Elementary Statistics - Collecting Data

Dr. Ab Mosca (they/them)

Slides based off slides courtesy of OpenIntro and John McGreedy of Johns Hopkins University

Plan for Today

- Populations vs samples
- Sampling
- Study design

Populations vs Samples

Study Example

Research Question: ?

PHYS ED | AUGUST 29, 2012, 12:01 AM | 21 Comments

Finding Your Ideal Running Form

By GRETCHEN REYNOLDS

David De Lossy/Getty Images

https://bit.ly/45vscjA

Study Example

PHYS ED | AUGUST 29, 2012, 12:01 AM | 21 Comments

Finding Your Ideal Running Form

By GRETCHEN REYNOLDS

David De Lossy/Getty Images

https://bit.ly/45vscjA

Research Question: Can people become better, more efficient runners on their own, merely by running?

Study Example

PHYS ED | AUGUST 29, 2012, 12:01 AM | 21 Comments

Finding Your Ideal Running Form

By GRETCHEN REYNOLDS

David De Lossy/Getty Images

https://bit.ly/45vscjA

Research Question: Can people become better, more efficient runners on their own, merely by running?

Population of Interest: ?

Study Example

PHYS ED | AUGUST 29, 2012, 12:01 AM | 21 Comments

Finding Your Ideal Running Form

By GRETCHEN REYNOLDS

David De Lossy/Getty Images

https://bit.ly/45vscjA

Research Question: Can people become better, more efficient runners on their own, merely by running?

Population of Interest: All people

Study Example

PHYS ED | AUGUST 29, 2012, 12:01 AM | 21 Comments

Finding Your Ideal Running Form

By GRETCHEN REYNOLDS

David De Lossy/Getty Images

https://bit.ly/45vscjA

Research Question: Can people become better, more efficient runners on their own, merely by running?

Population of Interest: All people

People in the study: ?

Study Example

PHYS ED | AUGUST 29, 2012, 12:01 AM | 21 Comments

Finding Your Ideal Running Form

By GRETCHEN REYNOLDS

David De Lossy/Getty Images

https://bit.ly/45vscjA

Research Question: Can people become better, more efficient runners on their own, merely by running?

Population of Interest: All people

People in the study: Group of adult women who recently joined a running group

Study Example

PHYS ED | AUGUST 29, 2012, 12:01 AM | 21 Comments

Finding Your Ideal Running Form

By GRETCHEN REYNOLDS

David De Lossy/Getty Images

https://bit.ly/45vscjA

Research Question: Can people become better, more efficient runners on their own, merely by running?

Population of Interest: All people

People in the study: Group of adult women who recently joined a running group

Population to which results can be generalized: ?

Study Example

PHYS ED | AUGUST 29, 2012, 12:01 AM | 21 Comments

Finding Your Ideal Running Form

By GRETCHEN REYNOLDS

David De Lossy/Getty Images

https://bit.ly/45vscjA

Research Question: Can people become better, more efficient runners on their own, merely by running?

Population of Interest: All people

People in the study: Group of adult women who recently joined a running group

Population to which results can be generalized: Adult women (assuming the data are randomly sampled)

Study Example

PHYS ED | AUGUST 29, 2012, 12:01 AM | 21 Comments

Finding Your Ideal Running Form

By GRETCHEN REYNOLDS

David De Lossy/Getty Images

https://bit.ly/45vscjA

Why don't our results generalize to the population of interest?

Research Question: Can people become better, more efficient runners on their own, merely by running?

Population of Interest: All people

People in the study: Group of adult women who recently joined a running group

Population to which results can be generalized: Adult women (assuming the data are randomly sampled)

Study Example

PHYS ED | AUGUST 29, 2012, 12:01 AM | 21 Comments

Finding Your Ideal Running Form

By GRETCHEN REYNOLDS

David De Lossy/Getty Images

https://bit.ly/45vscjA

Why don't our results generalize to the population of interest?

Research Question: Can people become better, more efficient runners on their own, merely by running?

Population of Interest: All people

≠

People in the study: Group of adult women who recently joined a running group

Population to which results can be generalized: Adult women (assuming the data are randomly sampled)

Study Example

PHYS ED | AUGUST 29, 2012, 12:01 AM | 21 Comments

Finding Your Ideal Running Form

By GRETCHEN REYNOLDS

David De Lossy/Getty Images

https://bit.ly/45vscjA

Why don't our results generalize to the population of interest?

Research Question: Can people become better, more efficient runners on their own, merely by running?

Population of Interest: All people

≠

~~*People in the study*~~ ***Sample***: Group of adult women who recently joined a running group

Population to which results can be generalized: Adult women (assuming the data are randomly sampled)

Populations and Samples

- In statistics…
  - a ***population*** is the group about which you are trying to answer a question
  - a ***sample*** is the portion of the population you are actually analyzing

Populations and Samples

- In statistics…
  - a ***population*** is the group about which you are trying to answer a question
  - a ***sample*** is the portion of the population you are actually analyzing

- Wouldn't it be better to just include everyone and "sample" the entire population?

Populations and Samples

- In statistics…
  - a ***population*** is the group about which you are trying to answer a question
  - a ***sample*** is the portion of the population you are actually analyzing

- Wouldn't it be better to just include everyone and "sample" the entire population?
  - This is called a *census*.

Populations and Samples

- Problems with a census
  - It can be difficult to complete a census: there always seem to be some individuals who are hard to locate or hard to measure. *And these difficult-to-find people may have certain characteristics that distinguish them from the rest of the population.*
  - Populations rarely stand still. Even if you could take a census, the population changes constantly, so it's never possible to get a perfect measure.
  - Taking a census may be more complex than sampling.

Populations and Samples

- (Some) Problems with the US census…
  - Does not reach all subpopulations equally
    - undercounts Hispanic, Black, and Native American residents
    - overcounts white and Asian American residents
    - undocumented residents
    - undercounts residents living in poor urban areas
    - undercounts unhoused people
  - Does not collect data in a representative way
    - sex (only binary options, no intersex option)
    - gender is not collected
    - race includes no multiracial option

Populations and Samples

- Because a census is problematic, we use samples and infer

Populations and Samples

- Because a census is problematic, we use samples and infer
- Think about cooking
  - You might taste a spoonful of soup and decide the spoonful you tasted isn't salty enough
  - If you generalize and conclude that your entire soup needs salt, that's an *inference*

Populations and Samples

- Because a census is problematic, we use samples and infer
- Think about cooking
  - You might taste a spoonful of soup and decide the spoonful you tasted isn't salty enough
  - 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).
  - If your spoonful comes only from the surface and the salt is collected at the bottom of the pot, what you tasted is probably not representative of the whole pot.
  - If you first stir the soup thoroughly before you taste, your spoonful will more likely be representative of the pot.

Sampling

Samples

- It is unethical to collect data from and/or perform research on someone who has not consented to participation
- Samples consist of volunteered information
- As a result, bias can occur

Sampling Bias

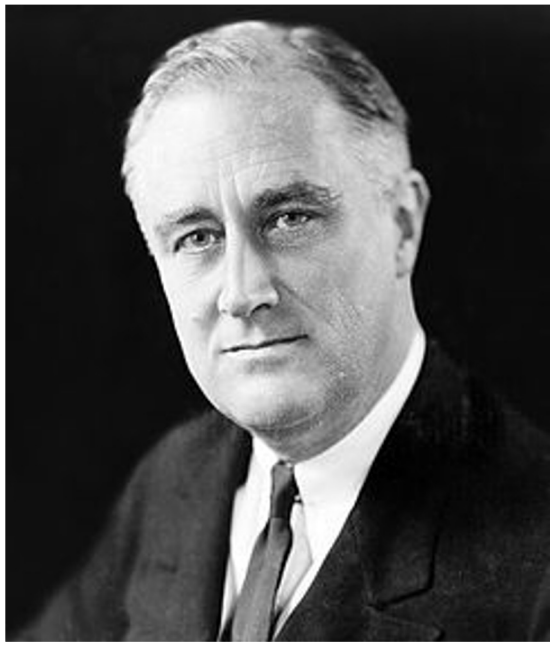
Bias is caused by…

- Non-response: If only a small fraction of sampled people choose to participate, the sample may no longer be representative of the population.
- Voluntary response: Occurs when the sample consists of people who volunteer to participate because they have strong opinions on the issue. 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.

Sampling Bias

An historical example of a biased sample yielding misleading results

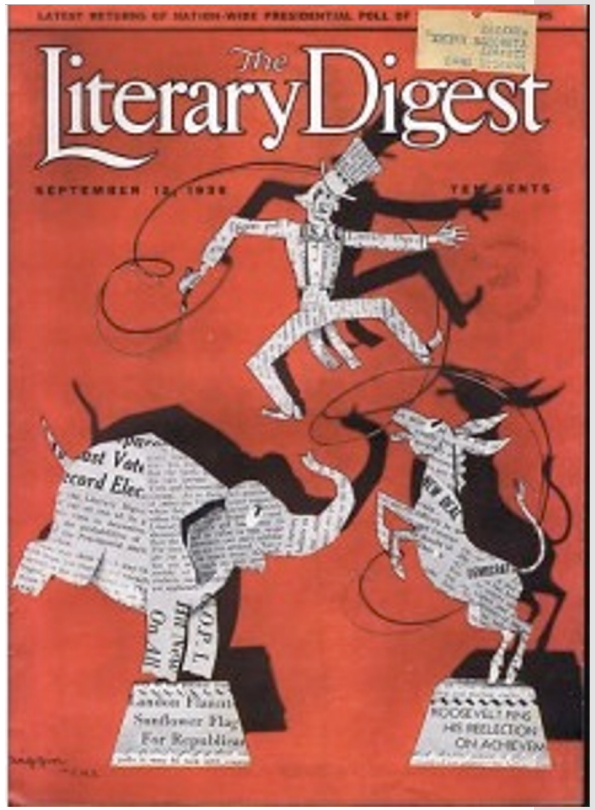
In 1936, Landon sought the Republican presidential nomination opposing the re-election of FDR.



Sampling Bias

The Literary Digest Poll

- The Literary Digest polled about 10 million Americans, and got responses from about 2.4 million.
- The poll showed that Landon would likely be the overwhelming winner and FDR would get only 43% of the votes.
- Election result: FDR won, with 62% of the votes.
- The magazine was completely discredited because of the poll, and was soon discontinued.



Sampling Bias

The Literary Digest Poll - what went wrong?

- The magazine had surveyed
  - its own readers,
  - registered automobile owners, and
  - registered telephone users.
- These groups had incomes well above the national average of the day (remember, this is Great Depression era) which resulted in lists of voters far more likely to support Republicans than a truly typical voter of the time, i.e. the sample was not representative of the American population at the time.

Sampling

Large samples are preferable, but...

- The Literary Digest election poll was based on a sample size of 2.4 million, which is huge, but since the sample was *biased*, the sample did not yield an accurate prediction.
- Back to the soup analogy: If the soup is not well stirred, it doesn't matter how large a spoon you have, it will still not taste right. If the soup is well stirred, a small spoon will suffice to test the soup.

Practice

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. 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**

Practice

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. 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**

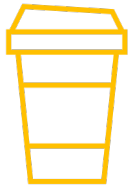
Studies

Studies

- Studies explore the (potential) relationship between a stimuli or treatment (independent or explanatory variable(s)) and an outcome or response (dependent or response variable(s))

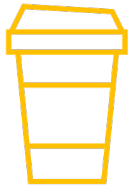
Studies

- Studies explore the (potential) relationship between a stimuli or treatment (independent or explanatory variable(s)) and an outcome or response (dependent or response variable(s))
- Ex. Caffeine and Sleep



Studies

- Studies explore the (potential) relationship between a stimuli or treatment (independent or explanatory variable(s)) and an outcome or response (dependent or response variable(s))
- Ex. Caffeine and Sleep



- There are two main types of studies
  - Observational
  - Experimental

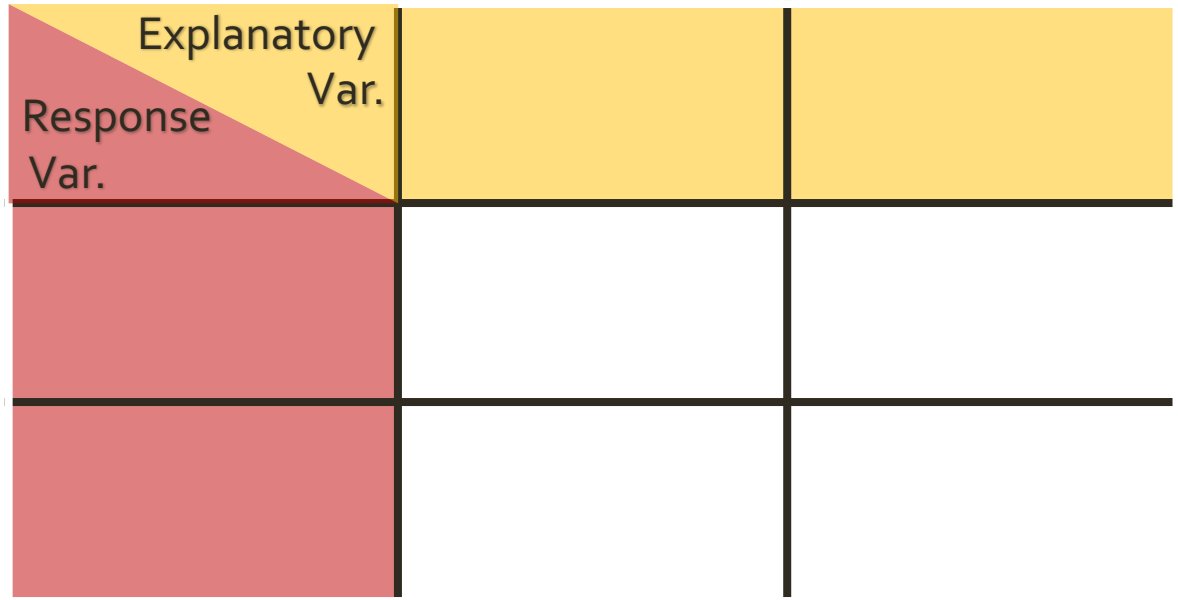
Observational Studies

Observational Studies

- Researchers collect data in a way that does not directly interfere with how the data arise.
  - Ex. Look at medical records and receipts to see if people who bought coffee went to the doctor for sleep related issues

Observational Studies

- Researchers collect data in a way that does not directly interfere with how the data arise.
  - Ex. Look at medical records and receipts to see if people who bought coffee went to the doctor for sleep related issues



Observational Studies

- Researchers collect data in a way that does not directly interfere with how the data arise.
  - Ex. Look at medical records and receipts to see if people who bought coffee went to the doctor for sleep related issues

Explanatory Var. / Response Var.	Coffee	No Coffee
Sleep Issues		
No Sleep Issues		

Observational Studies

- Researchers collect data in a way that does not directly interfere with how the data arise.
  - Ex. Look at medical records and receipts to see if people who bought coffee went to the doctor for sleep related issues

Explanatory Var. / Response Var.	Coffee	No Coffee
Sleep Issues		
No Sleep Issues		

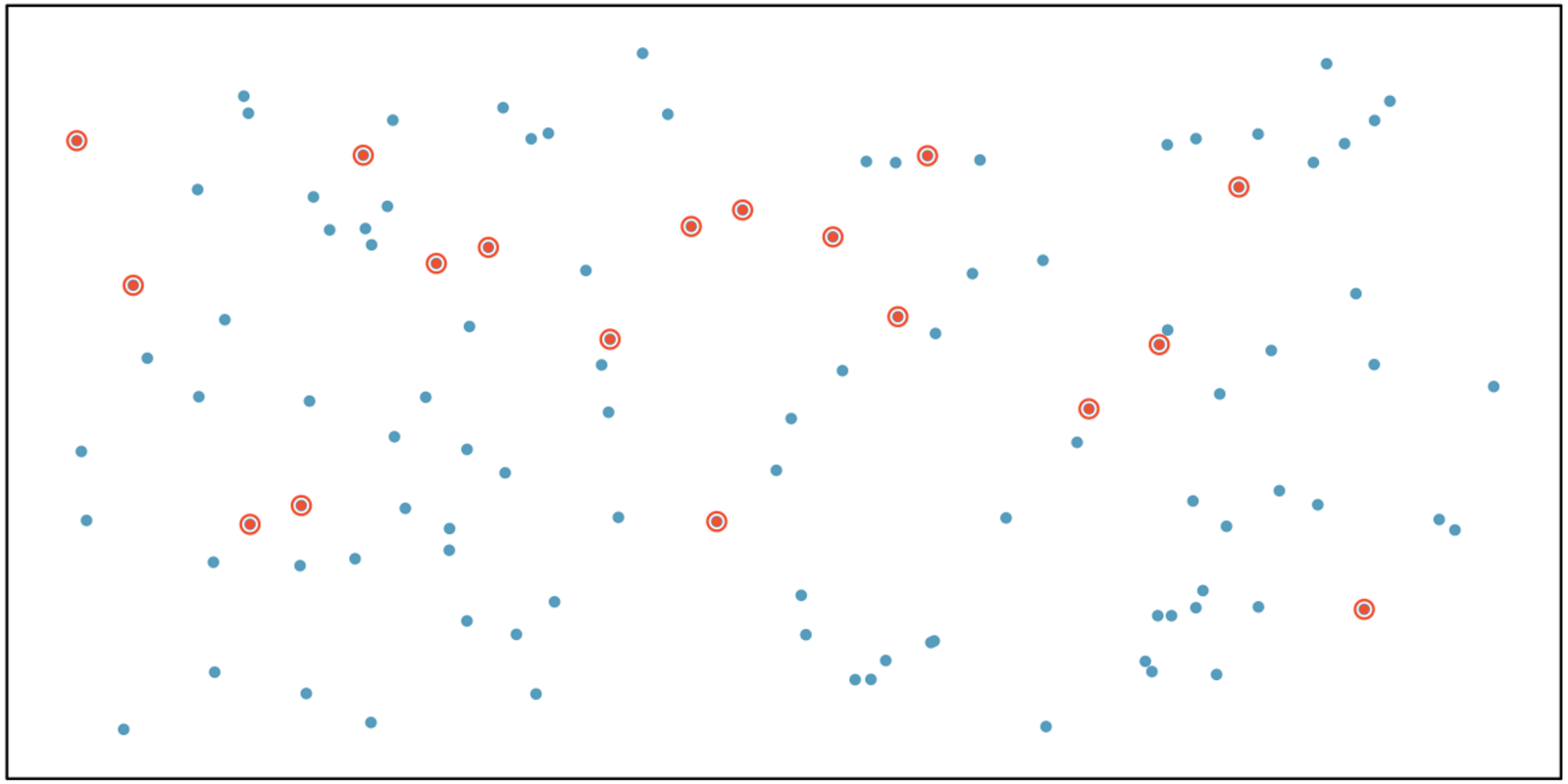
- Can show an association, but not causal connection

Sampling for Observational Studies

- There are different methods for sampling to reduce bias
- Remember, the goal is to get as representative a sample as possible

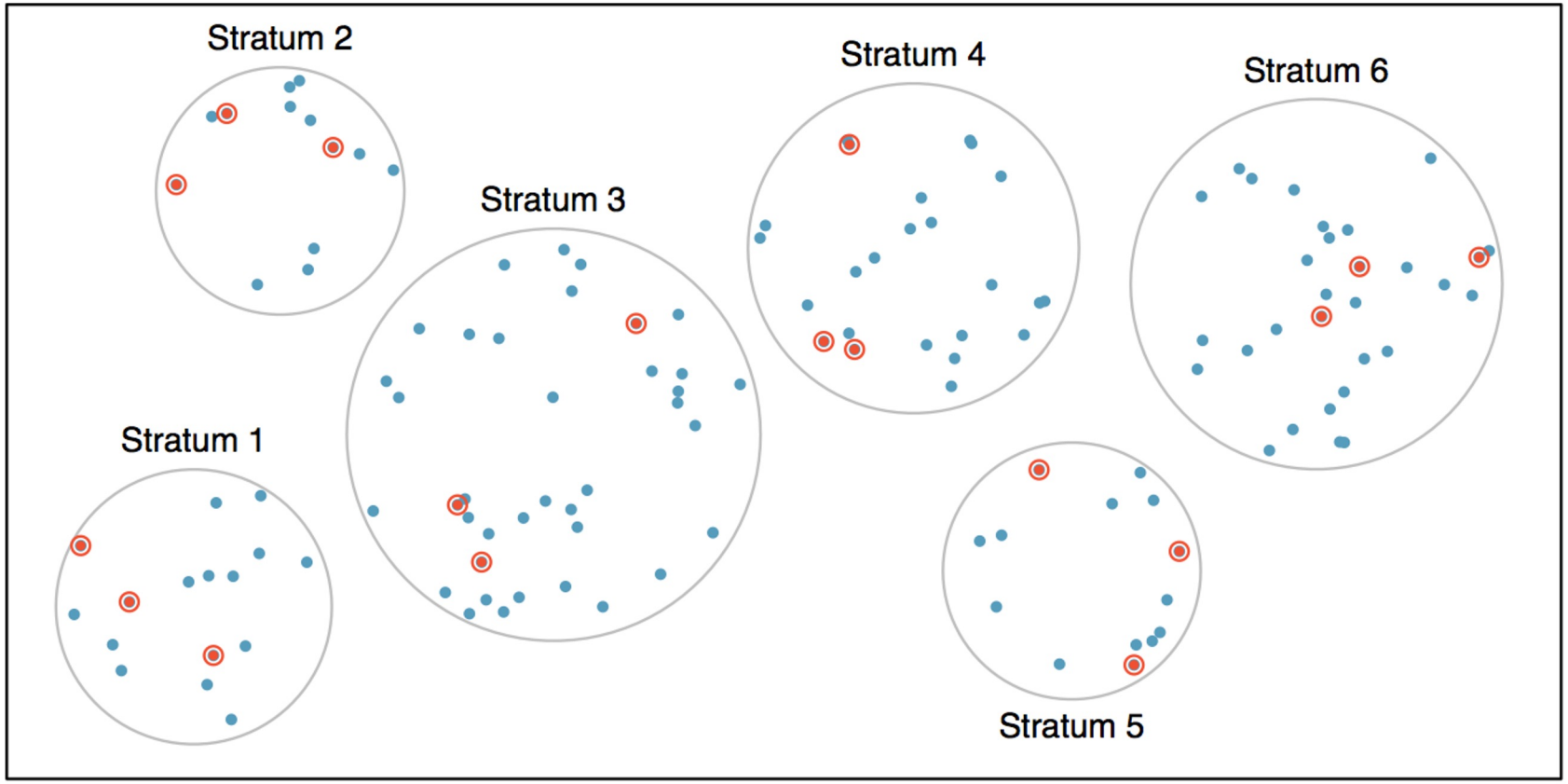
Simple Random Sample

Randomly select cases from the population, where there is no implied connection between the points that are selected.



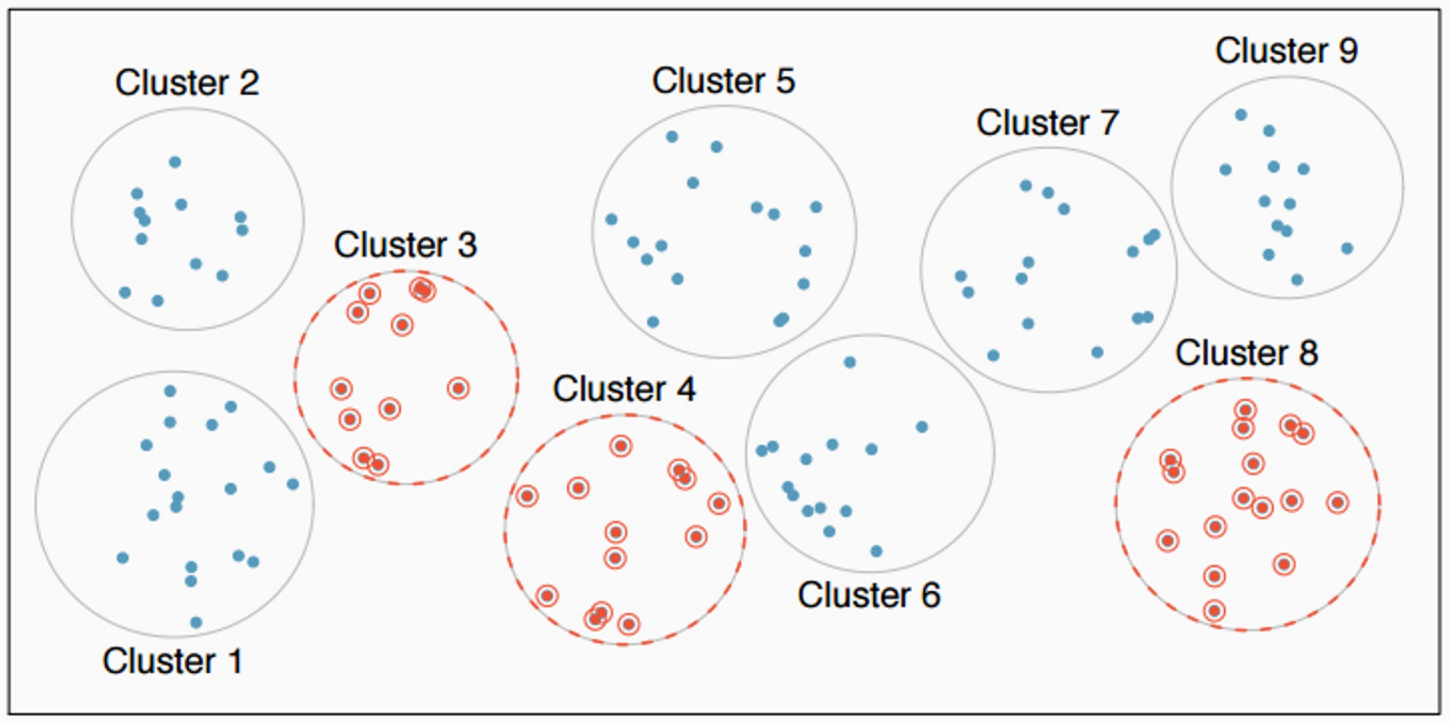
Stratified Sample

Strata are made up of similar observations. We take a simple random sample from each stratum.



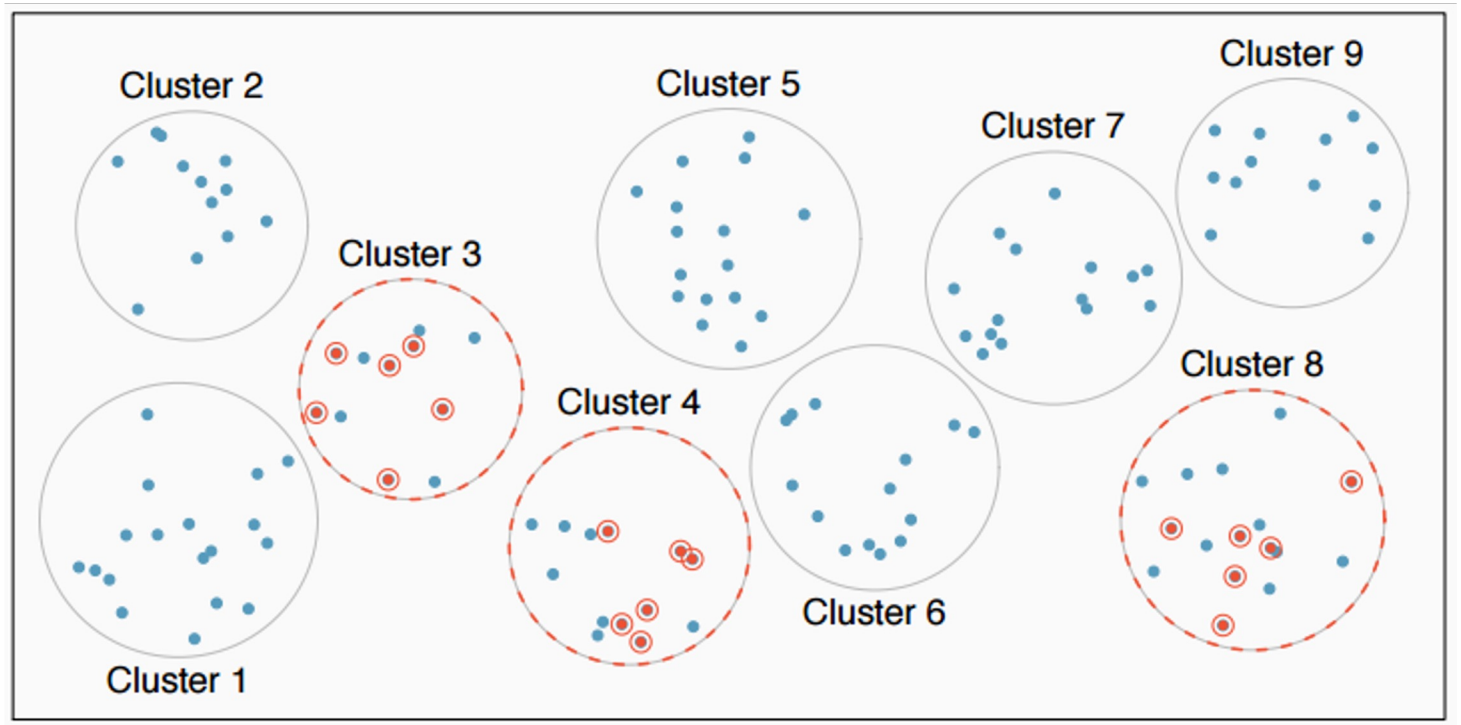
Cluster Sample

Clusters are usually not made up of homogeneous observations. We take a simple random sample of clusters, and then sample all observations in that cluster. Usually preferred for economical reasons.



Multistage Sample

Clusters are usually not made up of homogeneous observations. We take a simple random sample of clusters, and then take a simple random sample of observations from the sampled clusters



Practice

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. Which approach would likely be the *least* effective?

(a) Simple random sampling
(b) Cluster sampling
(c) Stratified sampling
(d) Multistage sampling

Practice

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. Which approach would likely be the *least* effective?

(a) Simple random sampling

(b) Cluster sampling

(c) Stratified sampling

(d) Multistage sampling

Experiments

Experimental Studies

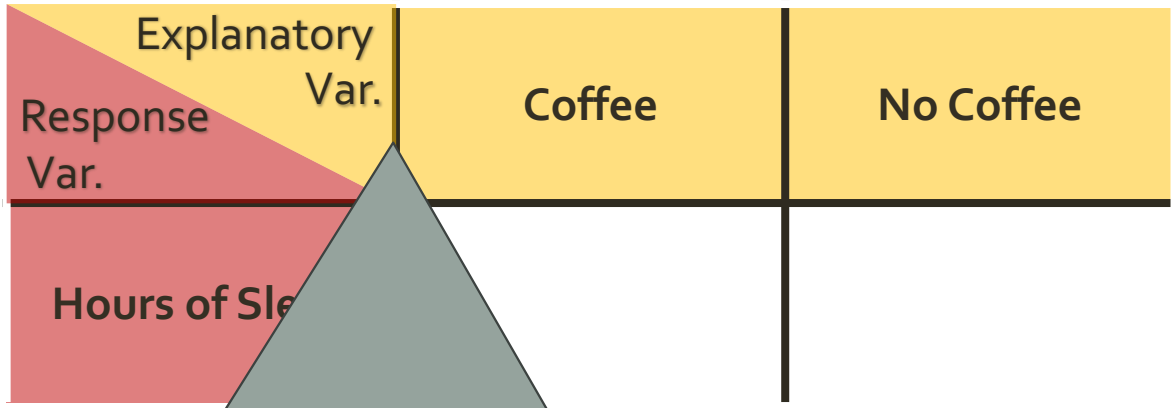
- Researchers collect data in a highly structured way.
  - Ex. Have some people drink coffee at 8am and noon every day for 3 months, record how many hours of sleep they get each night. Compare to people who drink no coffee

Experimental Studies

- Researchers collect data in a highly structured way.
  - Ex. Have some people drink coffee at 8am and noon every day for 3 months, record how many hours of sleep they get each night. Compare to people who drink no coffee

Experimental Studies

- Researchers collect data in a highly structured way.
  - Ex. Have some people drink coffee at 8am and noon every day for 3 months, record how many hours of sleep they get each night. Compare to people who drink no coffee

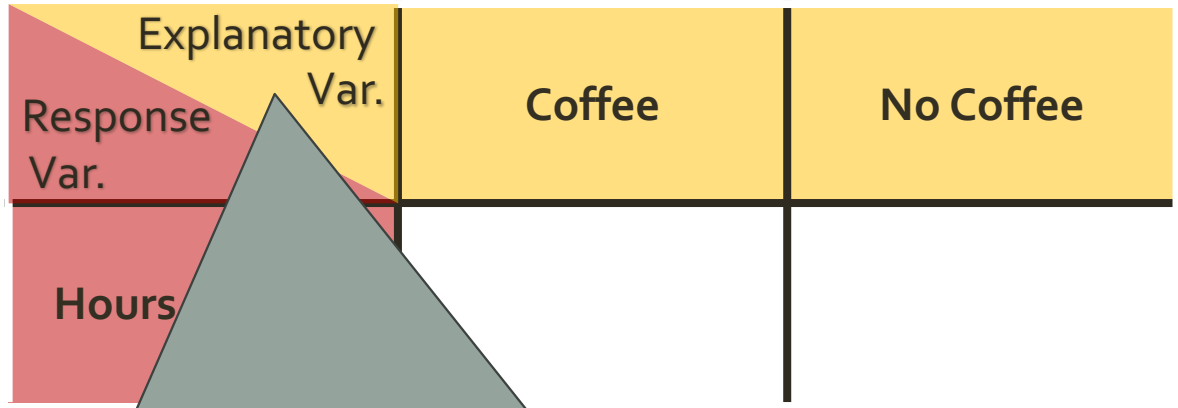


Explanatory variable is also called a **factor** in experiments. **Factors have** distinct **levels**, in this case Coffee and No Coffee.

Experimental Studies

- Researchers collect data in a highly structured way.

What other levels could the coffee factor have?

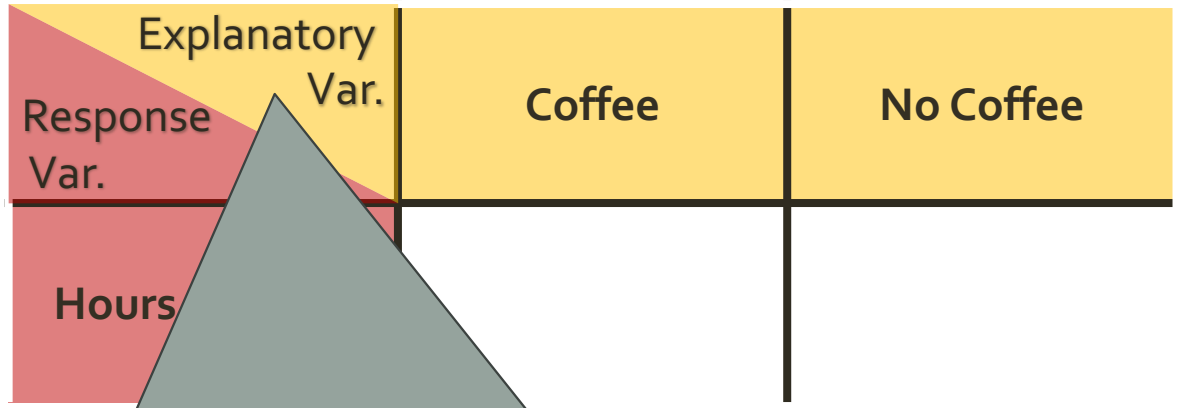


Explanatory variable is also called a **factor** in experiments. **Factors have** distinct **levels**, in this case Coffee and No Coffee.

Experimental Studies

- Researchers collect data in a highly structured way.

What other factors could we add to this experiment?

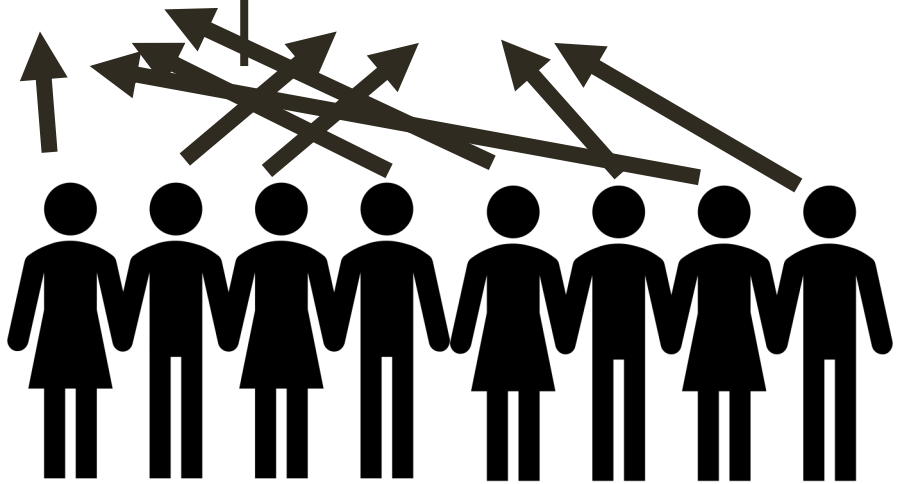


Explanatory variable is also called a **factor** in experiments. **Factors have** distinct **levels**, in this case Coffee and No Coffee.

Random Assignment

- Researchers collect data in a highly structured way.
  - Ex. Have some people drink coffee at 8am and noon every day for 3 months, record how many hours of sleep they get each night. Compare to people who drink no coffee

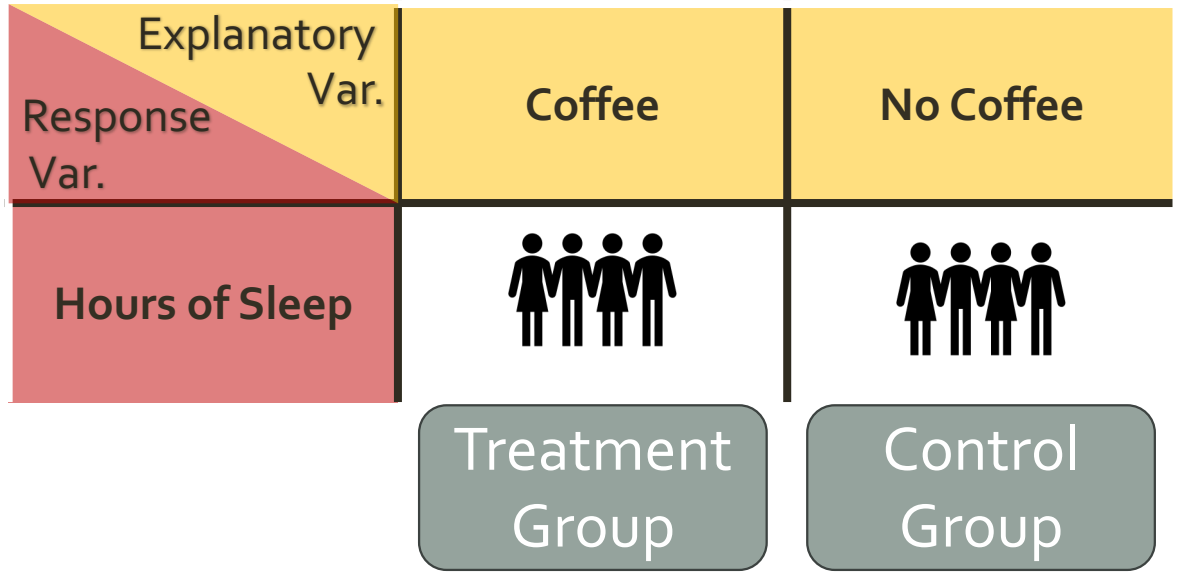
Explanatory Var. / Response Var.	Coffee	No Coffee
Hours of Sleep		



Participants get randomly assigned to a group

Experimental Studies

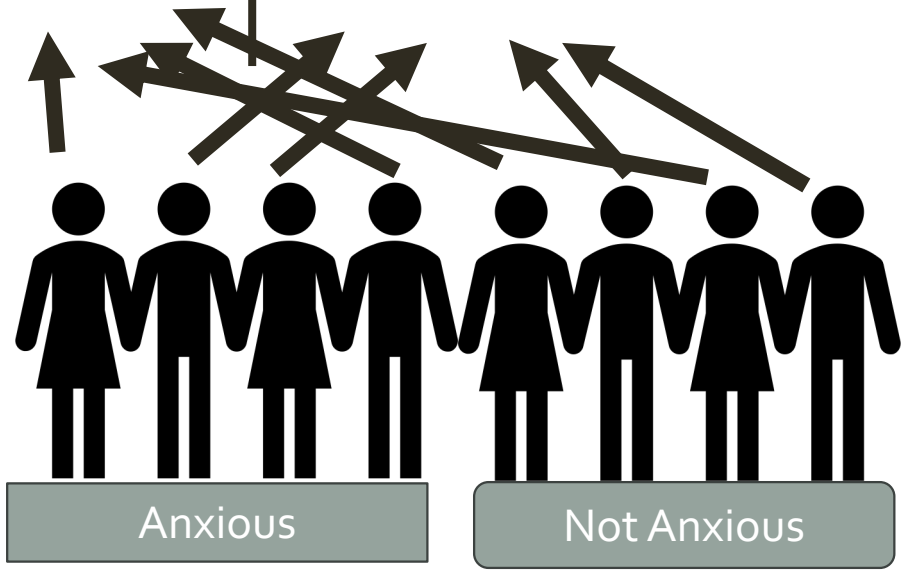
- Researchers collect data in a highly structured way.
  - Ex. Have some people drink coffee at 8am and noon every day for 3 months, record how many hours of sleep they get each night. Compare to people who drink no coffee



Blocking

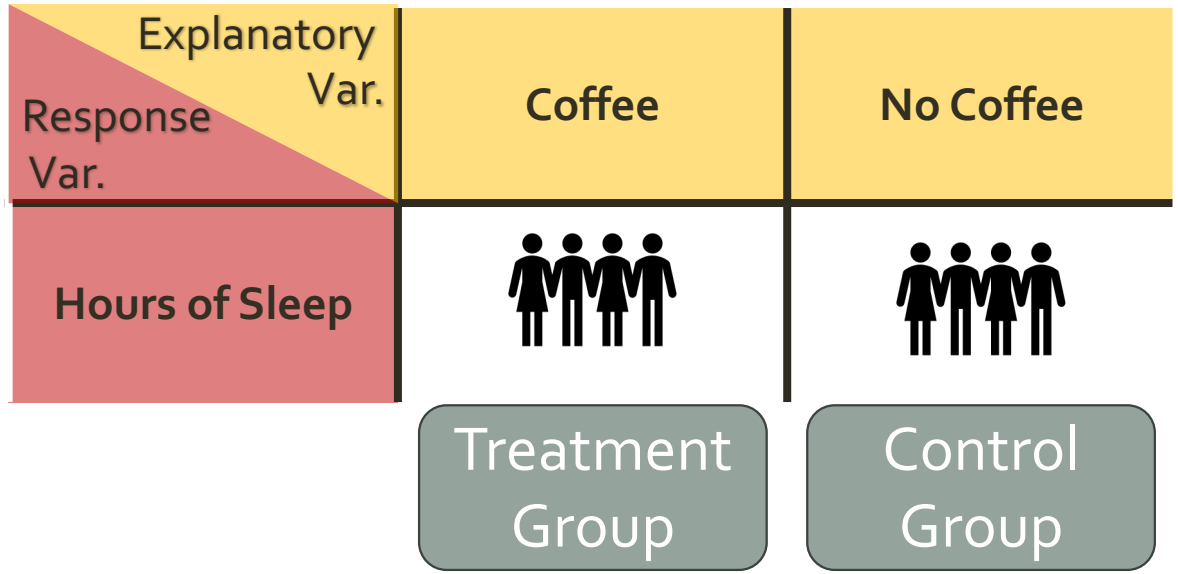
- Researchers collect data in a highly structured way.
  - Ex. Have some people drink coffee at 8am and noon every day for 3 months, record how many hours of sleep they get each night. Compare to people who drink no coffee

Participants are ***blocked*** then ***randomly assigned*** to a group



Experimental Studies

- Researchers collect data in a highly structured way.
  - Ex. Have some people drink coffee at 8am and noon every day for 3 months, record how many hours of sleep they get each night. Compare to people who drink no coffee



  - Can show a causal connection

Principles of Experiment Design

1. **Control**: Compare treatment of interest to a control group.
2. **Randomize**: Randomly assign subjects to treatments, and randomly sample from the population whenever possible.
3. **Replicate**: Within a study, replicate by collecting a sufficiently large sample. Or replicate the entire study.
4. **Block**: If there are variables that are known or suspected to affect the response variable, first group subjects into blocks based on these variables, and then randomize cases within each block to treatment groups.

Experiment Design

- Design an experiment to investigate if energy gels makes you run faster:
  - Treatment: **?**
  - Control: **?**

Experiment Design

- Design an experiment to investigate if energy gels makes you run faster:
  - Treatment: energy gel
  - Control: no energy gel

Experiment Design

- Design an experiment to investigate if energy gels makes you run faster:
  - Treatment: energy gel
  - Control: no energy gel
- It is suspected that energy gels might affect pro and amateur athletes differently.
- **How do you design your experiment?**

Experiment Design

- Design an experiment to investigate if energy gels makes you run faster:
  - Treatment: energy gel
  - Control: no energy gel
- It is suspected that energy gels might affect pro and amateur athletes differently, therefore we block for pro status:
  - Divide the sample to pro and amateur
  - Randomly assign pro athletes to treatment and control groups
  - Randomly assign amateur athletes to treatment and control groups
  - Pro/amateur status is equally represented in the resulting treatment and control groups

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

Practice

A study is designed to test the effect of light level and noise level on exam performance of students. The researcher also believes that light and noise levels might have different effects on high and low income students, so wants to make sure both economic statuses are equally represented in each group. Which of the below is correct?

A. There are 3 explanatory variables (light, noise, economic status) and 1 response variable (exam performance)

B. There are 2 explanatory variables (light and noise), 1 blocking variable (economic status), and 1 response variable (exam performance)

C. There is 1 explanatory variable (economic status) and 3 response variables (light, noise, exam performance)

D. There are 2 blocking variables (light and noise), 1 explanatory variable (economic status), and 1 response variable (exam performance)

Experiment Design

A study is designed to test the effect of light level and noise level on exam performance of students. The researcher also believes that light and noise levels might have different effects on high and low income students, so wants to make sure both economic statuses are equally represented in each group. Which of the below is correct?

A. There are 3 explanatory variables (light, noise, economic status) and 1 response variable (exam performance)

B. There are 2 explanatory variables (light and noise), 1 blocking variable (economic status), and 1 response variable (exam performance)

C. There is 1 explanatory variable (economic status) and 3 response variables (light, noise, exam performance)

D. There are 2 blocking variables (light and noise), 1 explanatory variable (economic status), and 1 response variable (exam performance)

Experiment Design

Difference Between Blocking and Explanatory Variables

- Stimuli are conditions we can impose on the experimental units.
- Blocking variables are characteristics that the experimental units come with, that we would like to control for. [Otherwise these characteristics could be **confounding factors**]
- Blocking is like stratifying, except used in experimental settings when randomly assigning, as opposed to when sampling.

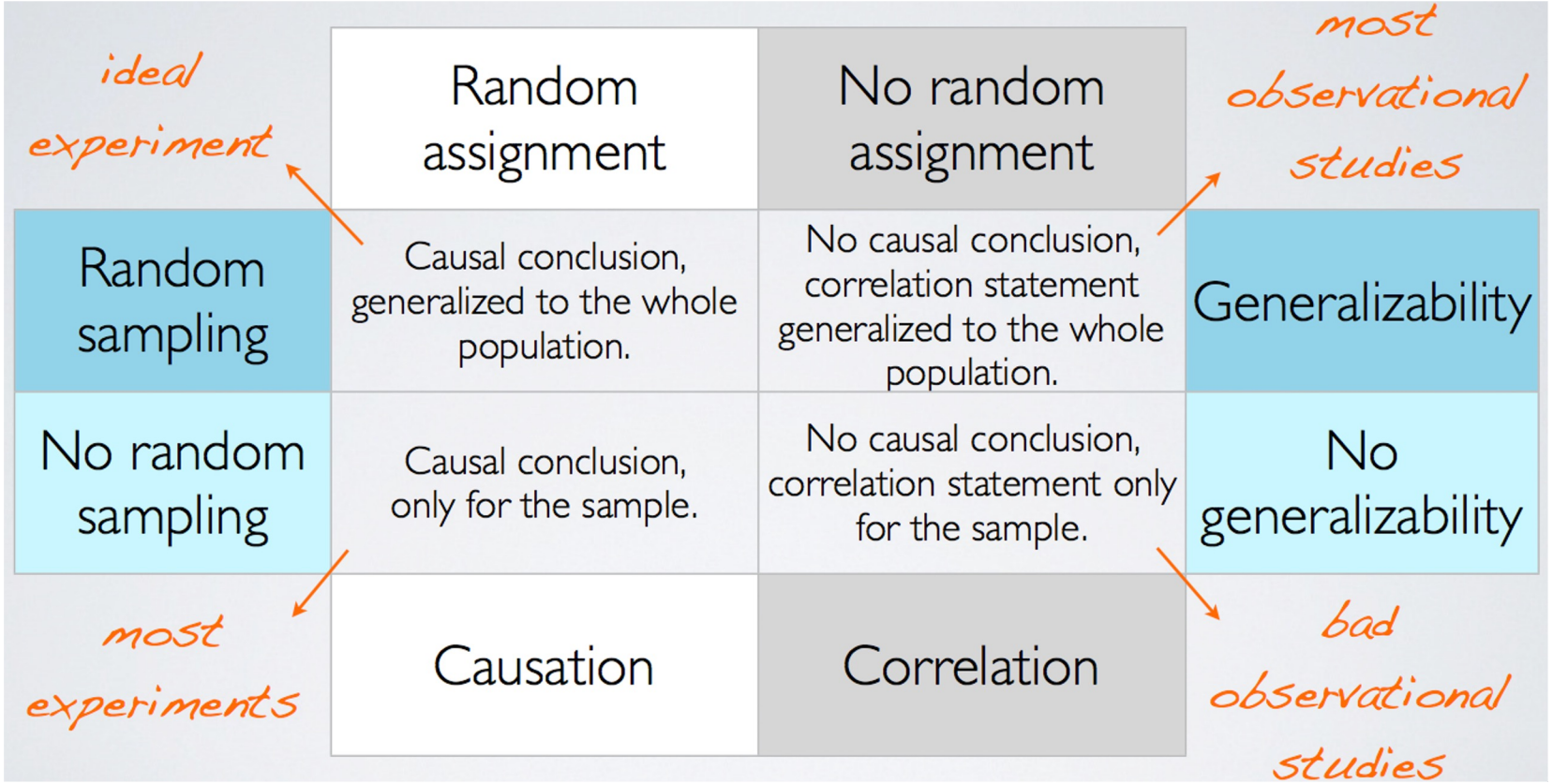
Experiment Design

More Experimental Design Terminology...

- Placebo: fake treatment, often used as the control group for medical studies
- Placebo effect: experimental units showing improvement simply because they believe they are receiving a special treatment
- Blinding: when experimental units do not know whether they are in the control or treatment group
- Double-blind: when both the experimental units and the researchers who interact with the patients do not know who is in the control and who is in the treatment group

Study Design

Random Assignment vs. Random Sampling



Practice

- Work with 2 other people
- Develop a research question of interest to you
- Design a study to answer that question
  - Observational or Experiment?
  - How will you collect a sample or design your experiment?
  - Etc..
- Be prepared to share with everyone