **Confounding variables** are variables that are correlated with (have a relationship with) both the explanatory and response variables.
- Confounding variables can help explain why unexpected relationships occur.
- It is almost impossible to guarantee that we've measured (or even thought of) all confounding variables.
- Note: just because a relationship makes sense, doesn't mean it is causal or that there are no confounding variables!

Types of Observational Studies

We can ask two types of observational questions.

- A **prospective study** collects data as the events unfold.
- A **retrospective study** collects data on events that have already taken place.

Some datasets may include variables taken both prospectively and retrospectively.

Sampling Methods

Simple random sampling is the "raffle method" we talked about earlier. Each case has an equal probability of being selected from the population.

Stratified sampling uses a "divide-and-conquer" approach.

- The population is divided into groups called **strata**, chosen so that similar cases are grouped together.
  - We could group based on variables like year in college, gender, team, etc.
  - We typically choose these grouping based on some variable that we think relates to our outcome.
- We then randomly sample a fixed number of cases from each strata.

Sampling Methods

Cluster sampling involves breaking the population into many groups, called **clusters**.

- We then randomly select some of the clusters and sample all of the cases in each of the selected clusters.

Multistage sampling is similar to cluster sampling, but instead of keeping all cases in each cluster, we randomly sample from each selected cluster (this is the "multistage" part).

Sampling Methods

Cluster and multistage sampling may be cheaper and easier to collect.

- If we wanted to sample individuals from 30 remote villages, it would be cheaper to cluster by village and only travel to 10 of them.

We may also use these approaches when within-cluster variability is high but the clusters are similar *on average*.

- For example, 5 economically diverse neighborhoods with similar average wages in each neighborhood.

The analyses discussed in this class will all pertain to simple random sampling. However, most of these analyses can be extended to work with a variety of sampling methods.