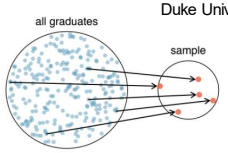
Unit 1: Introduction to data

1. Data Collection + Observational studies & experiments

Sta 101 – Spring 19

Duke University, Department of Statistical Science



Dr. Ellison

Slides posted at
https://www2.stat.duke.edu/courses/Spring19/sta101.001/

In-Class Lecture Notes Key

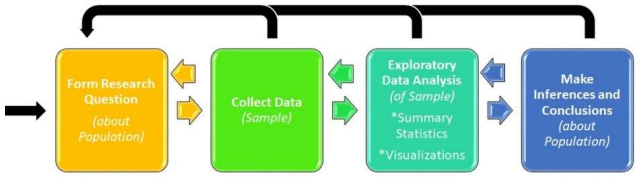
- Extra Focus
- Concept from Different Angle
- NEW Completely New Concept
- Inner Workings/Why does this work?
- Relationships Between Concepts
- Building Intuition
- Hands-On Exercises
- Common Misconceptions
- Coding
- Game

Outline

1. Main ideas
   - **A. Course Goal:** NEW Use a sample (data) to make inferences about the population
   - **B. Data Collection Principles**
     1. Random Sampling:
        - i. NEW Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier
        - ii. NEW Sampling schemes can suffer from a variety of biases
     2. Random Assignment: Randomly assign observations to each independent variable group.
        1. NEW Four principles of experimental design: randomize, control, block, replicate
   - **C. Types of Studies:** NEW Experiments use random assignment to treatment groups observational studies do not
   - **D. Types of Inferences we can Make and How:**
     1. NEW Random sampling helps generalizability,
     2. NEW Random assignment helps causality (two or more variables)
2. Summary

NEW 1. Use a sample to make inferences about the population

- **Ultimate goal**: make inferences about populations

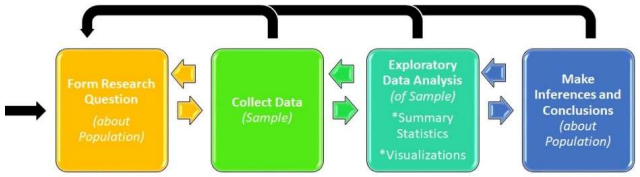


1

NEW

1. Use a sample to make inferences about the population

- **Ultimate goal**: make inferences about populations
- **Caveat**: populations are difficult or impossible to access

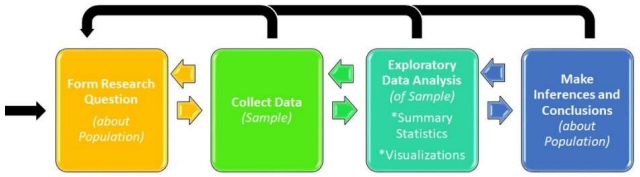


1

NEW

1. Use a sample to make inferences about the population

- **Ultimate goal**: make inferences about populations
- **Caveat**: populations are difficult or impossible to access
- **Solution**: use a sample from that population, and use *statistics* from that sample to make inferences about the unknown population *parameters*

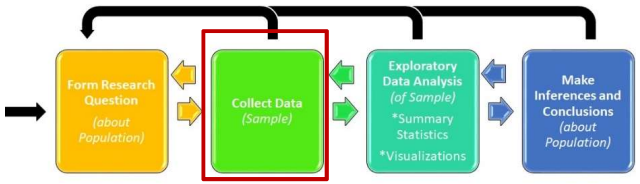


1

NEW

1. Use a sample to make inferences about the population

- **Ultimate goal**: make inferences about populations
- **Caveat**: populations are difficult or impossible to access
- **Solution**: use a sample from that population, and use *statistics* from that sample to make inferences about the unknown population *parameters*
- The better (more *representative*) sample we have, the more reliable our estimates and more accurate our inferences will be



1

NEW

1. Use a sample to make inferences about the population

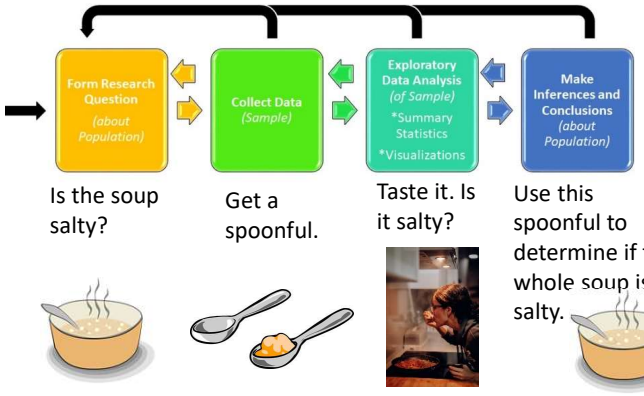
- Ultimate goal: make inferences about populations
- Caveat: populations are difficult or impossible to access
- Solution: use a sample from that population, and use *statistics* from that sample to make inferences about the unknown population *parameters*
- The better (more *representative*) sample we have, the more reliable our estimates and more accurate our inferences will be

Suppose we want to know how many offspring female lemurs have, on average. It's not feasible to obtain offspring data from on all female lemurs, so we use data from the Duke Lemur Center. We use the sample mean from these data as an estimate for the unknown population mean. Can you see any limitations to using data from the Duke Lemur Center to make inferences about all lemurs?

1

NEW

Sampling is natural



2

NEW

Sampling is natural

- When you taste a spoonful of soup and decide the spoonful you tasted isn't salty enough, that's *exploratory analysis*
- If you generalize and conclude that your entire soup needs salt, that's an *inference*
- For your inference to be valid, the spoonful you tasted (the sample) needs to be *representative* of the entire pot (the population)

2

NEW

Outline

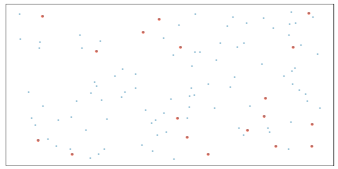
1. Main ideas
   - **A. Course Goal:** NEW Use a sample (data) to make inferences about the population
   - **B. Data Collection Principles**
     1. Random Sampling:
        - i. NEW Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier
        - ii. NEW Sampling schemes can suffer from a variety of biases
     2. Random Assignment: Randomly assign observations to each independent variable group.
        1. NEW Four principles of experimental design: randomize, control, block, replicate
   - **C. Types of Studies**: NEW Experiments use random assignment to treatment groups, observational studies do not
   - **D. Types of Inferences we can Make and How:**
     1. NEW Random sampling helps generalizability,
     2. NEW Random assignment helps causality (two or more variables)
2. Summary

2. Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier

NEW

Simple random:

Drawing names from a hat



3

2. Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier

NEW

Assess: We want to know how many hours the average adult American spends watching TV. We have log data from TV set-top-boxes. However, adults age groups 18-29, 30-45, and 55+ often have different TV viewing habits and 18-29 adults don't often have TV set-top-boxes.

- **What is the population of interest in this example?**
- **What are some potential issues that might be encountered in conducting a simple random sample of this population?**

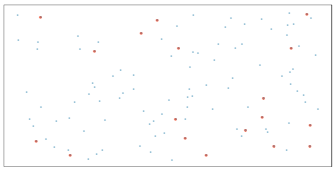
3

2. Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier

NEW

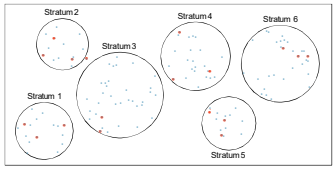
Simple random:

Drawing names from a hat



Stratified: homogenous strata

Stratify to control for age group



3

2. Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier

Assess: There exist 57 distinct colonies of chimpanzees spread across a large country. Researcher wish to assess the average gestation period of female chimpanzees in the country. Each of these colonies are spread out across the country and are difficult to find, but the chimpanzees in each colony are all interspersed with young, old, sick, and healthy chimpanzees.

- **What is the population of interest in this example?**
- **What are some potential issues that might be encountered in conducting a simple random sample of this population?**

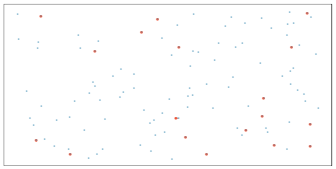
3

2. Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier

NEW

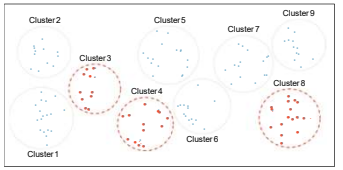
Simple random:

Drawing names from a hat



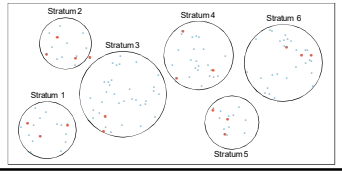
Cluster: heterogenous clusters

Sample all chosen clusters



Stratified: homogenous strata

Stratify to control for age group

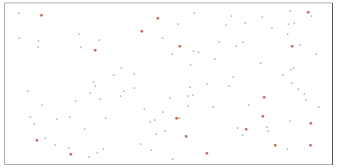


3

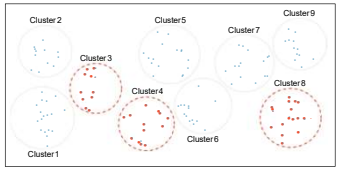
2. Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier

NEW

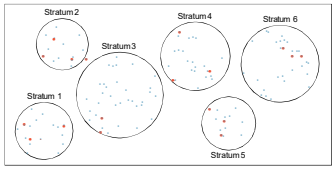
Simple random:
Drawing names from a hat



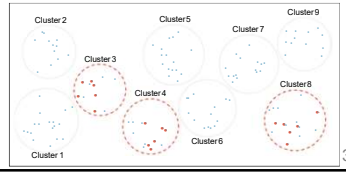
Cluster: heterogenous clusters
Sample all chosen clusters



Stratified: homogenous strata
Stratify to control for age group



Multistage: heterogeneous clusters
Random sample in chosen clusters



3

2. Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier

NEW

	Sampling Method Name		
	Stratified Sampling	**Cluster Sampling**	**Multistage Sampling**
How to sample this way.	Step 1: Assign each observation in a population into groups. Each group is called a...		
	Stratum (Strata plural),	Cluster,	Cluster,
	where the objects in a given group are...		
	homogeneous.	heterogeneous.	heterogeneous.
	Step 2:Then, select...		
	ALL the groups.	a random selection of groups	a random selection of groups
	Step 3: Then from each of these selected groups, select...		
	a random sample of observations	ALL the observations	a random sample of observations

Clicker question

NEW

A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes and some with only apartments. Which approach would likely be the least effective?

(a) Simple random sampling
(b) Stratified sampling, where each stratum is a neighborhood
(c) Cluster sampling, where each cluster is a neighborhood

4

Clicker question

NEW

A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes and some with only apartments. Which approach would likely be the least effective?

(a) Simple random sampling
(b) Stratified sampling, where each stratum is a neighborhood
(c) Cluster sampling, where each cluster is a neighborhood

4

Clicker question

A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes and some with only apartments. Which approach would likely be the least effective?

(a) Simple random sampling

(b) Stratified sampling, where each stratum is a neighborhood

(c) Cluster sampling, where each cluster is a neighborhood

- *Groups = neighborhoods*
- *Neighborhoods are unique => homogeneous within*
- *Clusters (in cluster sampling) need to be heterogeneous.*

4

Outline

Random sampling HELPS generalize sample results to the population, what types of sampling make it HARDER to generalize to the population?

Outline

1. Main ideas
   - **A. Course Goal:** NEW Use a sample (data) to make inferences about the population
   - **B. Data Collection Principles**
     1. Random Sampling:
        - i. NEW Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier
        - ii. NEW Sampling schemes can suffer from a variety of biases
     2. Random Assignment: Randomly assign observations to each independent variable group.
        1. NEW Four principles of experimental design: randomize, control, block, replicate
   - **C. Types of Studies:** NEW Experiments use random assignment to treatment groups, observational studies do not
   - **D. Types of Inferences we can Make and How:**
     1. NEW Random sampling helps generalizability,
     2. NEW Random assignment helps causality (two or more variables)
2. Summary

3. Sampling schemes can suffer from a variety of biases NEW

Assess: Out of a list of 30 political issues (ex: abortion, healthcare) what issue do Americans most desire to hear discussed in the 2016 presidential debate?

Some sample ideas… how could these samples be NOT REPRESENTATIVE of the population?

- Poll everyone in this STA101 class?
- Randomly select 300 addresses, send them a poll, and request that they return it?
- Open up a national online poll?

1

3. Sampling schemes can suffer from a variety of biases

NEW

- *Non-response:* If only a non-random fraction of the randomly sampled people choose to respond to a survey, the sample may no longer be representative of the population

5

3. Sampling schemes can suffer from a variety of biases

NEW

Assess: Out of a list of 30 political issues (ex: abortion, healthcare) what issue do Americans most desire to hear discussed in the 2016 presidential debate?

Some sample ideas… what type of biases do these sampling methods suffer from?

- Poll everyone in this STA101 class.
- Randomly select 300 addresses, send them a poll, and request that they return it
- Open up a national online poll.

1

3. Sampling schemes can suffer from a variety of biases

NEW

Assess: Out of a list of 30 political issues (ex: abortion, healthcare) what issue do Americans most desire to hear discussed in the 2016 presidential debate?

Some sample ideas… what type of biases do these sampling methods suffer from?

- Poll everyone in this STA101 class.
- Randomly select 300 addresses, send them a poll, and request that they return it. *(Non-response bias)*
- Open up a national online poll.

1

3. Sampling schemes can suffer from a variety of biases

NEW

- *Non-response:* If only a non-random fraction of the randomly sampled people choose to respond to a survey, the sample may no longer be representative of the population
- *Voluntary response:* Occurs when the sample consists of people who volunteer to respond because they have strong opinions on the issue since such a sample will also not be representative of the population

5

3. Sampling schemes can suffer from a variety of biases

NEW

Assess: Out of a list of 30 political issues (ex: abortion, healthcare) what issue do Americans most desire to hear discussed in the 2016 presidential debate?

Some sample ideas… what type of biases do these sampling methods suffer from?

- Poll everyone in this STA101 class.
- Randomly select 300 addresses, send them a poll, and request that they return it. *(Non-response bias)*
- Open up a national online poll.

1

3. Sampling schemes can suffer from a variety of biases

NEW

Assess: Out of a list of 30 political issues (ex: abortion, healthcare) what issue do Americans most desire to hear discussed in the 2016 presidential debate?

Some sample ideas… what type of biases do these sampling methods suffer from?

- Poll everyone in this STA101 class
- Randomly select 300 addresses, send them a poll, and request that they return it *(Non-response bias)*
- Open up a national online poll *(Voluntary response bias)*

1

3. Sampling schemes can suffer from a variety of biases

NEW

- *Non-response:* If only a non-random fraction of the randomly sampled people choose to respond to a survey such that the sample may be no longer be representative of the population
- *Voluntary response:* Occurs when the sample consists of people who volunteer to respond because they have strong opinions on the issue since such a sample will also not be representative of the population
- *Convenience sample:* Individuals who are easily accessible are more likely to be included in the sample

5

3. Sampling schemes can suffer from a variety of biases

NEW

Assess: Out of a list of 30 political issues (ex: abortion, healthcare) what issue do Americans most desire to hear discussed in the 2016 presidential debate?

Some sample ideas… what type of biases do these sampling methods suffer from?

- Poll everyone in this STA101 class
- Randomly select 300 addresses, send them a poll, and request that they return it *(Non-response bias)*
- Open up a national online poll *(Voluntary response bias)*

1

3. Sampling schemes can suffer from a variety of biases

NEW

Assess: Out of a list of 30 political issues (ex: abortion, healthcare) what issue do Americans most desire to hear discussed in the 2016 presidential debate?

Some sample ideas… what type of biases do these sampling methods suffer from?

- Poll everyone in this STA101 class? *(convenience sample bias)*
- Randomly select 300 addresses, send them a poll, and request that they return it? *(Non-response bias)*
- Open up a national online poll? *(voluntary response bias)*

1

NEW

Clicker question

A school district is considering whether it will no longer allow high school students to park at school after two recent accidents where students were severely injured. As a first step, they survey parents by mail, asking them whether or not the parents would object to this policy change. Of 6,000 surveys that go out, 1,200 are returned. Of these 1,200 surveys that were completed, 960 agreed with the policy change and 240 disagreed. Which of the following statements are true?

I. Some of the mailings may have never reached the parents.
II. Overall, the school district has strong support from parents to move forward with the policy approval.
III. It is possible that majority of the parents of high school students disagree with the policy change.
IV. The survey results are unlikely to be biased because all parents were mailed a survey.

(a) Only I (b) I and II (c) I and III (d) III and IV (e) Only IV

6

NEW

Clicker question

A school district is considering whether it will no longer allow high school students to park at school after two recent accidents where students were severely injured. As a first step, they survey parents by mail, asking them whether or not the parents would object to this policy change. Of 6,000 surveys that go out, 1,200 are returned. Of these 1,200 surveys that were completed, 960 agreed with the policy change and 240 disagreed. Which of the following statements are true?

I. Some of the mailings may have never reached the parents.
II. Overall, the school district has strong support from parents to move forward with the policy approval.
III. It is possible that majority of the parents of high school students disagree with the policy change.
IV. The survey results are unlikely to be biased because all parents were mailed a survey.

(a) Only I (b) I and II (c) *I and III* (d) III and IV (e) Only IV

6

NEW

Clicker question

A school district is considering whether it will no longer allow high school students to park at school after two recent accidents where students were severely injured. As a first step, they survey parents by mail, asking them whether or not the parents would object to this policy change. Of 6,000 surveys that go out, 1,200 are returned. Of these 1,200 surveys that were completed, 960 agreed with the policy change and 240 disagreed. Which of the following statements are true?

I. Some of the mailings may have never reached the parents. (Convenience sample)
II. Overall, the school district has strong support from parents to move forward with the policy approval.
III. It is possible that majority of the parents of high school students disagree with the policy change. (Non-response bias)
IV. The survey results are unlikely to be biased because all parents were mailed a survey.

(a) Only I (b) I and II (c) *I and III* (d) III and IV (e) Only IV

6

Outline

1. Main ideas
   - **A. Course Goal:** NEW Use a sample (data) to make inferences about the population
   - **B. Data Collection Principles**
     1. Random Sampling:
        - i. NEW Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier
        - ii. NEW Sampling schemes can suffer from a variety of biases
     2. Random Assignment: Randomly assign observations to each independent variable group.
        1. NEW Four principles of experimental design: randomize, control, block, replicate
   - **C. Types of Studies**: NEW Experiments use random assignment to treatment groups, observational studies do not
   - **D. Types of Inferences we can Make and How:**
     1. NEW Random sampling helps generalizability,
     2. NEW Random assignment helps causality (two or more variables)
2. Summary

Inferences about Relationships between two or more Variables:

"Is there an association between age and political affiliation?"

Vs.

"Does age affect/cause political affiliation?" or "Does being older cause you vote Republican?"

Sample Data

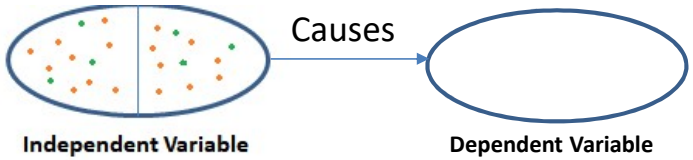
Gender (Categorical with 2 Levels)	Political Affiliation (Categorical with 3 Levels)	Voted in 2018? (Categorical with 2 Levels)	Age (Numerical - Continuous)	Income (Numerical - Continuous)	Number of Dependents (Numerical - Discrete)
Male	Democrat	Yes	22	$20,000	0
Female	Republican	Yes	39	$40,000	3
Male	Independent	Yes	47	$400,000	0
Male	Republican	Yes	18	$129,000	0
Female	Republican	No	29	$85,000	2
Female	Democrat	No	80	$72,000	1
Female	Democrat	Yes	56	$55,000	0
Male	Independent	Yes	72	$34,000	0
...	...	...	...	...	...

Outline

1. Main ideas
   - **A. Course Goal:** NEW Use a sample (data) to make inferences about the population
   - **B. Data Collection Principles**
     1. Random Sampling:
        - i. NEW Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier
        - ii. NEW Sampling schemes can suffer from a variety of biases
     2. Random Assignment: Randomly assign observations to each independent variable group.
        1. NEW Four principles of experimental design: randomize, control, block, replicate
   - **C. Types of Studies**: NEW Experiments use random assignment to treatment groups, observational studies do not
   - **D. Types of Inferences we can Make and How:**
     1. NEW Random sampling helps generalizability,
     2. NEW Random assignment helps causality (two or more variables)
2. Summary

Outline

Why does random assignment allow for us to conclude the independent variable causes/influences the dependent variable?



What type of study is this? What is the scope of inference (causality / generalizability)?

Facebook Tinkers With Users' Emotions in News Feed Experiment, Stirring Outcry

By VINDU GOEL JUNE 29, 2014 The New York Times

In an academic paper published in conjunction with two university researchers, the company reported that, for one week in January 2012, it had altered the number of positive and negative posts in the news feeds of 689,003 randomly selected users to see what effect the changes had on the tone of the posts the recipients then wrote.

The researchers found that moods were contagious. The people who saw more positive posts responded by writing more positive posts. Similarly, seeing more negative content prompted the viewers to be more negative in their own posts.

Were the users allowed to assign themselves to the positive or negative newsfeed groups or were they randomly assigned?

http://www.nytimes.com/2014/06/30/technology/facebook-tinkers-with-users-emotions-in-news-feed-experiment-stirring-outcry.html

7

NEW

What type of study is this? What is the scope of inference (causality / generalizability)?

Facebook Tinkers With Users' Emotions in News Feed Experiment, Stirring Outcry

By VINDU GOEL JUNE 29, 2014 The New York Times

In an academic paper published in conjunction with two university researchers, the company reported that, for one week in January 2012, it had altered the number of positive and negative posts in the news feeds of 689,003 randomly selected users to see what effect the changes had on the tone of the posts the recipients then wrote.

The researchers found that moods were contagious. The people who saw more positive posts responded by writing more positive posts. Similarly, seeing more negative content prompted the viewers to be more negative in their own posts.

Were the users allowed to assign themselves to the positive or negative newsfeed groups or were they randomly assigned?

→Study is an Experiment, not an observational study!

http://www.nytimes.com/2014/06/30/technology/facebook-tinkers-with-users-emotions-in-news-feed-experiment-stirring-outcry.html

7

NEW

What type of study is this? What is the scope of inference (causality / generalizability)?

Facebook Tinkers With Users' Emotions in News Feed Experiment, Stirring Outcry

By VINDU GOEL JUNE 29, 2014 The New York Times

In an academic paper published in conjunction with two university researchers, the company reported that, for one week in January 2012, it had altered the number of positive and negative posts in the news feeds of 689,003 randomly selected users to see what effect the changes had on the tone of the posts the recipients then wrote.

The researchers found that moods were contagious. The people who saw more positive posts responded by writing more positive posts. Similarly, seeing more negative content prompted the viewers to be more negative in their own posts.

What is the scope of inference:

- If it's a Random Experiment:
  - "Does newsfeed positivity *affect/cause* posting positivity?"
- If it's an Observational Study
  - "Is there an *association* between newsfeed positivity and posting positivity?"

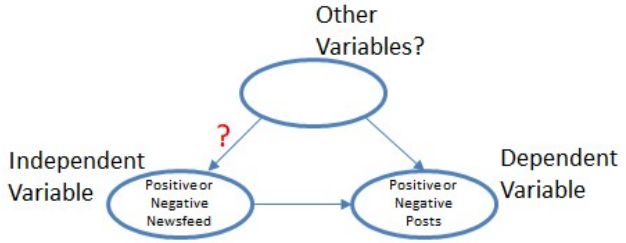
http://www.nytimes.com/2014/06/30/technology/facebook-tinkers-with-users-emotions-in-news-feed-experiment-stirring-outcry.html

7

NEW

What type of study is this? What is the scope of inference (causality / generalizability)?

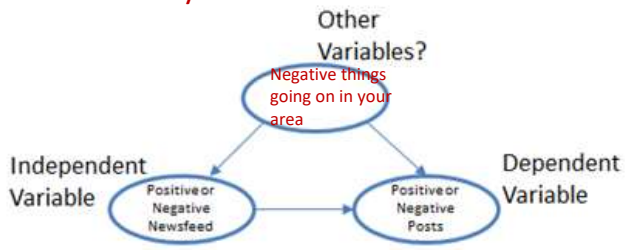
Experiment:
Facebook Study



7

What type of study is this? What is the scope of inference (causality / generalizability)?

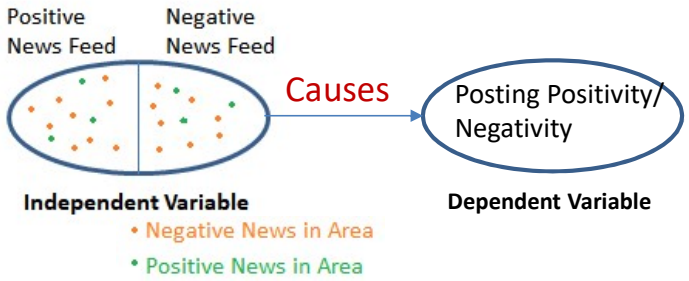
Experiment:
Facebook Study



7

What type of study is this? What is the scope of inference (causality / generalizability)?

Experiment:
Facebook Study



7

4. Experiments use random assignment to treatment groups, observational studies do not

A study that surveyed a random sample of otherwise healthy adults found that people are more likely to get muscle cramps when they're stressed. The study also noted that people drink more coffee and sleep less when they're stressed. What type of study is this?

What is the conclusion of the study?

Can this study be used to conclude a causal relationship between increased stress and muscle cramps?

8

4. Experiments use random assignment to treatment groups, observational studies do not

NEW

A study that surveyed a random sample of otherwise healthy adults found that people are more likely to get muscle cramps when they're stressed. The study also noted that people drink more coffee and sleep less when they're stressed. What type of study is this?

Observational

Independent Variable

Dependent Variable

What is the conclusion of the study?

Can this study be used to conclude a causal relationship between increased stress and muscle cramps?

8

4. Experiments use random assignment to treatment groups, observational studies do not

NEW

A study that surveyed a random sample of otherwise healthy adults found that people are more likely to get muscle cramps when they're stressed. The study also noted that people drink more coffee and sleep less when they're stressed. What type of study is this?

Observational

What is the conclusion of the study?

There is an association between increased stress & muscle cramps.

Can this study be used to conclude a causal relationship between increased stress and muscle cramps?

8

4. Experiments use random assignment to treatment groups, observational studies do not

NEW

Can this study be used to conclude a causal relationship between increased stress and muscle cramps?

Muscle cramps might also be due to increased caffeine consumption or sleeping less – these are potential confounding variables.

extraneous variables that affect both the independent and the dependent variable and that make it seem like there's a relationship between them.

High Stress Low Stress

Causes?

Muscle Cramps

Independent Variable

Dependent Variable

8

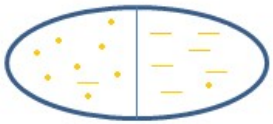
4. Experiments use random assignment to treatment groups, observational studies do not

NEW

Can this study be used to conclude a causal relationship between increased stress and muscle cramps?

Muscle cramps might also be due to increased caffeine consumption or sleeping less – these are potential confounding variables.

High Stress Low Stress



Causes?

Muscle Cramps

Independent Variable

Dependent Variable

• High Caffeine
— Low Caffeine

8

Outline

What are the four principles of an experimental design?

Outline

1. Main ideas
   - **A. Course Goal:** NEW Use a sample (data) to make inferences about the population
   - **B. Data Collection Principles**
     1. Random Sampling:
        - i. NEW Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier
        - ii. NEW Sampling schemes can suffer from a variety of biases
     2. Random Assignment: Randomly assign observations to each independent variable group.
        1. NEW Four principles of experimental design: randomize, control, block, replicate
   - C. **Types of Studies**: NEW Experiments use random assignment to treatment groups, observational studies do not
   - **D. Types of Inferences we can Make and How:**
     1. NEW Random sampling helps generalizability,
     2. NEW Random assignment helps causality (two or more variables)
2. Summary

5. Four principles of experimental design: randomize, control, block, replicate

NEW

- We would like to design an experiment to investigate if increased stress causes muscle cramps, *randomly assign* subjects to either:
  - Treatment: increased stress
  - Control: no or baseline stress

9

5. Four principles of experimental design: randomize, control, block, replicate

NEW

- We would like to design an experiment to investigate if increased stress causes muscle cramps, *randomly assign* subjects to either:
  - Treatment: increased stress
  - Control: no or baseline stress
- It is suspected that the effect of stress might be different on younger and older people: *block* for age.

Why is this important? Can you think of other variables to block for?

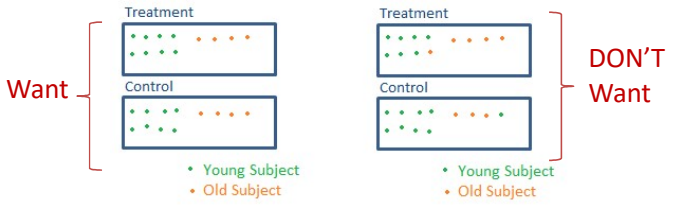
9

5. Four principles of experimental design: randomize, control, block, replicate

NEW

- We would like to design an experiment to investigate if increased stress causes muscle cramps, *randomly assign* subjects to either:
  - Treatment: increased stress
  - Control: no or baseline stress
- It is suspected that the effect of stress might be different on younger and older people: *block* for age.

Why is this important? Can you think of other variables to block for?



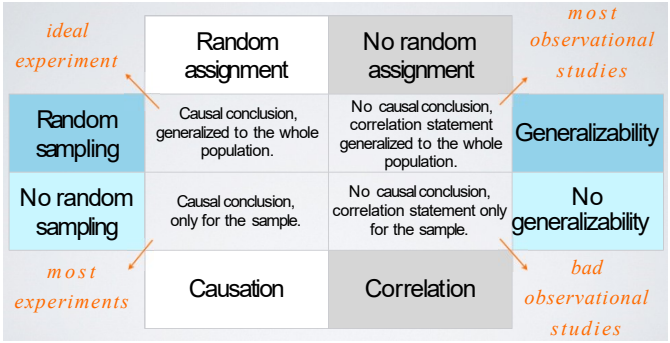
9

Outline

1. Main ideas
   - **A. Course Goal:** NEW Use a sample (data) to make inferences about the population
   - **B. Data Collection Principles**
     1. Random Sampling:
        - i. NEW Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier
        - ii. NEW Sampling schemes can suffer from a variety of biases
     2. Random Assignment: Randomly assign observations to each independent variable group.
        1. NEW Four principles of experimental design: randomize, control, block, replicate
   - **C. Types of Studies**: NEW Experiments use random assignment to treatment groups, observational studies do not
   - **D. Types of Inferences we can Make and How:**
     1. NEW Random sampling helps generalizability,
     2. NEW Random assignment helps causality (two or more variables)
2. Summary

6. Random sampling helps generalizability, random assignment helps causality (two or more variables) NEW

If two or more variables in research question



10

Outline

1. Main ideas
   - **A. Course Goal:** NEW Use a sample (data) to make inferences about the population
   - **B. Data Collection Principles**
     1. Random Sampling:
        - i. NEW Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier
        - ii. NEW Sampling schemes can suffer from a variety of biases
     2. Random Assignment: Randomly assign observations to each independent variable group.
        1. NEW Four principles of experimental design: randomize, control, block, replicate
   - **C. Types of Studies**: NEW Experiments use random assignment to treatment groups, observational studies do not
   - **D. Types of Inferences we can Make and How:**
     1. NEW Random sampling helps generalizability,
     2. NEW Random assignment helps causality (two or more variables)
2. Summary

TO-DO:

1. Get clickers if you haven't already and bring to class on Monday.
2. Download or purchase book.
3. Complete the Pre-test and Getting to Know you Survey.
4. Watch Unit 1 videos

Thursday:

1. Go to lab and work on Lab Assignment 0.

Next Monday:

1. Begin registering clickers (roll-call).
2. Take our unit 1 Readiness Assessment (not graded.)
3. Lesson 1.2 and Application Exercises.

11