Chapter 1 Introduction to Data

Instructor: Joyce Fu

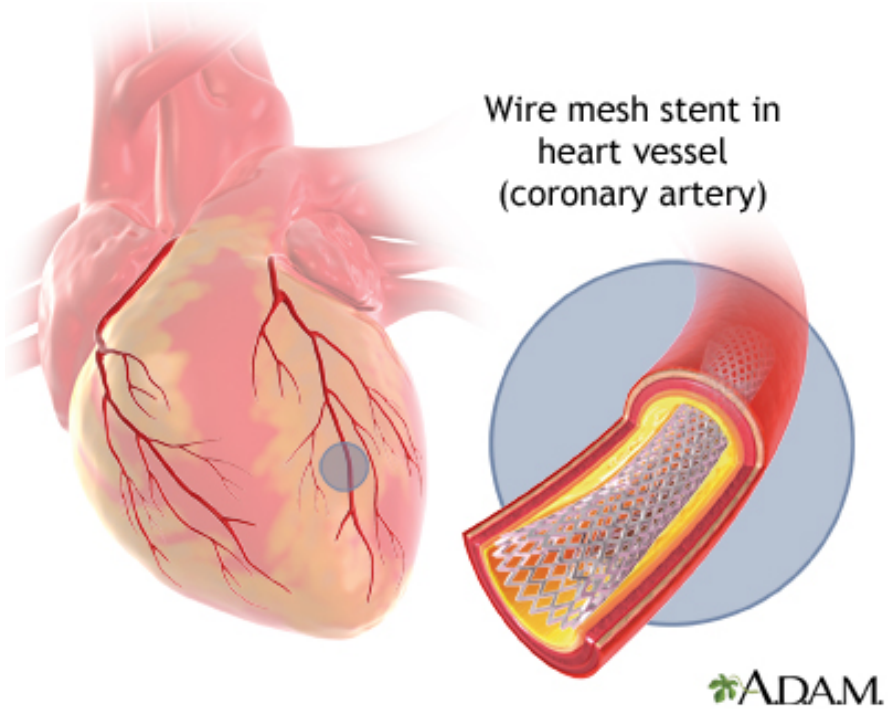
University of California, Riverside
materials adapted from and kindly shared by Lauren Cappiello

January 7, 2020

1.1 Case Study: Using Stents to Prevent Strokes

A classic challenge in statistics is evaluating _____. Here is one promising treatment under study.

- Stents are medical devices used to assist patients after cardiac events like strokes.



1.1 Case Study: Using Stents to Prevent 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 _____ to answer this question.

1.1 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).

1.1 Case Study

Fast forward to the research results

- Proportion who had a stroke in treatment (stent) group : 20%
- Proportion who had a stroke in control (no stent) group : 12%

What conclusion can we draw? Is stent effective? Do we have enough evidence to get stent treatment approved by regulations?
When we are half way through this course, we might be able to answer the questions.

1.2.1 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

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 _____ because they can *vary* between observations.

1.2.1 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?

1.2.1 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"?

1.2.2 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.

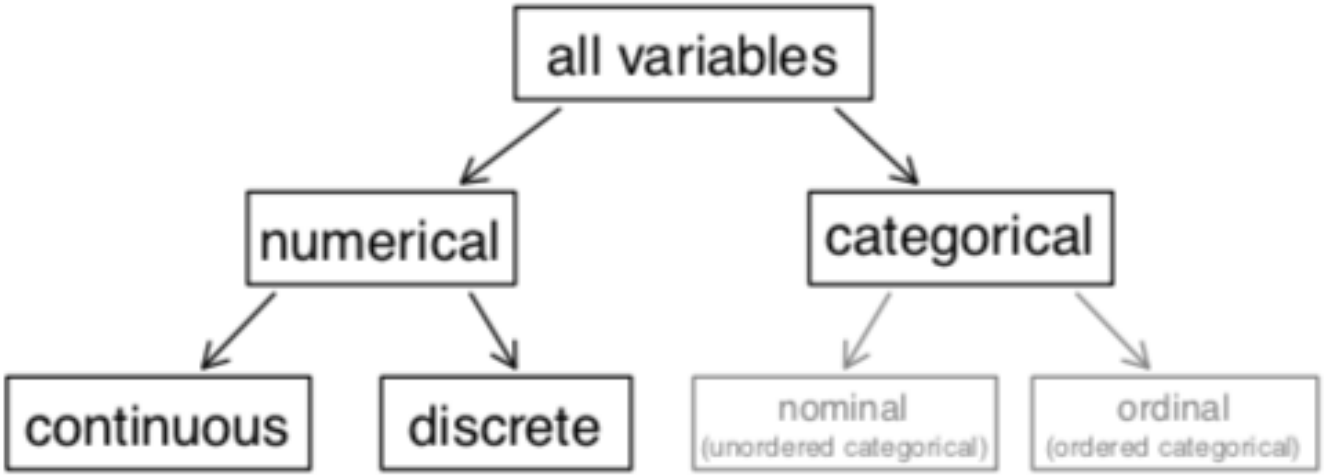
1.2.2 Types of Variables

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

1.2.2 Types of Numeric Variables

- _____
  - The responses can take on only whole number values.
  - Population count is a discrete variable.
- _____
  - 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.

1.2.2 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).

1.2.3 Relationships Between Variables

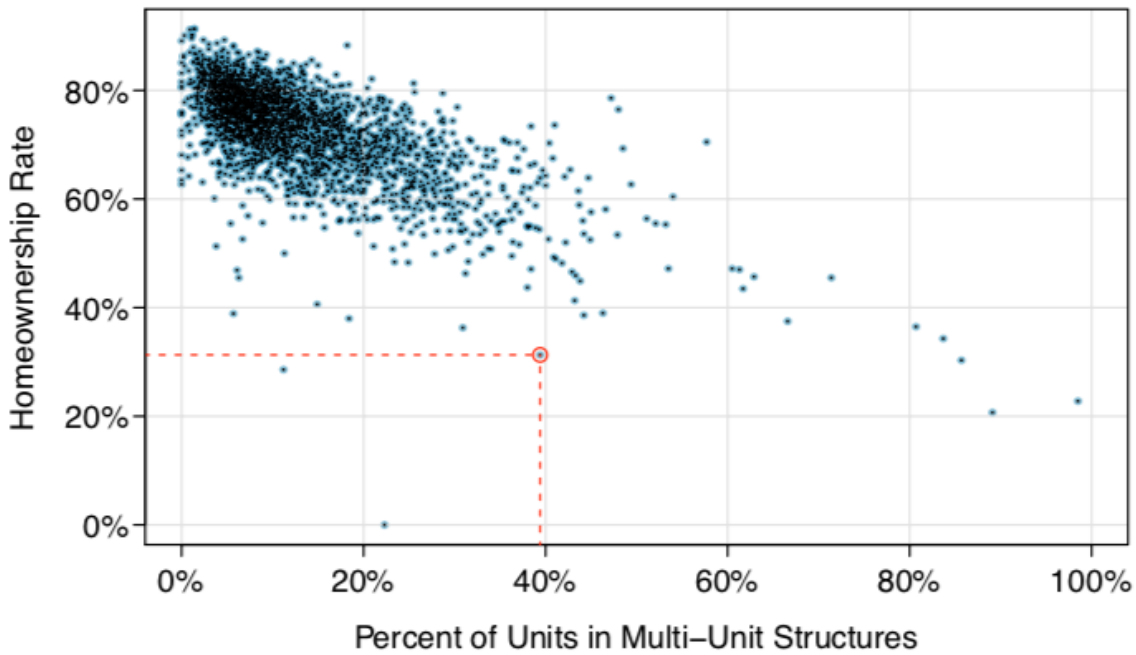
Statistics takes these kinds of questions about how variables relate to one another and formalizes them so that we can make sound scientific claims.

	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

For example, how was the interest rate determined for each case?

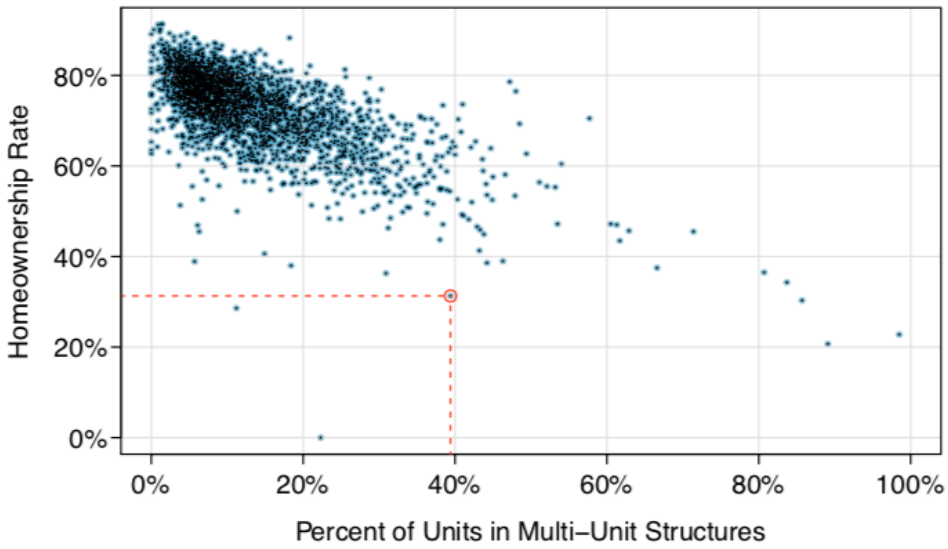
1.2.3 Relationships Between Variables

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?

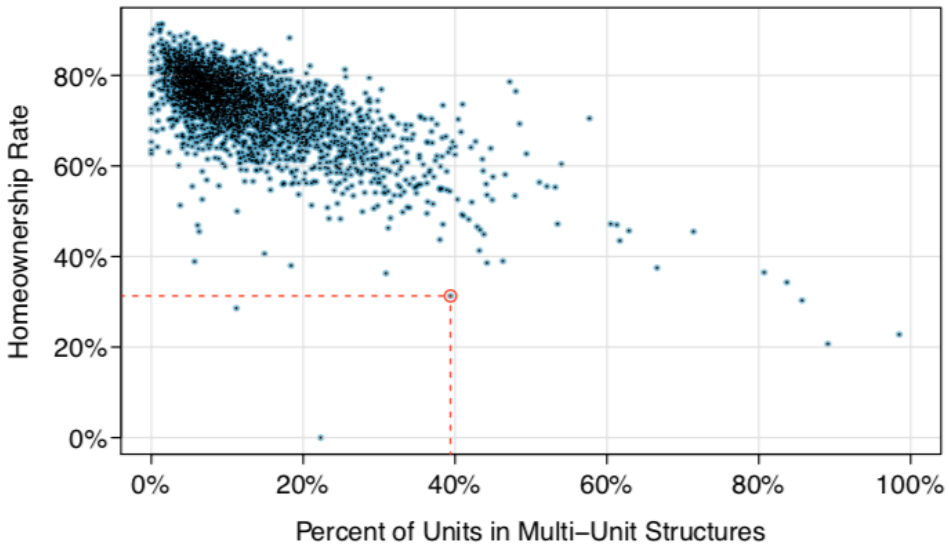
1.2.3 Relationships Between Variables



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

Sometimes we say these two variable are **correlated**. We will talk in more details about correlation in the later part of this course. For now, you don't need to worry about the difference between these two terms.

1.2.3 Trend



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

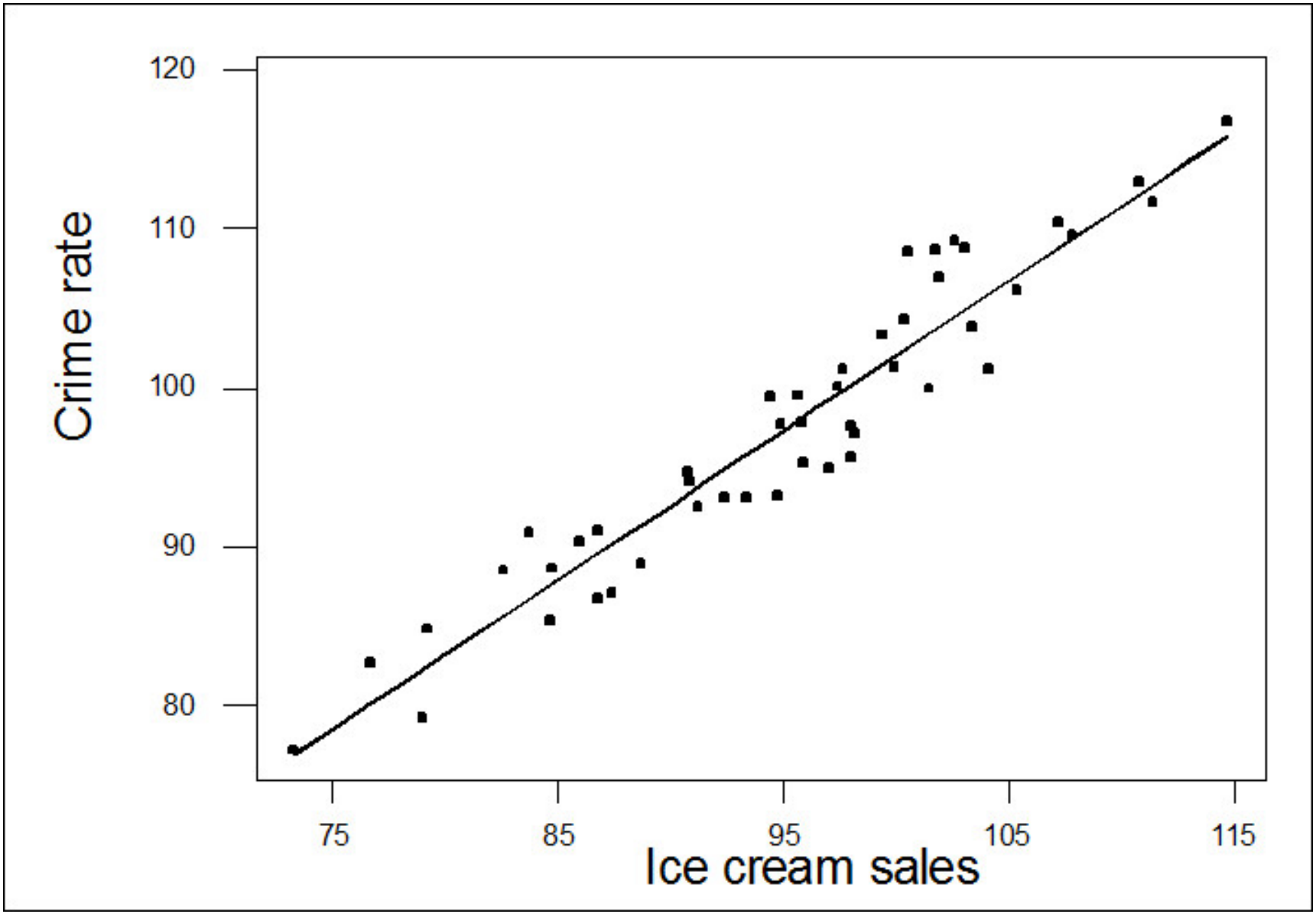
- Here, there is a downward trend, suggesting that these two variables are _____.
- When we see an upward trend, we say that the variables are _____.

1.2.3 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?

1.2.3 Correlation vs Causation



Explanatory and Response Variables

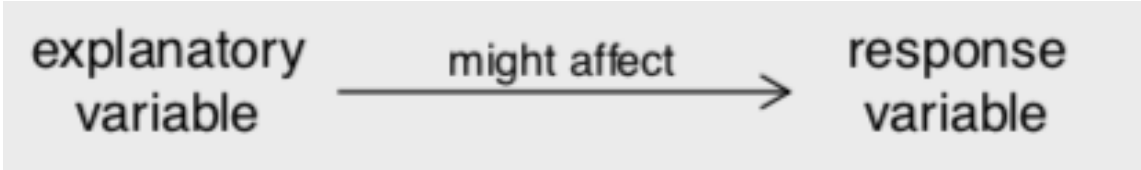
Sometimes we do have causal questions. Suppose have the following question:

> *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 _____.
  - We want to know if increase in household income *explains* population increase
- Population increase is _____..
  - 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.



Example:

- Use total number of customertotal number of customer in Costco per day to predict total revenue per day
- Use Loan amount, household income, loan term to predict loan interest rate.

1.3.1 Some Questions

Let's think about some possible research questions:

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

What makes these research questions clear and specific?

1.3.1 Populations and Samples

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

- What is the target population?
- What represent an individual case?

1.3.1 Populations and Samples

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

- What is the target population?
  -

- What represent an individual case?
  -

1.3.1 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 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?

1.3.1 Populations and Samples

1. Over the last 5 years, what is the average time to complete a degree for UCR undergrads?

- Population:

- Individual case:

1.3.1 Populations and Samples

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

- Population:

- Individual case:

1.3.2 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.

1.3.2 Anecdotal Evidence

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

- Sample sizes:

- Representing the population well? Just extreme cases?
- We often remember only the extreme caseecause 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?

1.3.3 How should we sample?

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

- *population* – all UCR undergrads
- *sample* – graduated students we selected.

1.3 How should we sample?

1.3 How should we sample?

What if sample everyone in an upper division physics course?

1.3.3 How should we sample?

A

---

sample! Only representing physics students.

We *bias* the sample toward however long it takes physics students to graduate.

1.3.3 How should we sample?

Use a raffle. One raffle ticket for each UCR student graduated from the past 5 years. Randomly pick 100 students.

1.3.3 How to Sample

This raffle is actually the

_____.

- Each individual case in the population has an equal probability of being included in the sample.

1.3.3 Sources of Bias

If students who took 6 or more years to graduate decide not to respond. We got biased sample!

Average time to graduate we obtained will be shorter than it actually is.

1.3 Sources of Bias - Convenience Sample

Another Bias: Sample out of convenience instead of going through all of the steps to get a truly random sample.

E.g., you sample everyone in a bookclub to obtain average reading time of all kids.

1.3.3 Bias in the Wild: Amazon Reviews

Suppose you're looking for a new outfit for your pet lizard. Reviews on Amazon are pretty mixed.

1.3.3 Bias in the Wild: Amazon Reviews

When do you think people are more likely to leave reviews?

1.3.5 Sampling Methods

_____, "raffle method" .
Each case has an equal probability of being selected from the population.

_____ uses a "divide-and-conquer" approach.

- Divided population into groups called **strata**, similar cases grouped together.
- Then randomly sample from each strata.

1.3.5 Sampling Methods

_____ breaking the population into many groups, called **clusters**.

- Randomly select some of the clusters
- Sample all of the cases in each of the selected clusters.

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

1.3.5 Sampling Methods

What is the following sampling method?

- We wanted to sample individuals from 30 remote villages.
- We randomly choose 5 fo them and sample all villagers within each.

- We wanted to sample individuals from 30 remote villages.
- We randomly choose 5 fo them and then randomly choose 10 within each.

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.