Unit 1: Introduction to Data

1. Data: Where it comes from, and why that matters

(Chapter 1.1-1.5)

12/15/2020

Sampling: The bridge between data and analysis

Key ideas

1. Using samples to make inferences about populations
2. The way you sample your data can change your inferences about the population
3. Experiments use random assignment to treatment groups, observational studies do not
4. Random samples help with **generalizability**, random assignment helps with **causality**

Do you approve of President Trump's drone strike that killed Soleimani?

Do you approve President Trump's drone strike that killed Soleimani?

Why did I have you close your eyes?

I wanted to get **independent** samples

Are there any other sources of **measurement error**?

You might want to give an answer that you think I will like.
This is a **Demand characteristic**. E.g. the Bradley effect

Do **Americans** support the strike?

Support :8%
Don't Support: 92%

24 votes cast

Support : 45%
Don't Support: 41%

1562 votes cast

Support : 58%
Don't Support: 34%

2510 votes cast

Each of these polls is a **sample**

But I want to make an inference to the **population**

When I draw a conclusion about the population from a sample, I make an **inference**.

The way I collect my sample can lead me to different inferences.

Which of these samples is the best?

Larger samples are better samples

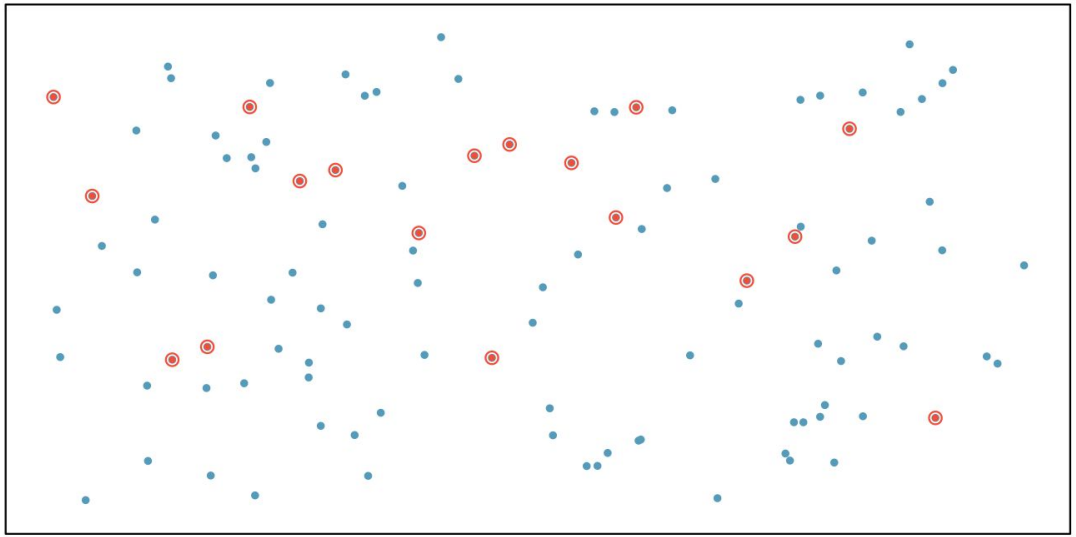
Why is bigger better?

Small samples are more **variable**.

There are 100 dots here, and 18 of them are red.

If I draw 3 dots, **more than half** the time 0 will be red.

If I draw 50 dots,
less than **1 out of 100 billion times** 0 will be red



For random samples, larger samples are more **representative**

Are these **random** samples?

Support :8%
Don't Support: 92%

24 votes cast

Support : 45%
Don't Support: 41%

1562 votes cast

Support : 58%
Don't Support: 34%

2510 votes cast

No! They are **convenience** samples

How representative is the Quinnipiac sample?

Quinnipiac University Poll

Support : 45%
Don't Support: 41%

1562 votes cast

Survey of 1562 self-identified registered voters conducted by telephone (landline and cell) by Quinnipiac from January 8-12, 2020. The margin of sampling error for results based on the total sample is plus or minus +/- percentage points.

The survey includes 651 Democratic voters and independent voters who lean Democratic with a margin of error of +/- 3.8 percentage points.

National average (according to Gallup):
Republican: 26, Democrat:27 Independent:44

How representative is the Defcon sample?

Support : 58%
Don't Support: 34%
2510 votes cast

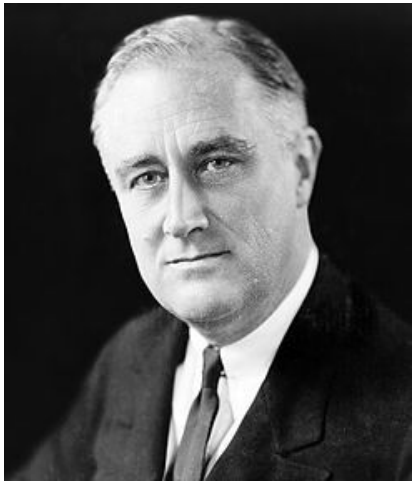
People could vote multiple times.
Why is this bad?

People who voted were likely to be regular readers of the forum.

People from outside the US could vote.
Why is this bad?

Sampling bias in the polls: Landon vs. FDR

Alf Landon

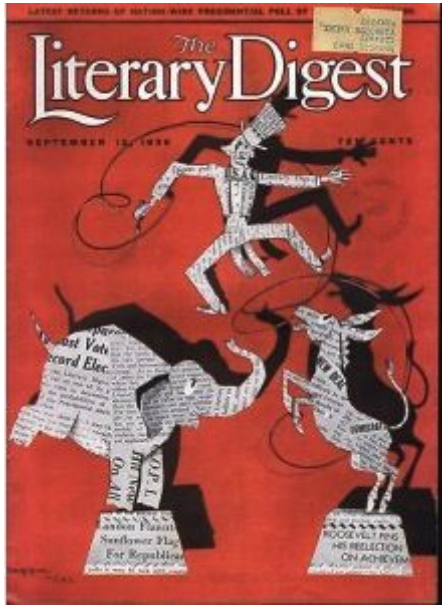


Franklin Delano Roosevelt

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

The Literary Digest poll

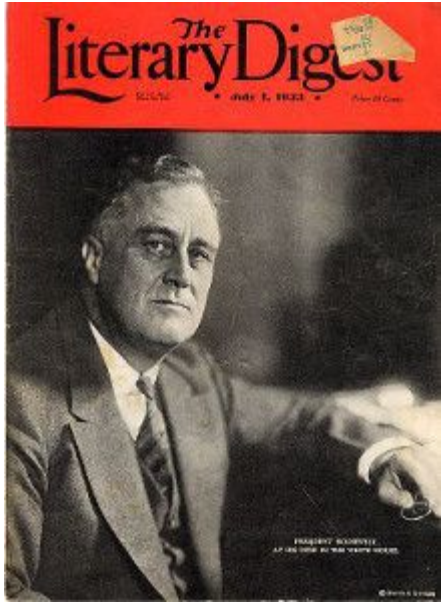
- The Literary Digest polled about 10 million Americans, and got responses from about 2.4 million.
- The poll showed that Landon would likely be the overwhelming winner and FDR would get only 43% of the votes.



- Election result: FDR won, with 62% of the votes.
- The magazine was completely discredited because of the poll, and was soon discontinued.

What went wrong?

- The magazine had surveyed
  - its own readers,
  - registered automobile owners,
  - registered telephone users, and
  - country club members
- These groups had incomes well above the national average—it was the Great Depression!
  - The sample was **not representative**
- This sample was huge—2.4 million people. But it was biased, and thus inaccurate.



A sampling metaphor

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**)

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.

Thanks Mine Çetinkaya-Rundel

Practice Question 1

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?

1. Some of the mailings may have never reached the parents.
2. The district has strong support from parents to move forward with the policy
3. It is possible that majority of the parents disagree with the policy change.
4. The survey results are unlikely to be biased because all parents were mailed a survey.

(a) Only 1 (b) 1 and 2 (c) 1 and 3 (d) 3 and 4 (e) Only 4

Practice Question 1

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?

1. Some of the mailings may have never reached the parents.
2. The district has strong support from parents to move forward with the policy
3. It is possible that majority of the parents disagree with the policy change.
4. The survey results are unlikely to be biased because all parents were mailed a survey.

(a) Only 1 (b) 1 and 2 **(c) 1 and 3** (d) 3 and 4 (e) Only 4

Practice Question 2

Support :8%
Don't Support: 92%

24 votes cast

Support : 45%
Don't Support: 41%

1562 votes cast

Support : 58%
Don't Support: 34%

2510 votes cast

I want to predict whether **Carnegie Mellon Students as a whole** support the drone strike.

Which sample should I use?

Do we know if watching the news played a **causal** role?

What if we ask non watchers?

Support : 8%
Don't Support: 92%

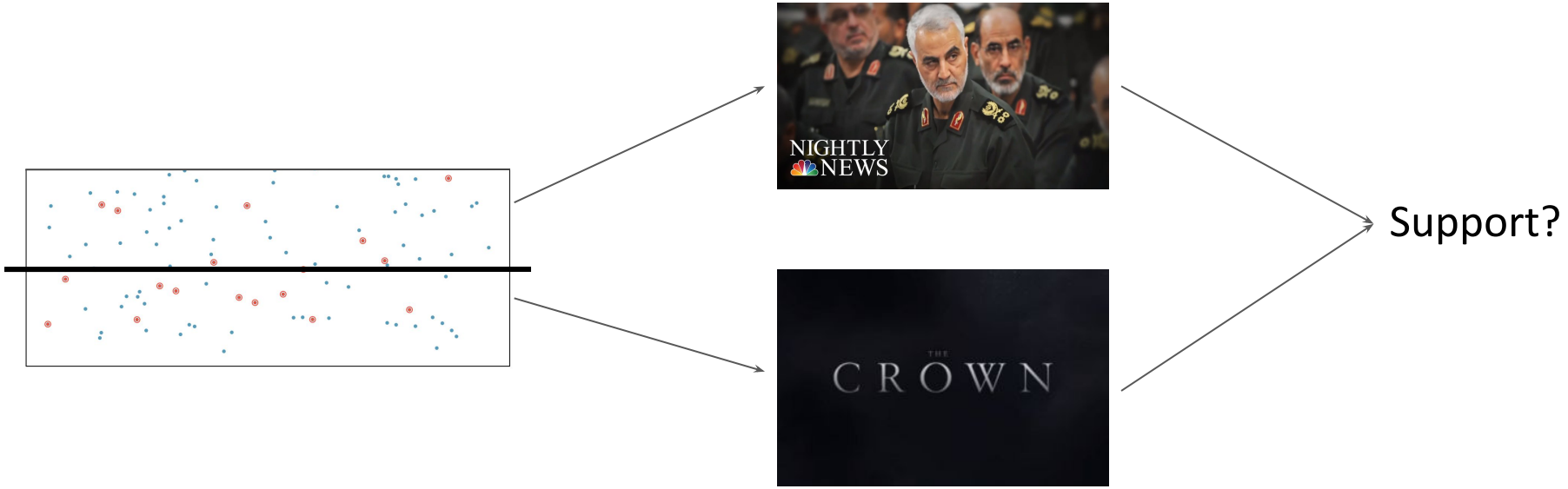
24 votes cast

Can I just compare the watchers and non-watchers to each-other?
Can the non-watchers be a **control** group for the watchers?

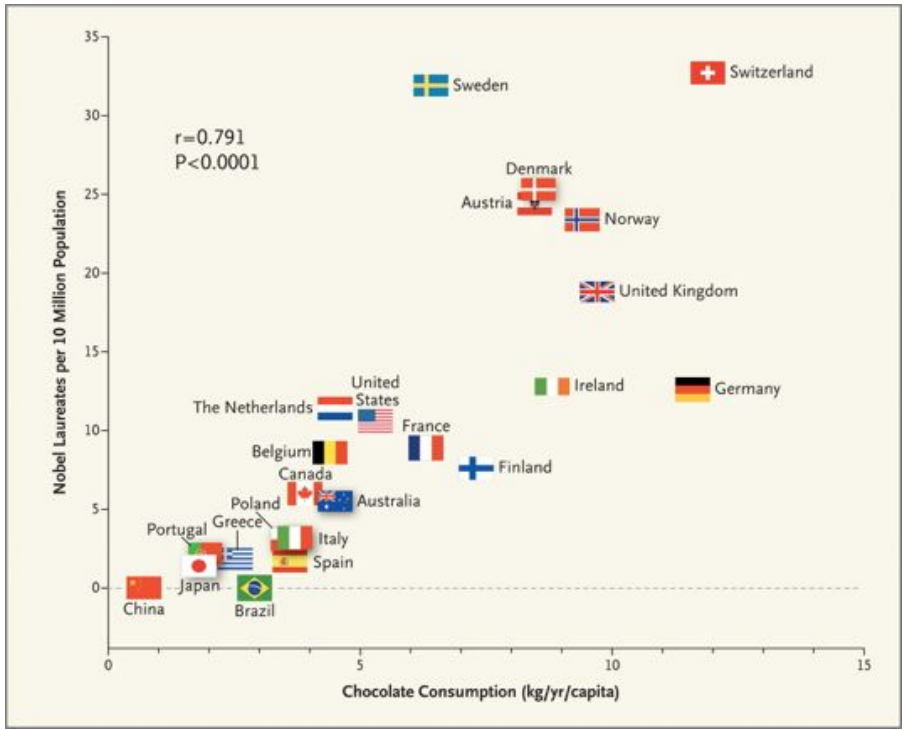
They might actually be from a different **population**.

We did an **observational study**, not an **experiment**

To know that my **treatment** was causal is to use **random assignment**



Consequences of non-random assignment



Chocolate makes you brilliant?

Brilliant people like chocolate?

What else could it be?

Messerili, F. H. (2012) NEJM

Practice Question 3

A study that surveyed a random sample of otherwise healthy adults found that people are more likely to get migraines 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 & migraines.

Can we conclude a **causal** relationship between increased stress and migraines?

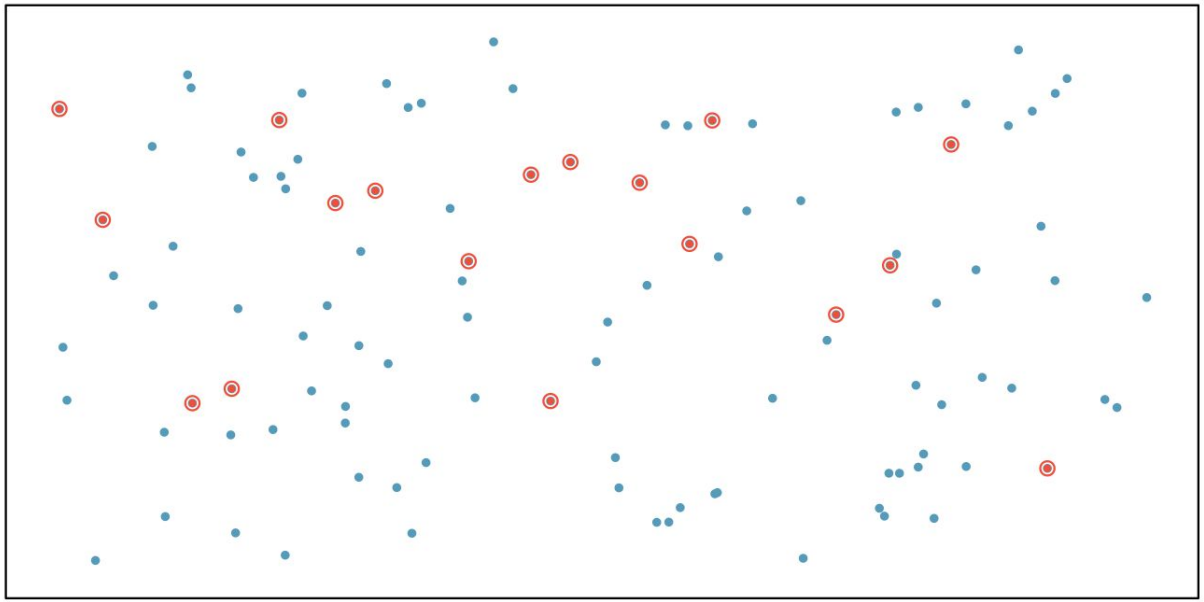
Migraines might also be due to increased caffeine consumption or sleeping less – these are potential **confounding** variables.

Getting good samples

- Almost all statistical methods are based on the notion of implied randomness.
- If observational data are not collected in a random framework from a population, these statistical methods – the estimates and errors associated with the estimates – are not reliable.
- Most commonly used random sampling techniques are **simple**, **stratified**, and **cluster** sampling.

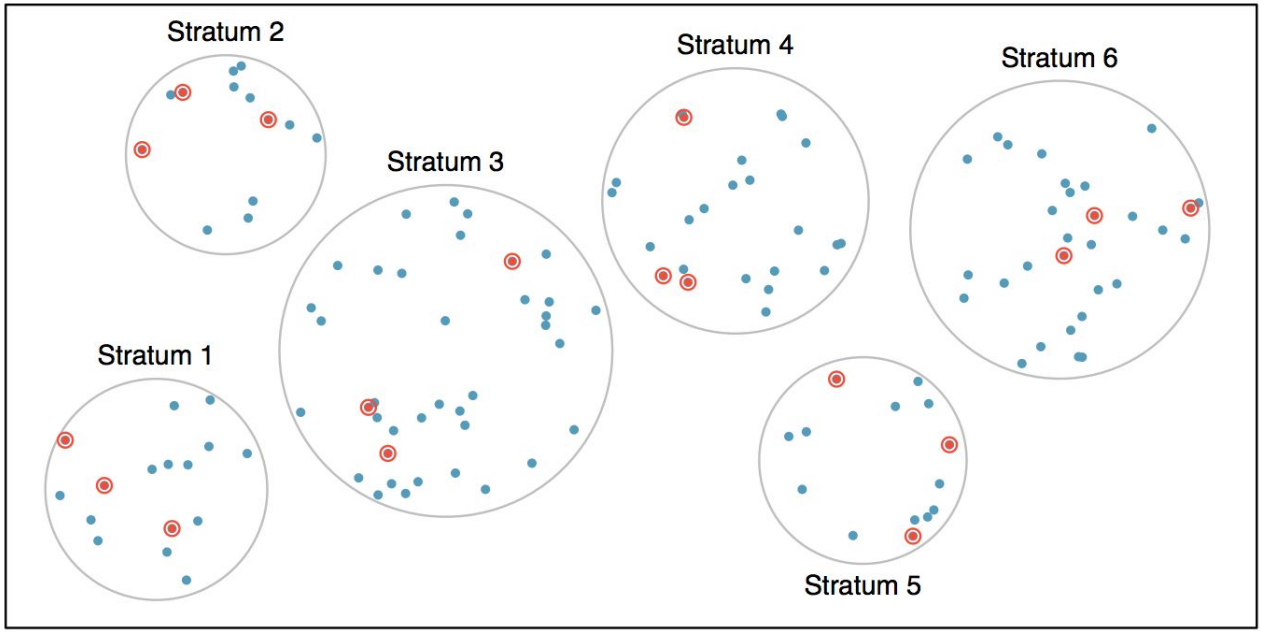
Simple random sample

If there are no dependencies between people,
you can just draw them random from the population



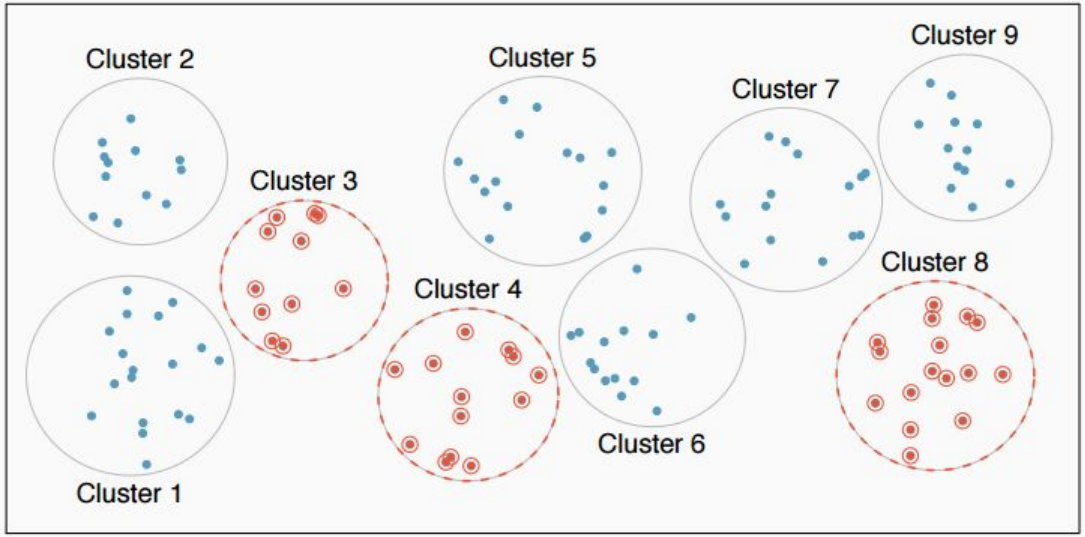
Stratified sample

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



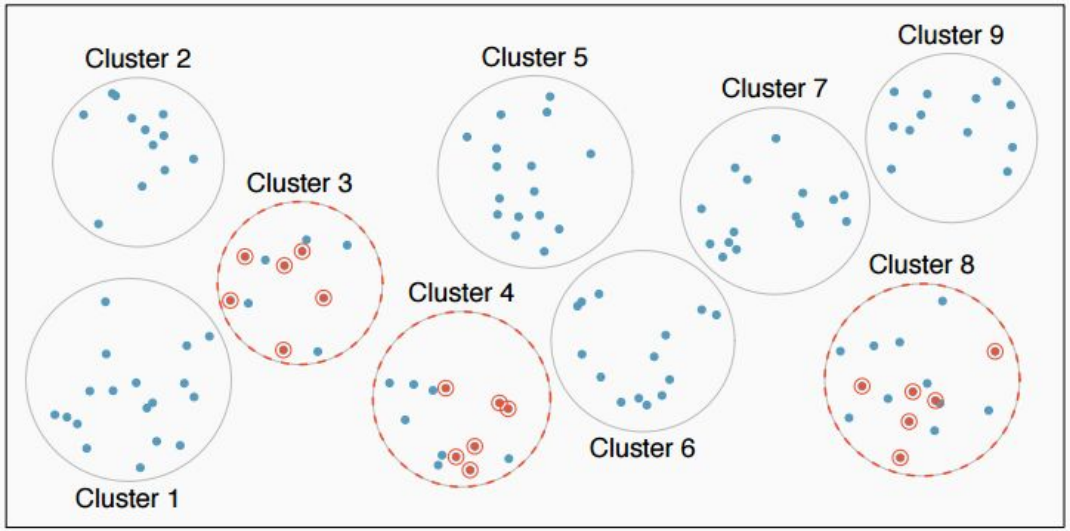
Cluster sample

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.



Multi-stage sample

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



Practice Question 4

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?

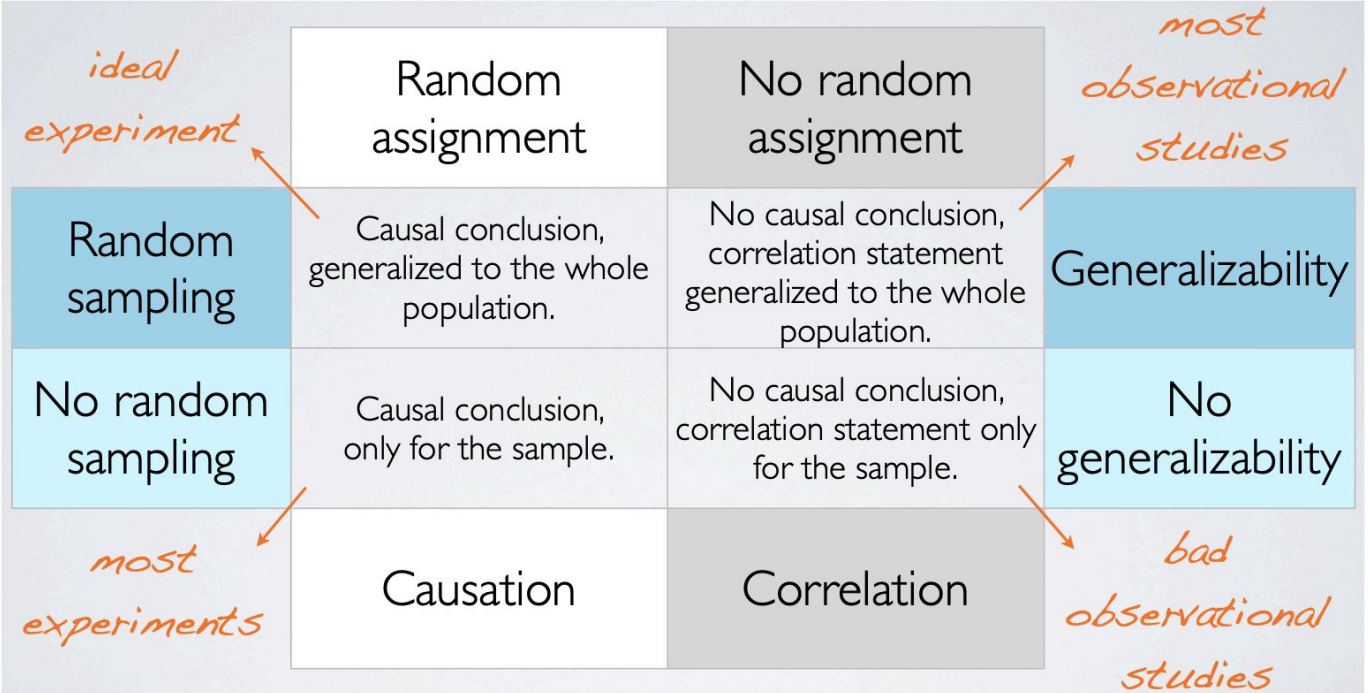
1. Simple random sampling
2. Stratified sampling, where each neighborhood is a stratum
3. Cluster sampling, where each neighborhood is a cluster

Practice Question 4

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?

1. Simple random sampling
2. Stratified sampling, where each neighborhood is a stratum
3. **Cluster sampling, where each neighborhood is a cluster**

Random assignment and random sampling



Key ideas

1. Using samples to make inferences about populations
2. The way you sample your data can change your inferences about the population
3. Experiments use random assignment to treatment groups, observational studies do not
4. Random samples help with **generalizability**, random assignment helps with **causality**

Things to do:

Take the CAOS Test. Due Friday Night!

Start thinking about the homework.

Online Resources

Course Website:
https://dyurovsky.github.io/85309/

- Find syllabus, slides, etc.

Canvas:
https://www.cmu.edu/canvas/

- Submit assignments

Piazza:
piazza.com/cmu/spring2020/85309/home

- Post and answer questions