Unit 1: Introduction to data

1. Data Collection + Observational studies & experiments

Sta 101 - Spring 2015

Duke University, Department of Statistical Science

January 12, 2015

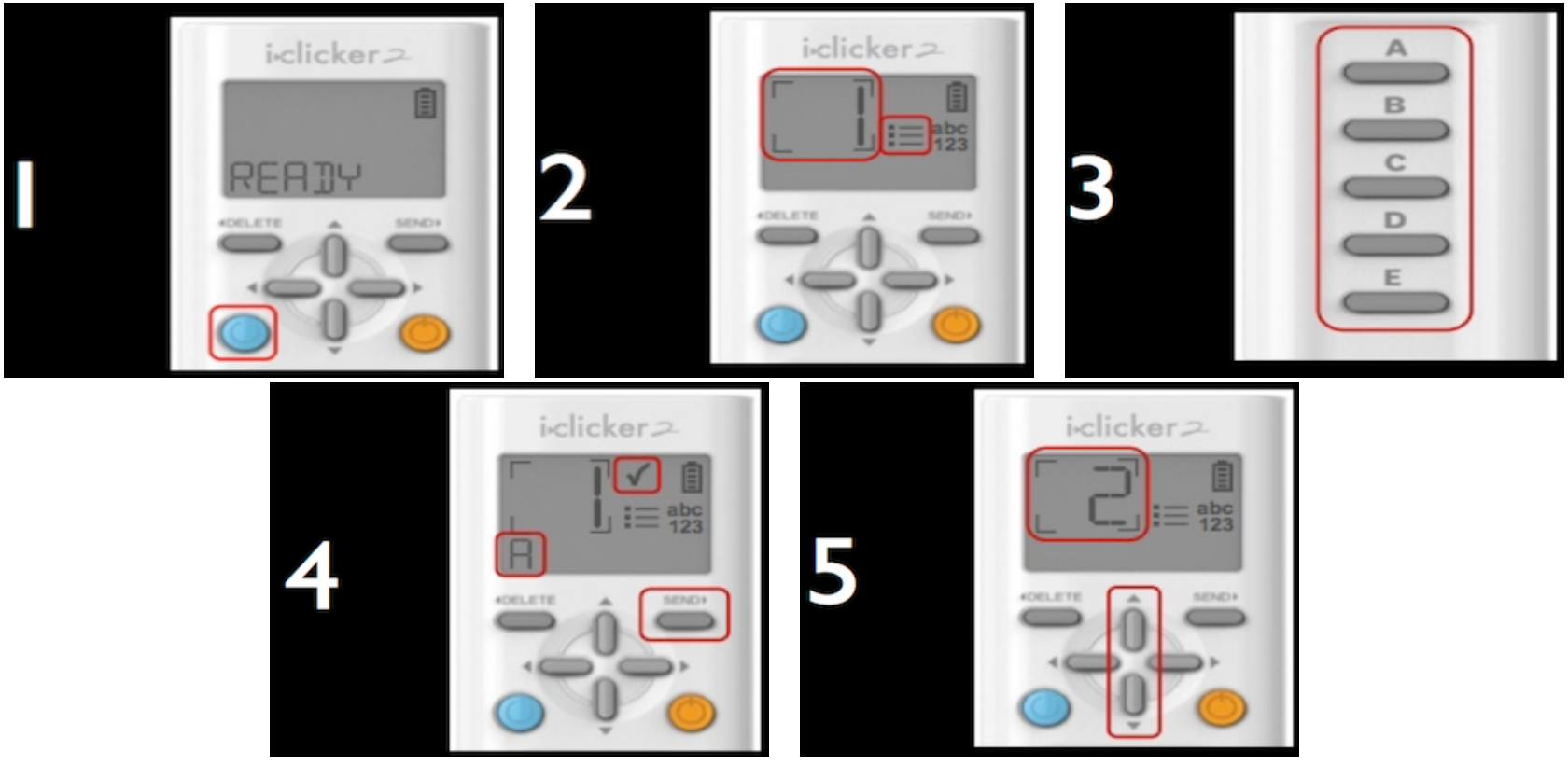
Dr. Çetinkaya-Rundel

Slides posted at http://bitly.com/sta101sp15

Outline

1. Readiness assessment
2. Housekeeping
3. Main ideas
    1. Use a sample to make inferences about the population
    2. Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier
    3. Sampling schemes can suffer from a variety of biases
    4. Experiments use random assignment to treatment groups, observational studies do not
    5. Four principles of experimental design: randomize, control, block, replicate
    6. Random sampling helps generalizability, random assignment helps causality
4. Summary

- *Individual:* 15 minutes, using clickers



- *Team:* 10 minutes, using scratch off sheets (1 per team)

Outline

1. Readiness assessment
2. Housekeeping
3. Main ideas
    1. Use a sample to make inferences about the population
    2. Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier
    3. Sampling schemes can suffer from a variety of biases
    4. Experiments use random assignment to treatment groups, observational studies do not
    5. Four principles of experimental design: randomize, control, block, replicate
    6. Random sampling helps generalizability, random assignment helps causality
4. Summary

- PS 1 due Wednesday on Sakai, by the beginning of class
- Lab tomorrow, sit with your teams

Outline

1. Readiness assessment
2. Housekeeping
3. Main ideas
    1. Use a sample to make inferences about the population
    2. Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier
    3. Sampling schemes can suffer from a variety of biases
    4. Experiments use random assignment to treatment groups, observational studies do not
    5. Four principles of experimental design: randomize, control, block, replicate
    6. Random sampling helps generalizability, random assignment helps causality
4. Summary

Outline

1. Readiness assessment
2. Housekeeping
3. Main ideas
    1. Use a sample to make inferences about the population
    2. Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier
    3. Sampling schemes can suffer from a variety of biases
    4. Experiments use random assignment to treatment groups, observational studies do not
    5. Four principles of experimental design: randomize, control, block, replicate
    6. Random sampling helps generalizability, random assignment helps causality
4. Summary

1. Use a sample to make inferences about the population

- Our ultimate goal is to make inferences about populations

1. Use a sample to make inferences about the population

- Our ultimate goal is to make inferences about populations
- However populations are difficult or impossible to access

1. Use a sample to make inferences about the population

- Our ultimate goal is to make inferences about populations
- However populations are difficult or impossible to access
- Therefore we use a sample from that population, and use *statistics* from that sample to make inferences about the unknown population *parameters*

1. Use a sample to make inferences about the population

- Our ultimate goal is to make inferences about populations
- However populations are difficult or impossible to access
- Therefore we use a sample from that population, and use *statistics* from that sample to make inferences about the unknown population *parameters*
  - We want to know how many offspring female lemurs have, on average

1. Use a sample to make inferences about the population

- Our ultimate goal is to make inferences about populations
- However populations are difficult or impossible to access
- Therefore we use a sample from that population, and use *statistics* from that sample to make inferences about the unknown population *parameters*
  - 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

1. Use a sample to make inferences about the population

- Our ultimate goal is to make inferences about populations
- However populations are difficult or impossible to access
- Therefore we use a sample from that population, and use *statistics* from that sample to make inferences about the unknown population *parameters*
  - 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

1. Use a sample to make inferences about the population

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

1. Use a sample to make inferences about the population

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

1. Use a sample to make inferences about the population

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

Can you see any limitations to using data from the Duke Lemur Center to make inferences about all lemurs?

- 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)

Outline

1. Readiness assessment
2. Housekeeping
3. Main ideas
    1. Use a sample to make inferences about the population
    2. Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier
    3. Sampling schemes can suffer from a variety of biases
    4. Experiments use random assignment to treatment groups, observational studies do not
    5. Four principles of experimental design: randomize, control, block, replicate
    6. Random sampling helps generalizability, random assignment helps causality
4. Summary

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

Demo: *http://bl.ocks.org/avimoondra*

- *Simple random sampling:* Randomly select cases from the population, each case is equally likely to be selected
- *Stratified sampling:* First divide the population into homogenous *strata*, then randomly sample from each stratum
  - e.g. Stratify to control for socio-economic status
- *Cluster sampling:* First randomly sample a few clusters, then randomly sample from within them
  - *Clusters* are not necessarily homogenous, but ideally they're not too different from each other
  - e.g. First sample a few schools from a school district, and then only sample students from within those schools
  - Usually preferred for economical reasons

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, some with only apartments, and others a diverse mixture of housing structures. Which approach would likely be the least effective?

(a) Simple random sampling

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

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

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, some with only apartments, and others a diverse mixture of housing structures. Which approach would likely be the least effective?

(a) Simple random sampling

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

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

1. Readiness assessment

2. Housekeeping

3. Main ideas
    1. Use a sample to make inferences about the population
    2. Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier
    3. Sampling schemes can suffer from a variety of biases
    4. Experiments use random assignment to treatment groups, observational studies do not
    5. Four principles of experimental design: randomize, control, block, replicate
    6. Random sampling helps generalizability, random assignment helps causality

4. Summary

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

3. Sampling schemes can suffer from a variety of biases

- *Non-response:* If only a small 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

- *Non-response:* If only a small 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
- *Convenience sample:* Individuals who are easily accessible are more likely to be included in the sample

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

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

Outline

1. Readiness assessment
2. Housekeeping
3. Main ideas
    1. Use a sample to make inferences about the population
    2. Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier
    3. Sampling schemes can suffer from a variety of biases
    4. Experiments use random assignment to treatment groups, observational studies do not
    5. Four principles of experimental design: randomize, control, block, replicate
    6. Random sampling helps generalizability, random assignment helps causality
4. Summary

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.

http://www.nytimes.com/2014/06/30/technology/

facebook-tinkers-with-users-emotions-in-news-feed-experiment-stirring-outcry.html

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?

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?

Observational

What is the conclusion of the study?

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

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?

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?

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?

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?

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

Outline

1. Readiness assessment
2. Housekeeping
3. Main ideas
    1. Use a sample to make inferences about the population
    2. Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier
    3. Sampling schemes can suffer from a variety of biases
    4. Experiments use random assignment to treatment groups, observational studies do not
    5. Four principles of experimental design: randomize, control, block, replicate
    6. Random sampling helps generalizability, random assignment helps causality
4. Summary

- We would like to design an experiment to investigate if increased stress causes muscle cramps:

- We would like to design an experiment to investigate if increased stress causes muscle cramps:
  - Treatment: increased stress
  - Control: no or baseline stress

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

- We would like to design an experiment to investigate if increased stress causes muscle cramps:
  - 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.

- We would like to design an experiment to investigate if increased stress causes muscle cramps:
  - 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?

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

- We would like to design an experiment to investigate if increased stress causes muscle cramps:
  - 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?

Demo: *http://bl.ocks.org/avimoondra*

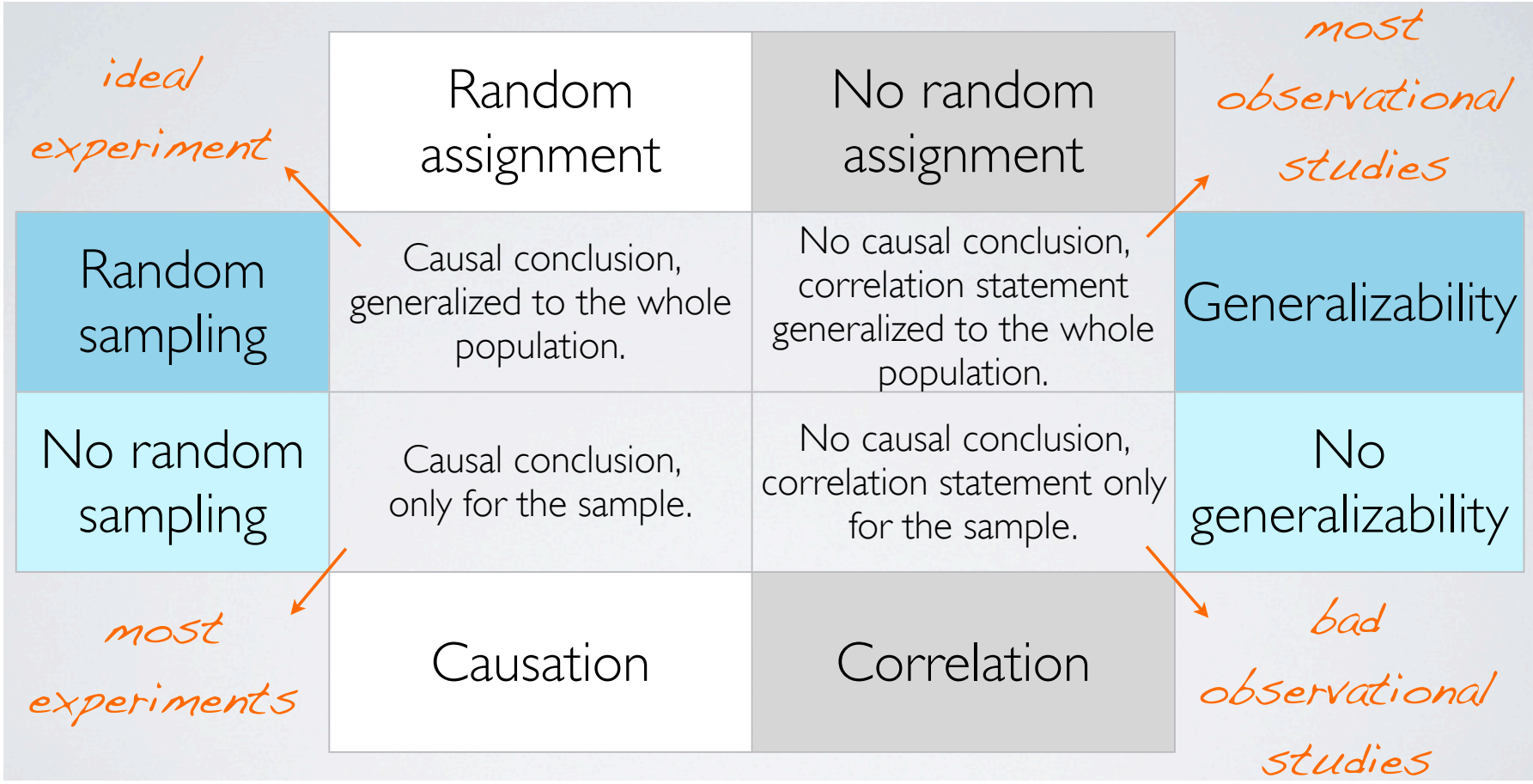
1. Readiness assessment

2. Housekeeping

3. Main ideas
    1. Use a sample to make inferences about the population
    2. Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier
    3. Sampling schemes can suffer from a variety of biases
    4. Experiments use random assignment to treatment groups, observational studies do not
    5. Four principles of experimental design: randomize, control, block, replicate
    6. Random sampling helps generalizability, random assignment helps causality

4. Summary

6. Random sampling helps generalizability, random assignment helps causality



Application exercise: 1.1 Scientific studies in the press

Read media coverage of a study titled "Haters Are Gonna Hate, Study Confirms" and answer the following questions. If the relevant information isn't in the article, refer to the original study.

1. What are the cases?
2. What is (are) the response variable(s) in this study?
3. What is (are) the explanatory variable(s) in this study?
4. Does the study employ random sampling? How about random assignment?
5. Is this an observational study or an experiment? Explain your reasoning.
6. Can we establish a causal link between the explanatory and response variables?
7. Can the results of the study be generalized to the population at large?

Outline

1. Readiness assessment
2. Housekeeping
3. Main ideas
    1. Use a sample to make inferences about the population
    2. Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier
    3. Sampling schemes can suffer from a variety of biases
    4. Experiments use random assignment to treatment groups, observational studies do not
    5. Four principles of experimental design: randomize, control, block, replicate
    6. Random sampling helps generalizability, random assignment helps causality
4. Summary

1. Use a sample to make inferences about the population
2. Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier
3. Sampling schemes can suffer from a variety of biases
4. Experiments use random assignment to treatment groups, observational studies do not
5. Four principles of experimental design: randomize, control, block, replicate
6. Random sampling helps generalizability, random assignment helps causality