Sampling and Observational Studies

Review from Last Unit

A sports psychologist wanted to test an intervention using visualization in the training of the athletes at a college. First, the psychologist wanted to get an idea what proportion of athletes were already using visualization as part of their regular training routine. She decided to ask a randomly selected group of 100 athletes if they currently used visualization in their training.

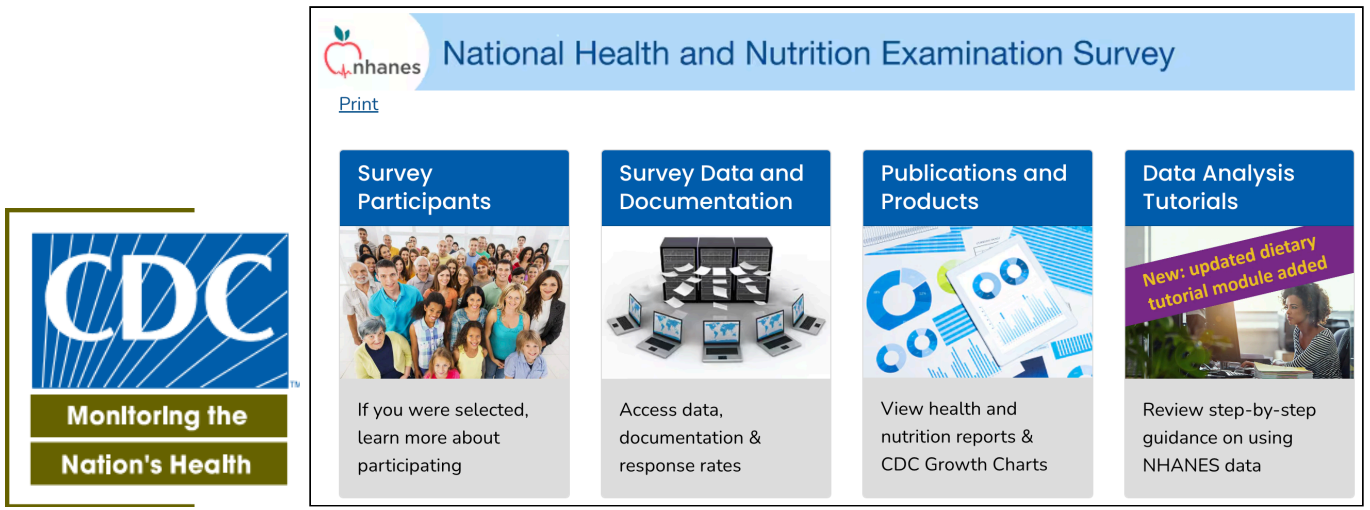
- Population
- Individual
- Parameter
- Sample
- Variable
- Statistic

Review from Last Unit

A sports psychologist wanted to test an intervention using visualization in the training of the athletes at a college. First, the psychologist wanted to get an idea what proportion of athletes were already using visualization as part of their regular training routine. She decided to ask a randomly selected group of 100 athletes if they currently used visualization in their training.

- Population: All athletes
- Individual: An athlete
- Parameter: The proportion (or percent) of all athletes who use visualization
- Sample: The 100 athletes
- Variable: Whether an athlete uses visualization or not
- Statistic: The proportion (or percent) of the 100 athletes who use visualization

A Real Survey



- Goal: Collect health data for a nationally representative sample of the resident civilian noninstitutionalized US population

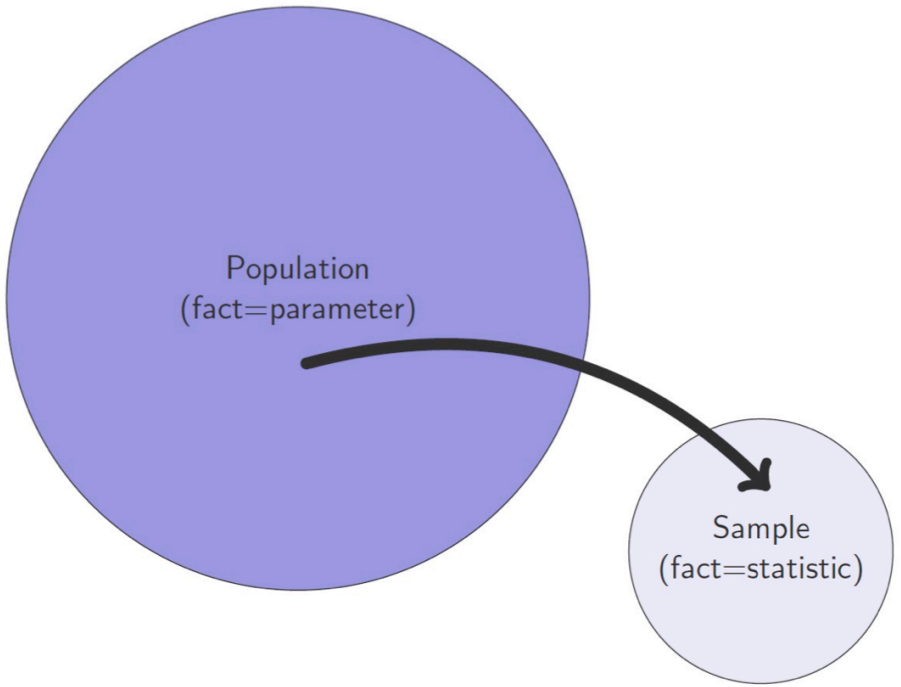
In this unit:

- How do we collect data in an appropriate way?

Two Strategies for Data Collection

1. Observational Studies
    - Sample a group of people and watch/observe their behavior (this subunit)
2. Experiments
    - Recruit a group of people and assign them a treatment (more on this in later subunits)

Why Sampling?



- Population often too big and it's cost effective BUT…
- Sample **must** represent the population (which can be hard)

Example of Sampling

To study the attitude of BYU students towards the beard policy, I stand on the quad and ask students who walk by how they feel about the policy.

- What issue(s), if any, does this sampling technique have?

- This is called **convenience sampling**.

(Another) Example of Sampling

Student ratings via ratemyprofessors.com.

- What issue(s), if any, does this sampling technique have?

- This is called **volunteer response sampling**.

(Another) Example of Sampling

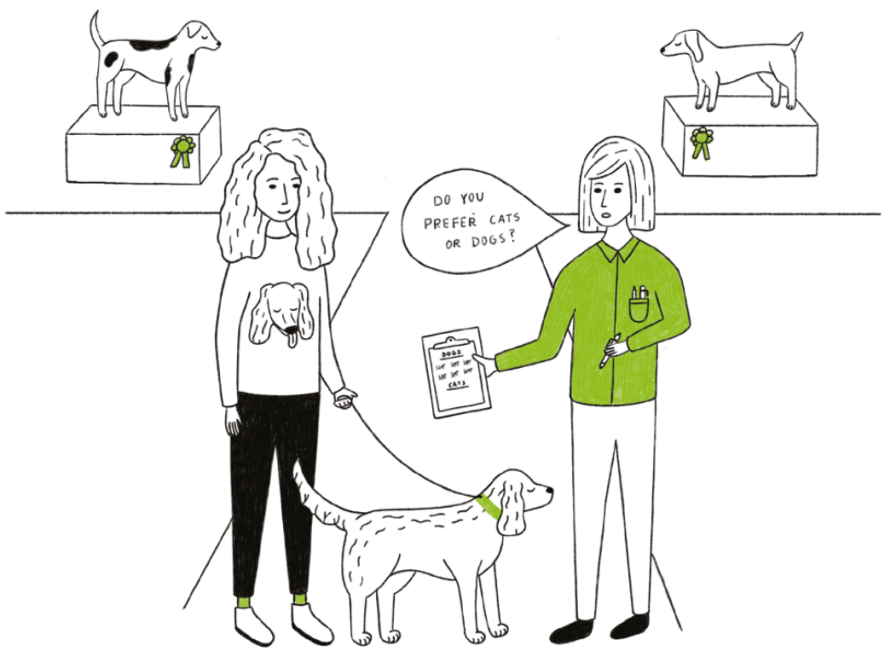
A researcher wants to survey individuals about what smartphone brand they prefer to use. The researcher considers a sample size of 500 respondents. The researcher samples a quota of the first 100 respondents between each of the ages of 16-20, 21-30, 31-40, 41-50, and 51+.

- What issue(s), if any, does this sampling technique have?

- This is called **quota sampling**.
- Issues: non-random sampling, proportions might not match the population

Effective Sampling

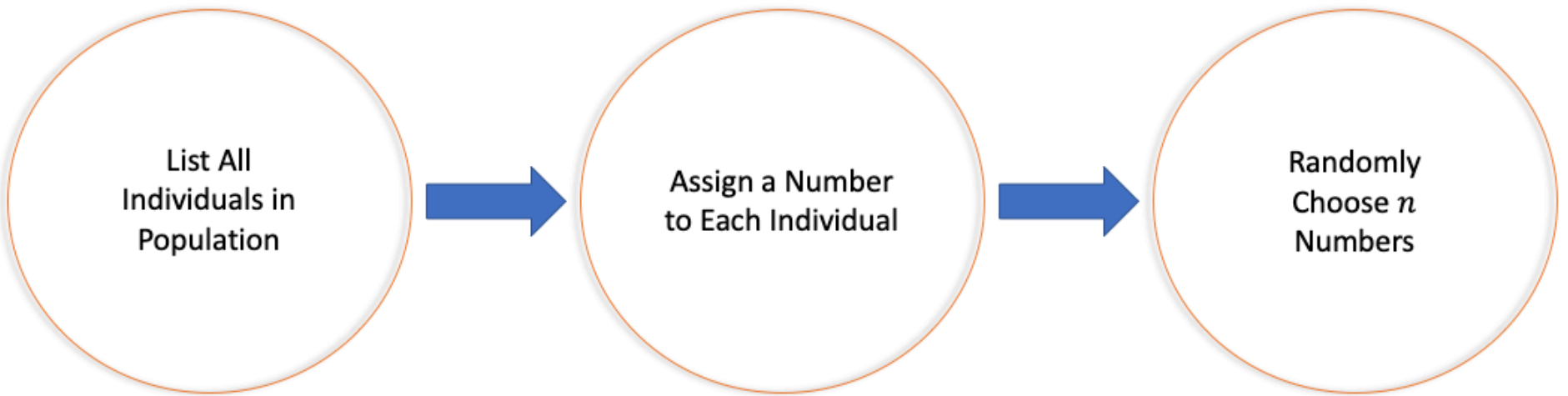
- Sample **must** represent the population
- Sample must not be **biased** (favor certain outcomes)



- Key to effective sampling: randomization

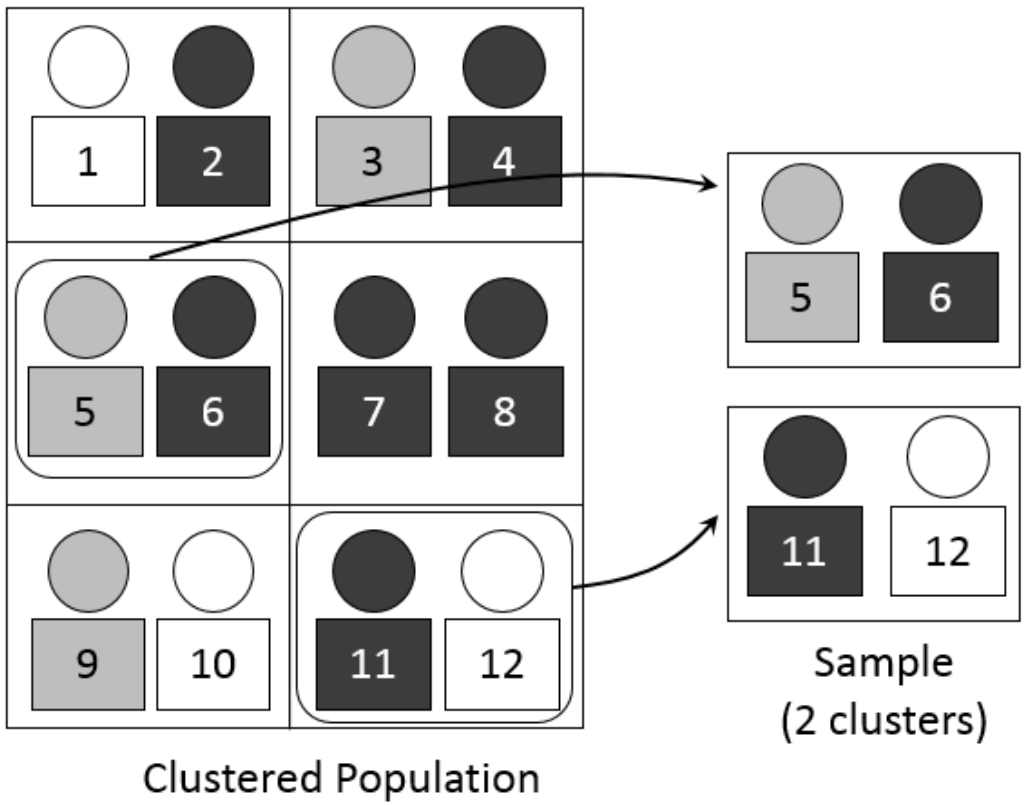
Ways to Effectively Sample

Simple random sample: randomly sample individuals so that all samples of same size have equal chance of being sampled.



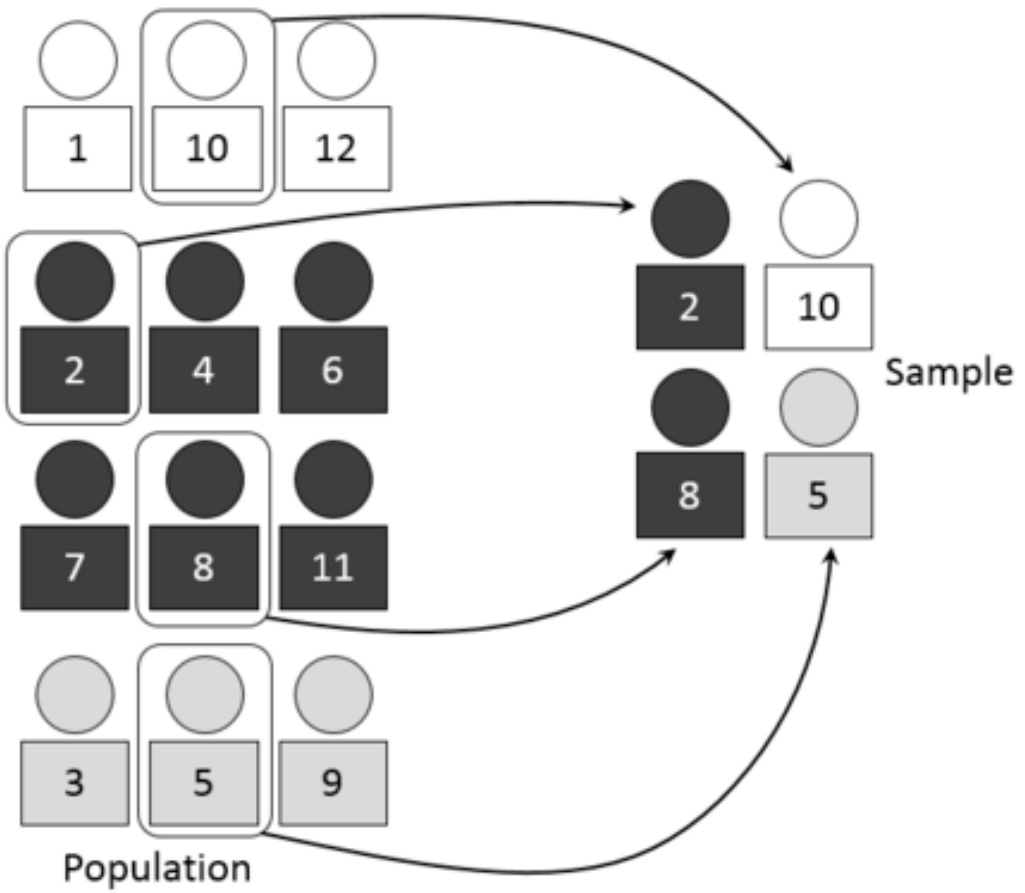
(Another) Way to Effectively Sample

Cluster sample: put population into groups where each group is representative of the population then randomly sample a few groups and include ALL individuals in the selected groups.



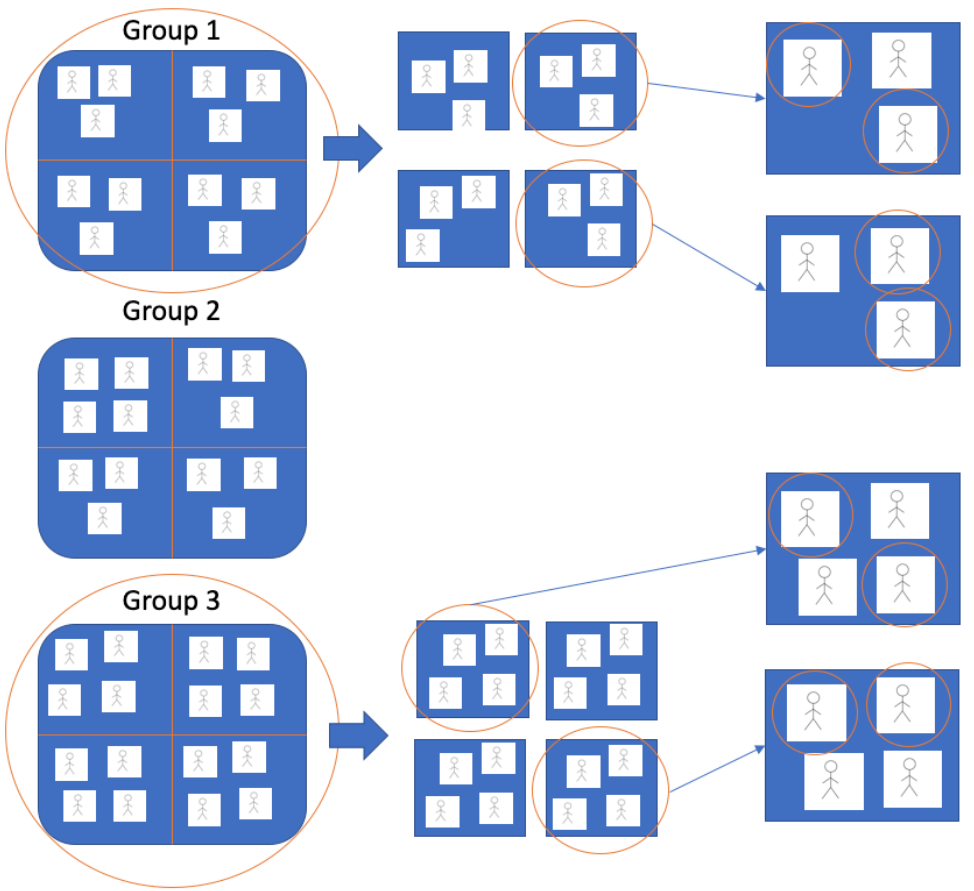
(Another) Way to Effectively Sample

Stratified sample: put population into groups where each group is similar within the group (but different across groups) then randomly sample *some* individuals in each group according to population proportions.



(Another) Way to Effectively Sample

Multi-stage sample: Successively sample groups until you sample individuals.



What type of sampling is this?

BYU creamery wants to find out what the most popular flavor of ice cream is. To do so, they record the next 50 people in each of the age groups of 0-10, 11-20, 21-30, 31-40, 41-50, and 51+. From this information, the creamery infers the most popular ice cream flavor.

- Volunteer Response
- Quota Sampling
- Simple Random Sampling (SRS)
- Cluster Sampling
- Stratified Sampling
- Multi-stage Sample

What type of sampling is this?

BYU creamery wants to find out what the most popular flavor of ice cream is. To do so, they record the next 50 people in each of the age groups of 0-10, 11-20, 21-30, 31-40, 41-50, and 51+. From this information, the creamery infers the most popular ice cream flavor.

- Volunteer Response
- **Quota Sampling**
- Simple Random Sampling (SRS)
- Cluster Sampling
- Stratified Sampling
- Multi-stage Sample

What type of sampling is this?

A researcher at a junior college wants to better understand the proportion of students who would like to attend a 4-year university. The researcher splits the population of interest into 3 groups based on credit hours so that we have 4,000 students with 0-24 hours, 3500 students with 25-48 hours and 2700 students with 49-72 hours. Using 1/5 as the sampling fraction, the researcher randomly selects 800 “first year”, 700 “second year” students and 540 “final year” students.

- Volunteer Response
- Quota Sampling
- Simple Random Sampling (SRS)
- Cluster Sampling
- Stratified Sampling
- Multi-stage Sample

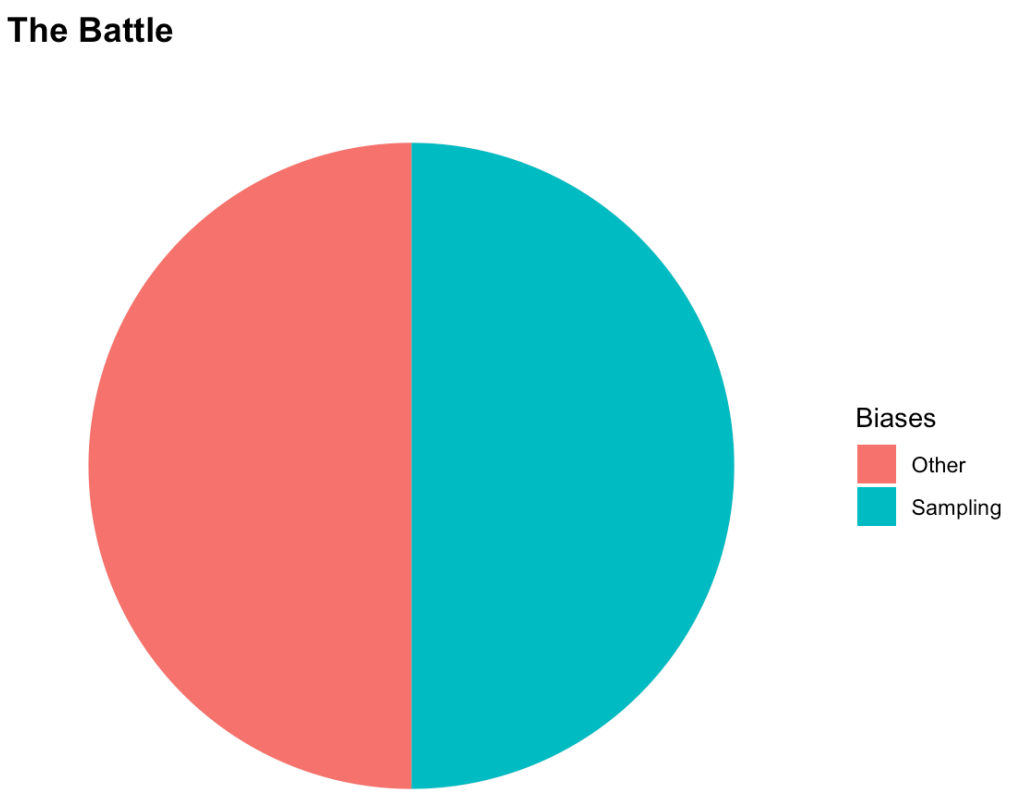
What type of sampling is this?

A researcher at a junior college wants to better understand the proportion of students who would like to attend a 4-year university. The researcher splits the population of interest into 3 groups based on credit hours so that we have 4,000 students with 0-24 hours, 3500 students with 25-48 hours and 2700 students with 49-72 hours. Using 1/5 as the sampling fraction, the researcher randomly selects 800 “first year”, 700 “second year” students and 540 “final year” students.

- Volunteer Response
- Quota Sampling
- Simple Random Sampling (SRS)
- Cluster Sampling
- **Stratified Sampling**
- Multi-stage Sample

Now that we have a sample...

Obtaining a good sample is only 1/2 the battle. We need to be careful how we extract the data from individuals.



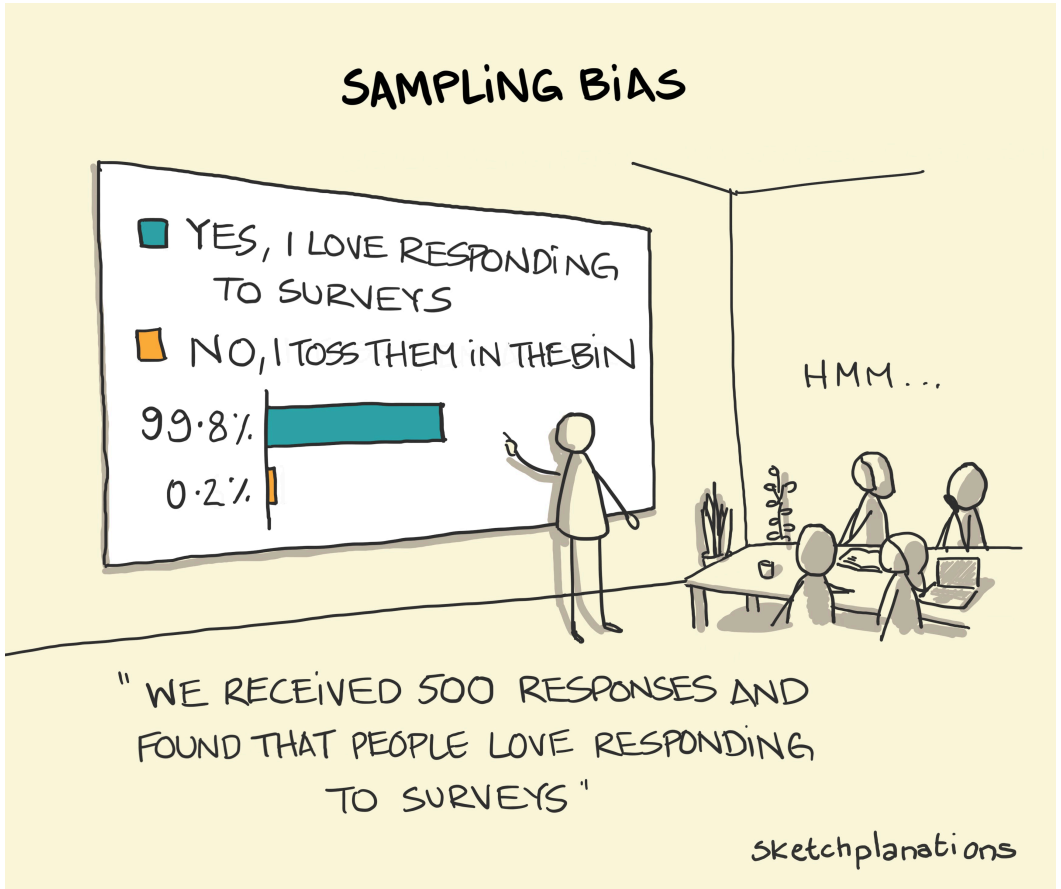
Undercoverage Bias

Undercoverage occurs when some individuals cannot be included in the sample. For example:

- people without a phone
- people without an address
- people with an incorrect email address

Non-response Bias

Select individuals refuse to answer.



Misleading Response

Individuals lie due to sensitive question or leading question.

- Have you ever cheated in a class?
- Do you struggle with mental health issues?
- How great is our hard-working customer service team?
- Doesn’t that make you feel good?

Interviewer Effect

Interviewer influences responses. For example,

- rude
- intimidating to some people
- subtle clues or gestures
- professor asking cheating questions

Question Order Effect

For example,

- What is the most important issue in your company right now?
- Do you approve or disapprove of the way your boss is handling their job?

Question Wording Biases

1. Double negatives
    - Do you oppose not allowing the house to pass Resolution 101?
2. Double barreled questions
    - How would you rate the training and onboarding process?
3. Jargon
    - How was face-time with your customer support rep?
4. Poor scale questions
    - How easy was it to login to the company website? (Yes/No)
    - How often do you check your email in a day? (0-1, 1-2, 2-3, 3+)

Open vs. Closed Questions

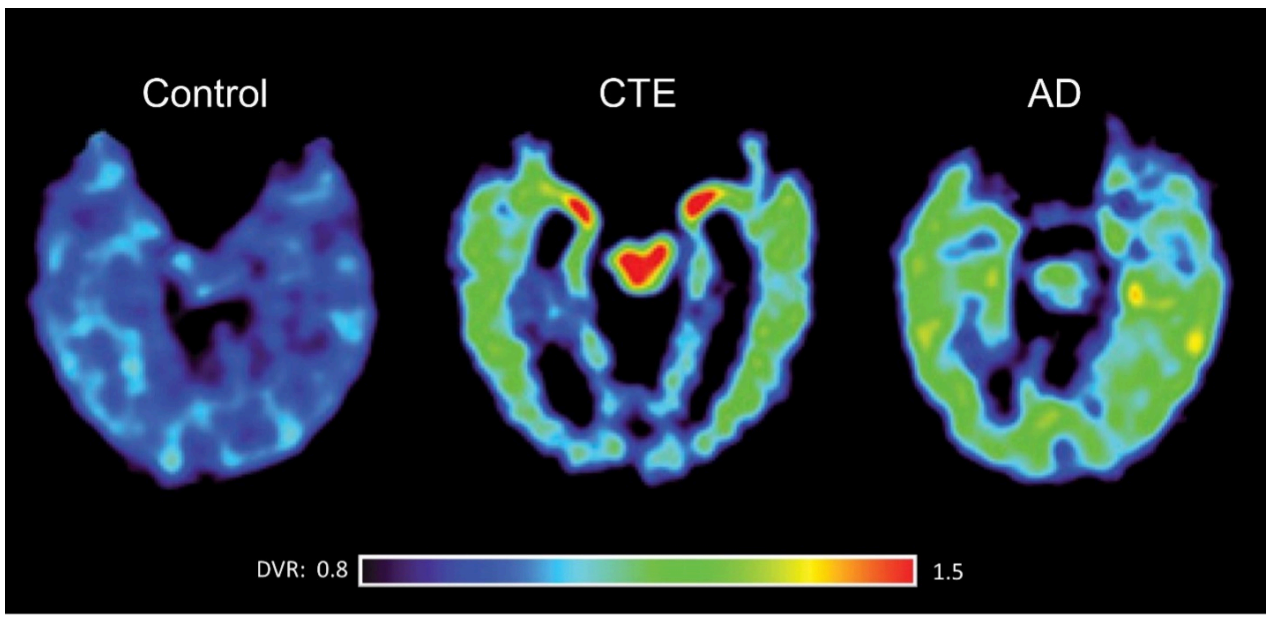
1. Open Questions:
    - What is your favorite kind of music?
    - Advantages: honest answer
    - Disadvantages: Lots of answers, difficult to summarize
2. Closed Questions:
    - Which of the following is your favorite kind of music?
    - Advantages: just a few answers
    - Disadvantages: biased due to not fitting an answer, should include “other”

Data Collection Principles for Observational Studies

1. Samples and statistics from the sample only *approximate* the population and the associated parameters.
2. Differences between the sample statistic and population parameter due to random chance can be accounted for via mathematical probability.
3. Differences between the sample statistic and population parameter due to biases or non-representation **can't** be accounted for.

Discussion Example

CTE (Chronic Traumatic Encephalopothy) is a disorder of the brain that is thought to be caused by multiple head traumas, such as concussions and subconcussive hits. However, causes and frequency of CTE in the general population are not well understood, in part because it can only be diagnosed posthumously by autopsy. In 2017, researchers at Boston University studied CTE in 111 NFL players. These were from specimens donated to the UNITE brain bank over an eight-year period. They found that 99 percent of the brains of former NFL players had CTE.



Discussion Example

CTE (Chronic Traumatic Encephalopothy) is a disorder of the brain that is thought to be caused by multiple head traumas, such as concussions and subconcussive hits. However, causes and frequency of CTE in the general population are not well understood, in part because it can only be diagnosed posthumously by autopsy. In 2017, researchers at Boston University studied CTE in 202 brains of football players, 111 of which played in the NFL. These were from specimens donated to the UNITE brain bank over an eight-year period. They found that 99 percent of the brains of former NFL players had CTE.

- Is this an observational study or experiment and why?
- What potential sources of bias are present in the study?
- How could we use the principles we have learned to design a study to estimate the proportion of former NFL players who have CTE at time of death?
- What would be the population? Parameter? Sample? Statistic? Variable?

Discussion Example

CTE (Chronic Traumatic Encephalopothy) is a disorder…

- Is this an observational study or experiment and why? Observational study because we didn't assign a treatment.
- What potential sources of bias are present in the study? No random sample, biased towards those experiencing CTE symptoms
- How could we use the principles we have learned to design a study to estimate the proportion of former NFL players who have CTE at time of death? Random sample of all past NFL players
- What would be the population? Parameter? Sample? Statistic? Variable? All former NFL players; proportion of former NFL players who have CTE; group of former NFL players who we collect data from; proportion of sampled former NFL players who have CTE; whether a player has CTE or not.

Key Terminology

- Observational Study
- Experiment
- Bad ways of sampling: convenience, quota, volunteer response
- Good ways of sampling: Simple random sample (SRS), cluster sample, stratified sample, multistage sample
- Types of biases: undercoverage, non-response, question order, question wording, interviewer effect, misleading response