UNIT 3: COLLECTING DATA

WHAT IS OUR GOAL FOR UNIT 3?

- Describe an appropriate method for gathering and representing data.
- Observational Studies vs. Experiments

TYPES OF STUDIES

- **Observational Studies (Samples)**
  - Cannot Imply causation: No treatments applied
- **Experiments**
  - Imply Causation: Treatments applied

Retrospective vs. Prospective Studies

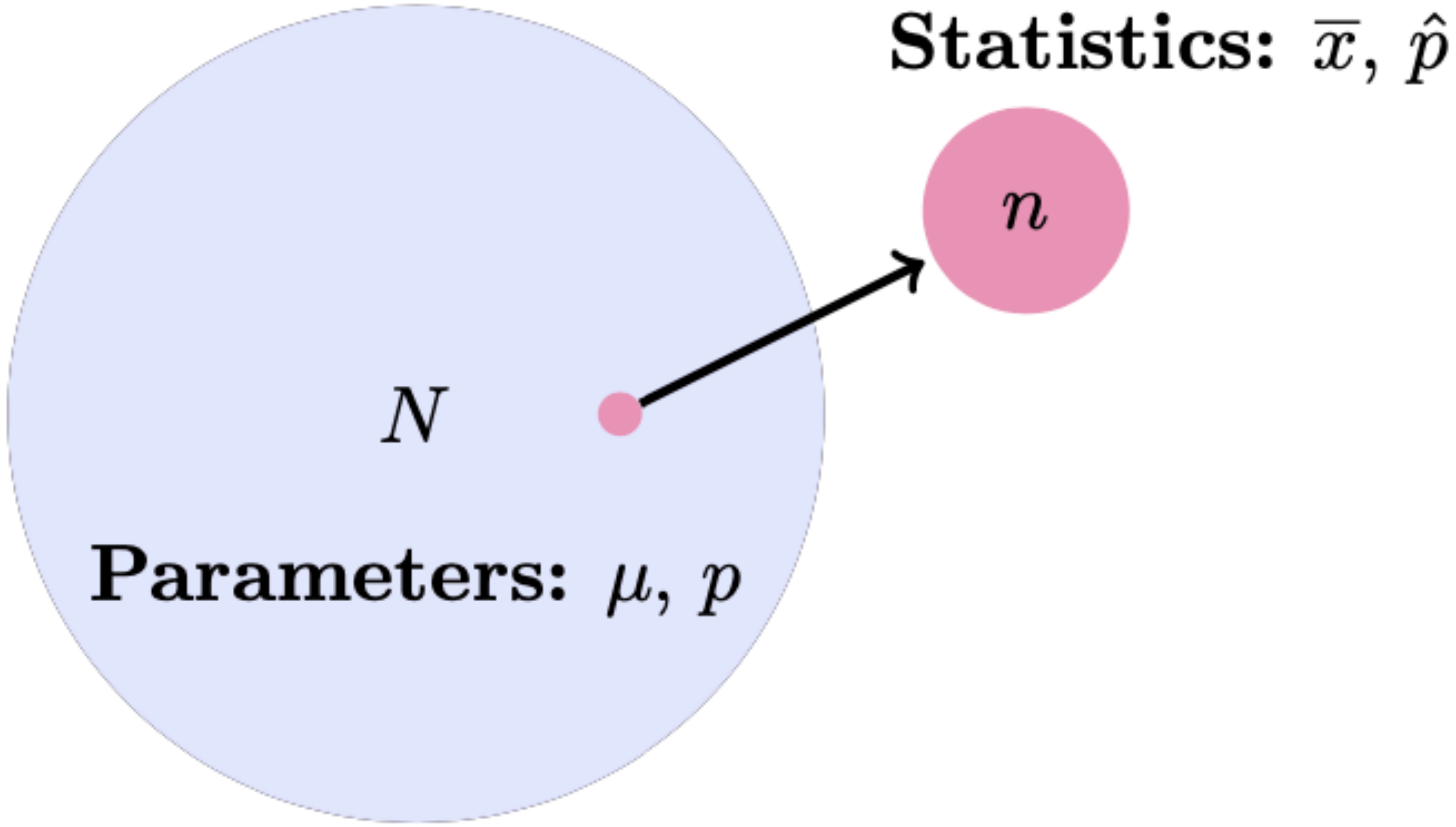
Prospective Studies - Watch for outcomes, track individuals into the future

- Usually have greater accuracy, less susceptible to subject recall error.

Retrospective Studies - Look backward at preexisting data

- Smaller in scale, less expensive, quicker to complete

Population vs. Sample


Statistics: $\overline{x}$, $\hat{p}$
n
N
Parameters: μ, p

Sampling Bias

The sampling Method over/under-predicts the true parameter of interest.

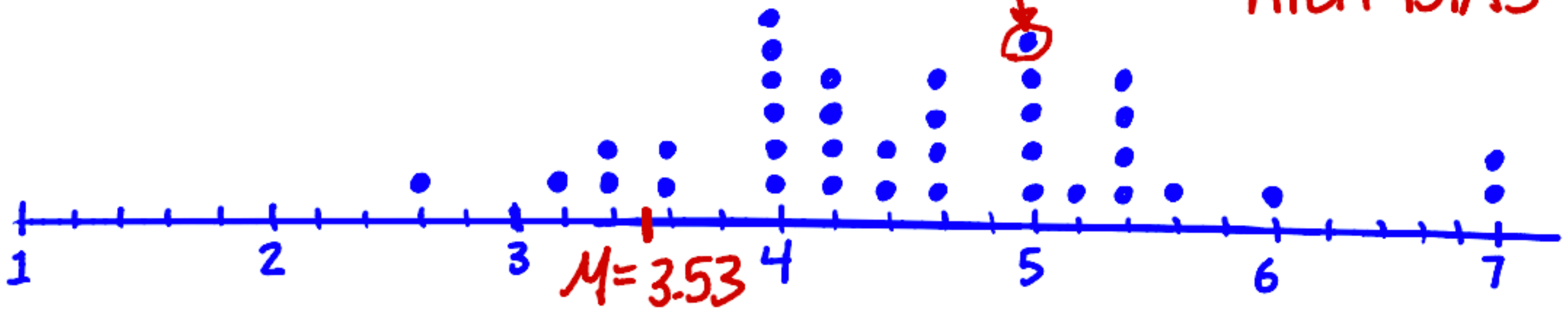
EX: Beyonce

3. Put your average on the dotplot on the white board at the front of the room. Copy the class dotplot below.

average word length from one sample

convenience sample

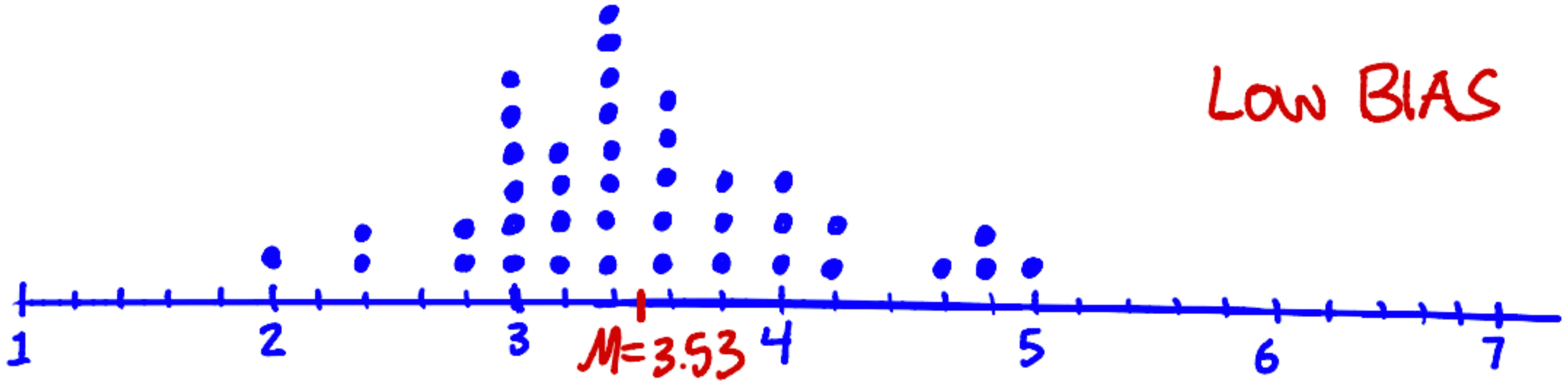
HIGH BIAS



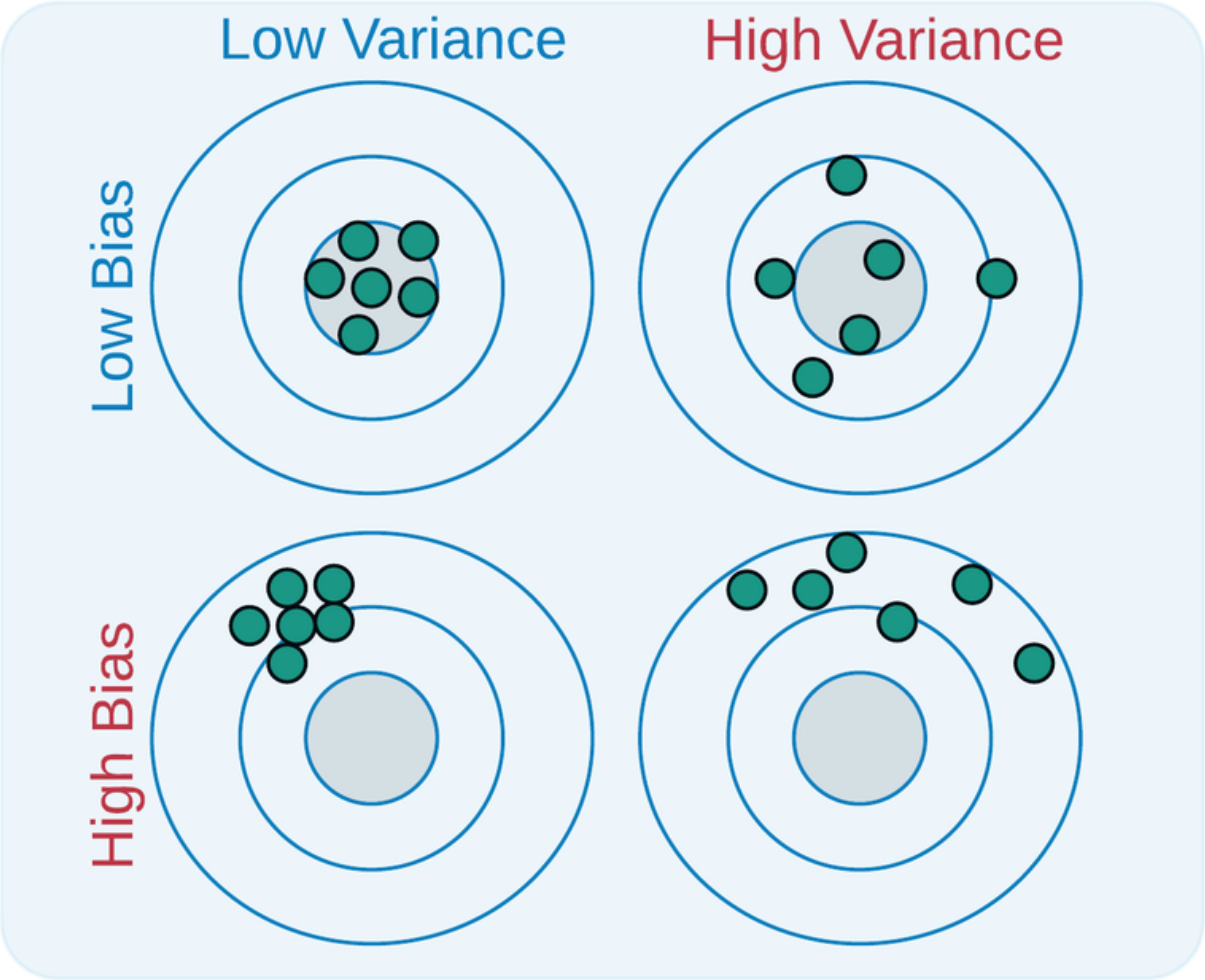
4. Find a new sample of 5 words using a random number generator. Put your average on the dotplot on the white board at the front of the room. Copy the class dotplot below.

simple random sample (SRS)

LOW BIAS



Sampling Bias and Variability



Bias

Voluntary Response - People voluntarily answer, but the volunteers often has a specific reason for volunteering.

Convenience Samples - Convenience leads to systematic pattern.

Under-coverage - Some part of the population has not be covered in the samples design.

Non-Response - People from the sample simply don’t respond

Sampling Methods (2 Examples)

Example 1 - A political scientist is interested in the proportion of Albertans who will vote NDP in the upcoming election.

Example 2 - A sloth researcher would like to estimate the lifespan of the three toed sloth.

Simple Random Sample

1. Label individuals in population from 1-N, or write individuals in population on separate cards

2. Randomly select n numbers from 1-N, or shuffle all cards in a hat and randomly pull n of them.

Benefits: Usually simple the implement. Minimal advanced knowledge of population. Allows us to make generalizations from sample to population

Disadvantages: Finding all subjects can suck/be impossible. Difficult to execute for large populations.

TO GET FULL MARKS ON THE TEST

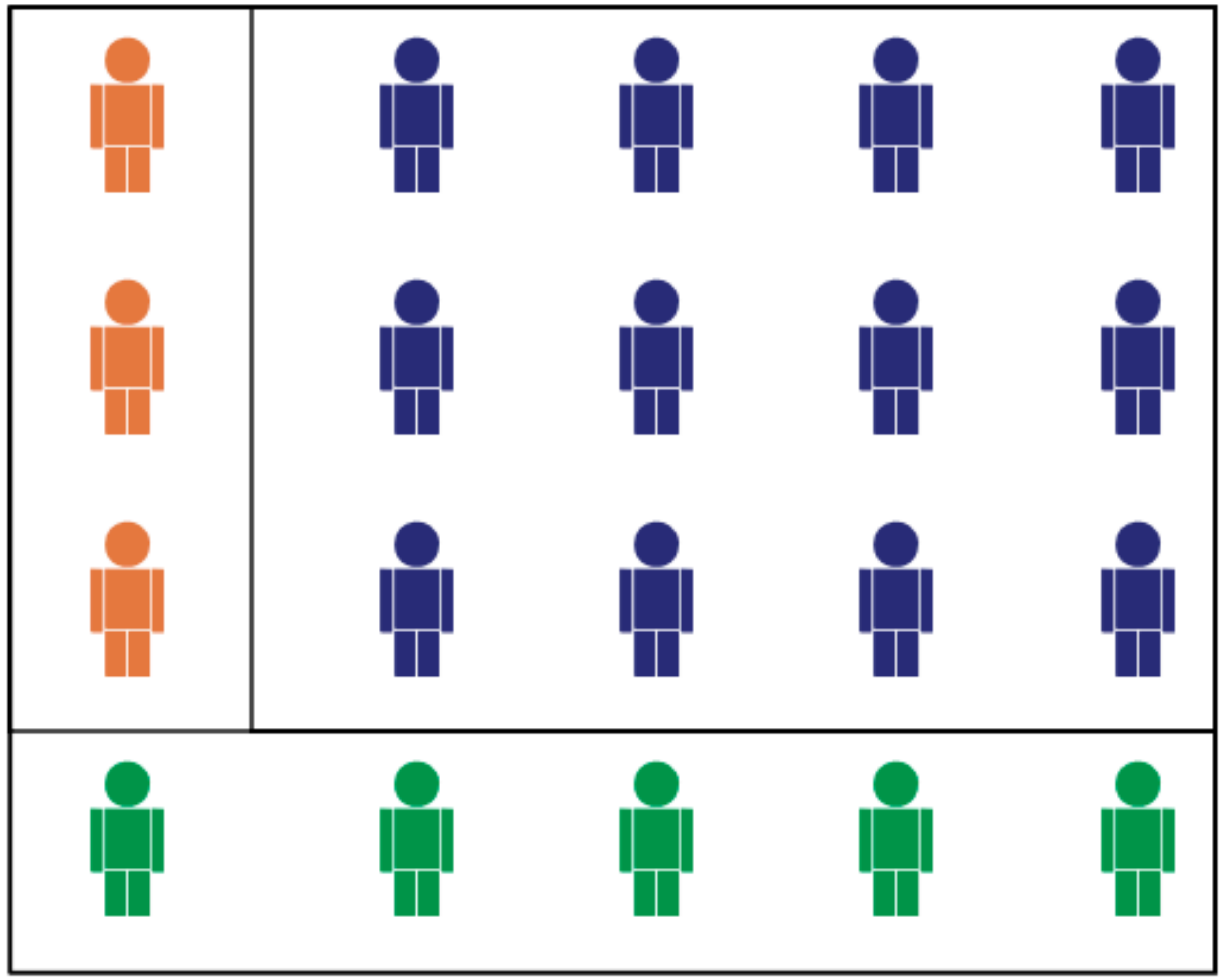
1. **Assigning Numbers**
2. **Using. Random number generated to generate distinct numbers in a given range**
3. **Linking selected numbers with corresponding individuals in the sample**

Simple Random Sample

Example 1 - Here we would look at the list of all N voters, assigning each voters a number from 1-N. Then use a random number generator to select n numbers that are used to select n individuals from the population.

Example 2 - Here we would need every living three toed sloth in existence to have a unique tag, or i.d. number. We would then generate random tag numbers and select the sloths to sample. Does this seem very practical? Not only is tagging all sloths probably absurd, what would be the point in taking a sample if we already had tagged and collected data on every single sloth?

Stratified Random Sample



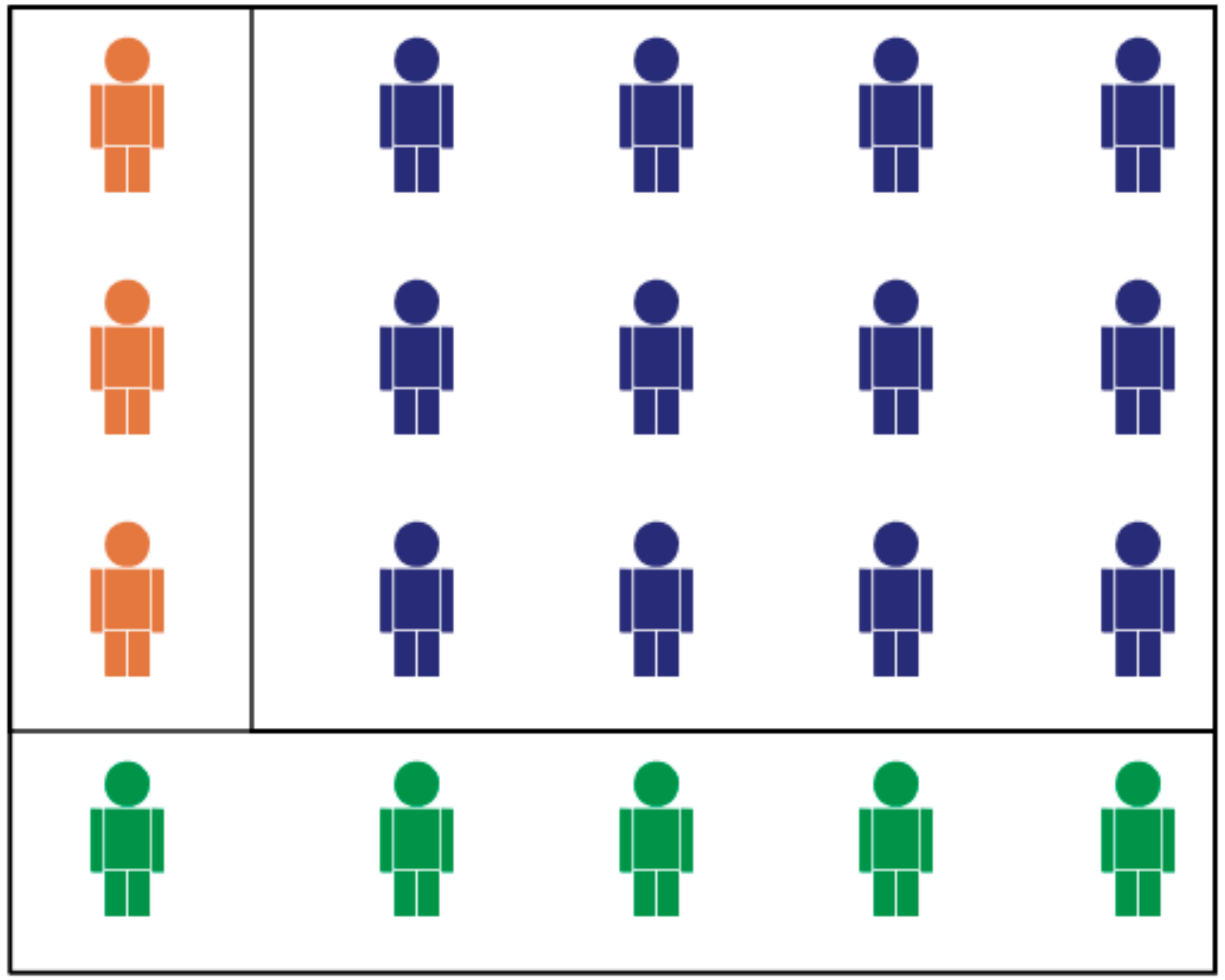
1. Partition population into homogenous groups called ‘strata’.

2. Take a simple random sample from each stratum.

Benefits: Can reduce variability of sampling distribution. More precise estimation.

Disadvantages: Difficult and expensive to implement in many circumstances. Cannot force homogeneous subdivisions.

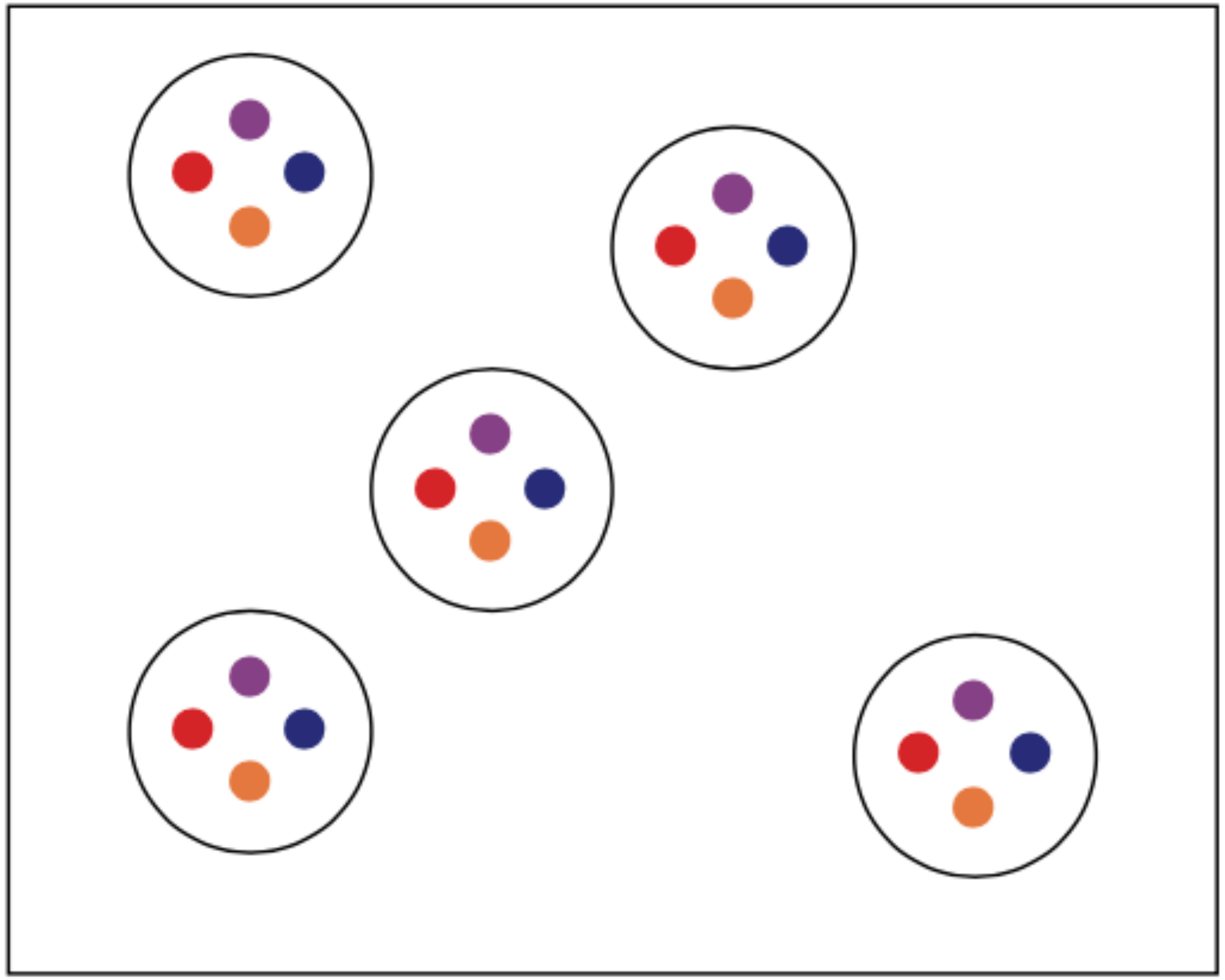
Stratified Random Sample



Example 1 - We can divide Albertan’s up by their registered voting status, and then take a simple random sample from each party.

Example 2 - We could divide the sloth population up into two strata; male and female sloths. We would then take a simple random sample from male sloths, and from female sloths. This is also a fairly contrived example. It would be easier if we were studying a less mysterious animal.

Cluster Random Sample

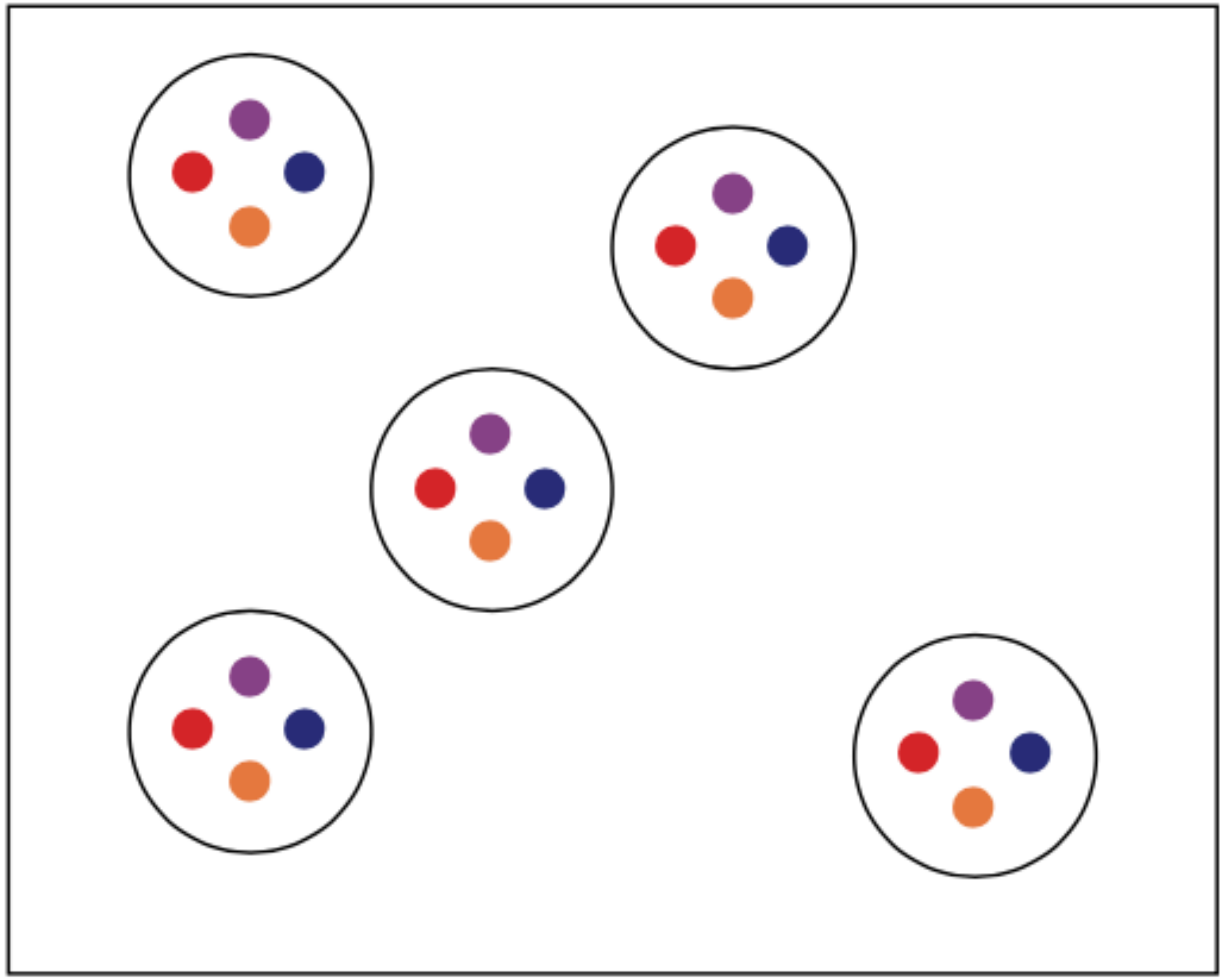


In cluster random sampling we sample naturally occurring clusters that have **heterogeneous** makeup. We randomly sample several clusters and sample **every single point from within**.

Benefits: If clusters exist it can make sampling really easy. Can often allow for a larger sample size since entire clusters are sampled

Disadvantages: Often has high variability due to patterns within clusters (AP Central Video). Clusters may not be truly heterogeneous.

Cluster Random Sample



Example 1 - Here may have a several small town, or counties across Alberta. We randomly select several of them and sample every person within. It's important to note that the clusters are heterogeneous in makeup.

Example 2 - Here we select several regions containing three toed sloths, and sample every sloth in the region. (Probably the most effective method so far).

Systematic Random Sample

In systematic random sampling a random point is selected from the population, and then every k^{th} point is selected until the desired sample size is reached.

Benefits: Really cheap easy to implement on a list of data. Can produce a representative sample for a list that does not exhibit patterns.

Disadvantages: Since we are introducing a systematic pattern to our sampling method, this can lead to BIAS if our list is grouped in certain ways.

Systematic Random Sample

Example 1 - Here we would need a list of every voter in Alberta. We would select a random voter and then select every kth th voter. Note that k here would likely be found using $\frac{N}{n}$ so sample points are spread over population.

Example 2 - For our sloth example, this sampling example is again quite unrealistic as we would need a list of every sloth in existence. If it could be obtained we would follow the same procedure as the last example.

Biased Sampling Methods

Convenience Sample: Sample points that are convenient to access

Example 1 - Call up the first n numbers you see in a phone book.

Example 2 - Simply sample the first n sloths you can find while hiking through a jungle. Why were these easy to sample? Slow? Sick?

Biased Sampling Methods

Voluntary Response Sample: Sample points ‘volunteer’

Example 1 - Put a political advertisement in paper to gather sample points. What kind of people might respond to this add?

Example 2 - Voluntary sampling doesn’t really make sense in this context. Do you sloths find you and asked to be sampled? Maybe you’re in Zootopia...

Homework

Read Barron’s Pages 147 - 153

Barron’s Quiz 10 - Quiz 11

Example 1: Last year a school offered an after school SAT prep class that students could volunteer to take. 44 students took the course and then took the SAT. The average SAT score for this group was 1220. The average SAT score for all students who did not take the prep class was 1050.

Is the situation described an observational study or experiment?

Example 1: Last year a school offered an after school SAT prep class that students could volunteer to take. 44 students took the course and then took the SAT. The average SAT score for this group was 1220. The average SAT score for all students who did not take the prep class was 1050.

Is this an observational study or experiment?

No Treatments Imposed, Observational Study

Example 1: Last year a school offered an after school SAT prep class that students could volunteer to take. 44 students took the course and then took the SAT. The average SAT score for this group was 1220. The average SAT score for all students who did not take the prep class was 1050.

Identify the Explanatory and Response Variables

Example 1: Last year a school offered an after school SAT prep class that students could volunteer to take. 44 students took the course and then took the SAT. The average SAT score for this group was 1220. The average SAT score for all students who did not take the prep class was 1050.

Identify the Explanatory and Response Variables

Explanatory: Whether or not student took SAT Course

Response: SAT Score

Example 1: Last year a school offered an after school SAT prep class that students could volunteer to take. 44 students took the course and then took the SAT. The average SAT score for this group was 1220. The average SAT score for all students who did not take the prep class was 1050.

Can you conclude that taking the prep course will cause a student's SAT score to increase? Why or why not?

Example 1: Last year a school offered an after school SAT prep class that students could volunteer to take. 44 students took the course and then took the SAT. The average SAT score for this group was 1220. The average SAT score for all students who did not take the prep class was 1050.

Can you conclude that taking the prep course will cause a student's SAT score to increase? Why or why not?

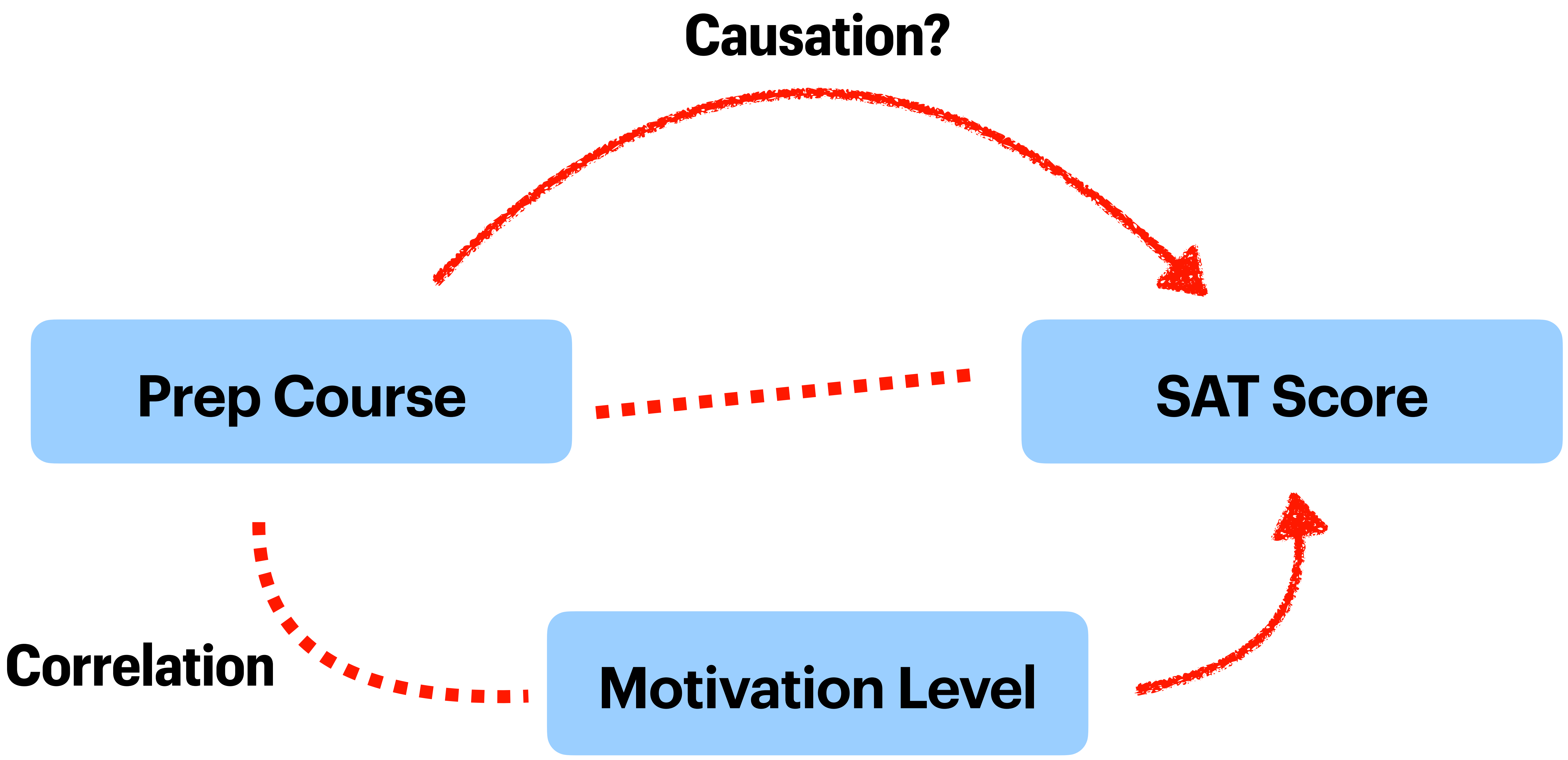
No, observational studies cannot imply causation. Students who take the prep course may be highly motivates so they study more anyways. Need an experiment to show causation.

Example 1: Last year a school offered an after school SAT prep class that students could volunteer to take. 44 students took the course and then took the SAT. The average SAT score for this group was 1220. The average SAT score for all students who did not take the prep class was 1050.

Identify as many other possible variables that you can that may explain why the SAT scores are higher for those who took the prep course than those who did not.

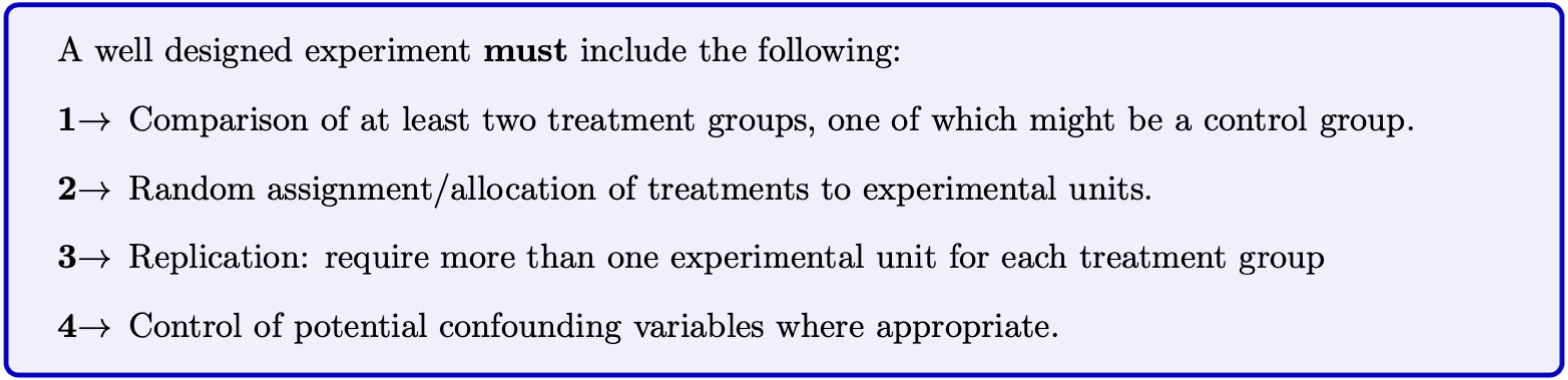
Identify as many other possible variables that you can that may explain why the SAT scores are higher for those who took the prep course than those who did not.

- Motivation Level
- College Bound
- Time After School
- Number of AP Classes



Example: Last year a school offered an after school SAT prep class that students could volunteer to take. 44 students took the course and then took the SAT. The average SAT score for this group was 1220. The average SAT score for all students who did not take the prep class was 1050.

Now let's look at an EXPERIMENT, that could be used to determine causation



A well designed experiment **must** include the following:

1→ Comparison of at least two treatment groups, one of which might be a control group.

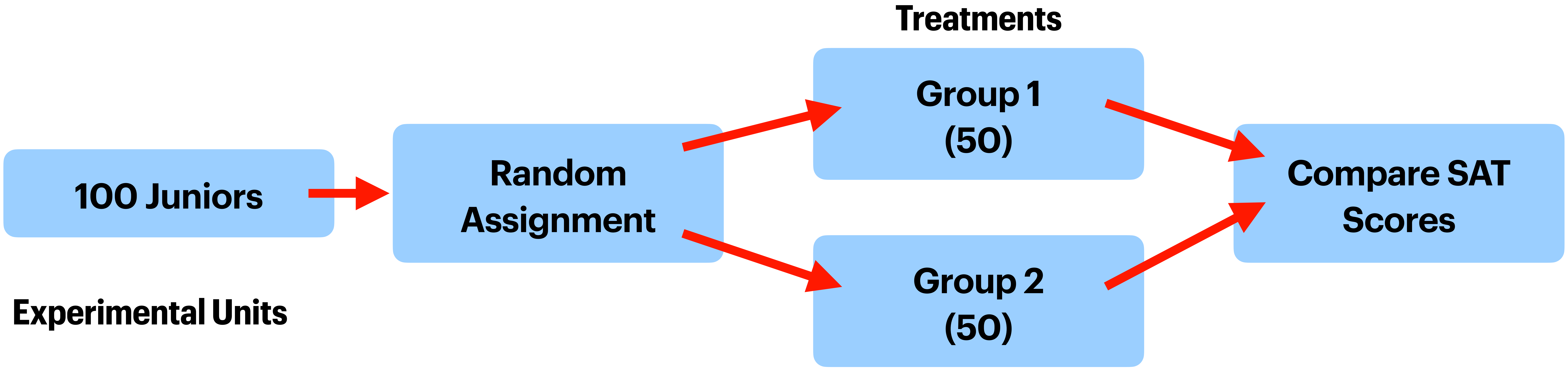
2→ Random assignment/allocation of treatments to experimental units.

3→ Replication: require more than one experimental unit for each treatment group

4→ Control of potential confounding variables where appropriate.

Completely Randomized Design

Randomly choose 100 juniors to be part of the experiment. Randomly split the group so half take the prep course and half do not. One month later have all the students take the SAT and compare scored between the two groups.



Random assignment: put 100 cards in box half say ‘course’, the other half say ‘no course’. Have each subject draw a card without replacement until all cards are assigned. Students with ‘course ‘ card take course, students with ‘no course’ cards do not take course.

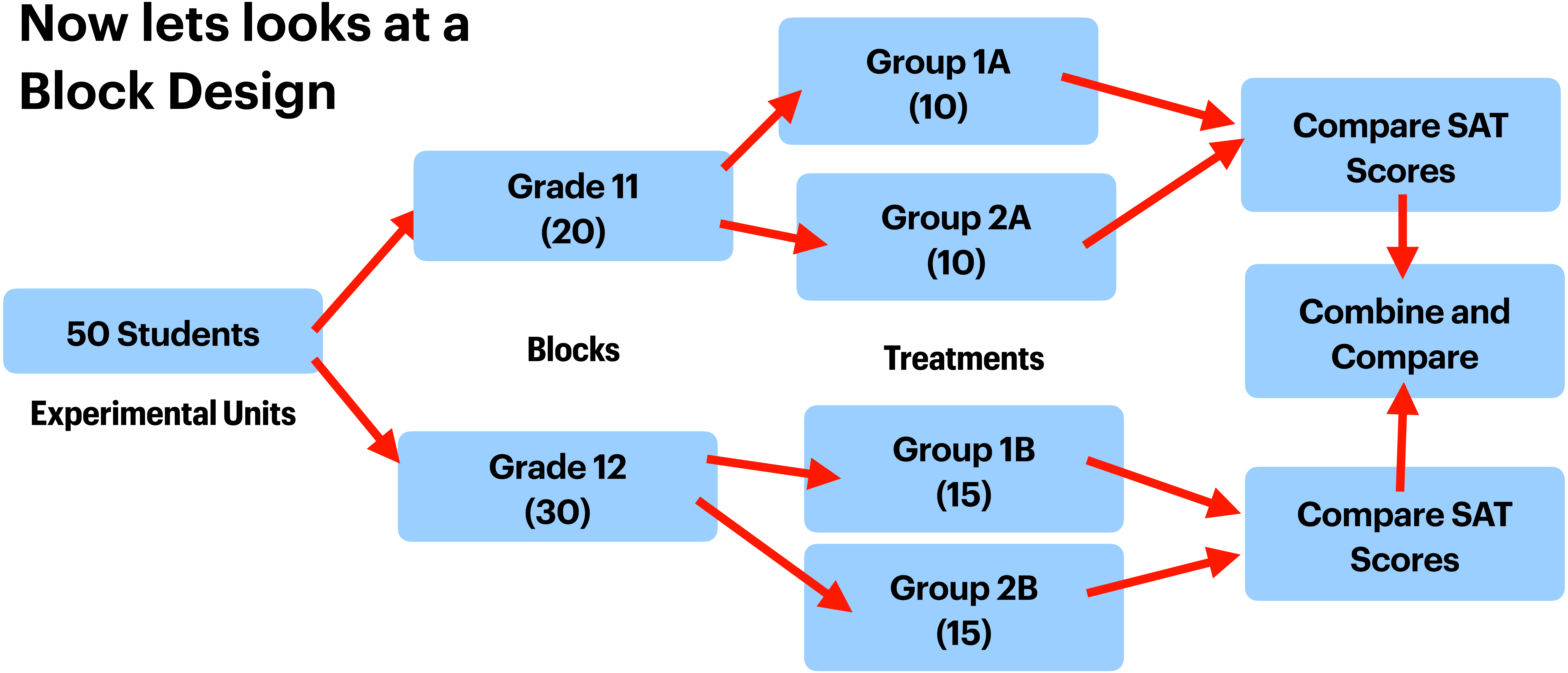
Example 2: Now Suppose a Diploma prep class is offered in two different formats: online or classroom teacher. The counsellors want to know which teaching method will yield higher Diploma scores so they have allowed us to set up an experiment. 50 students have signed up to take some form of the Diploma prep class. (20 Grade 11 and 30 Grade 12)

Write how you would conduct a Completely Randomized Design

Completely Randomized Design

Treatments
50 Students
Random Assignment
Group 1 (25)
Group 2 (25)
Online Course
Teacher Course
Compare SAT Scores
Experimental Units

Now lets looks at a Block Design



50 Students

Experimental Units

Grade 11 (20)

Grade 12 (30)

Blocks

Group 1A (10)

Group 2A (10)

Group 1B (15)

Group 2B (15)

Treatments

Compare SAT Scores

Compare SAT Scores

Combine and Compare

Example 3: The counsellors are now concerned that GPA is certainly going to affect a students Diploma scores. Looking only at the 30 grade 12 students how could we conduct an experiment so that the GPA's are evenly distributed among treatment groups.

Now Let's Look at the Matched Pairs Design

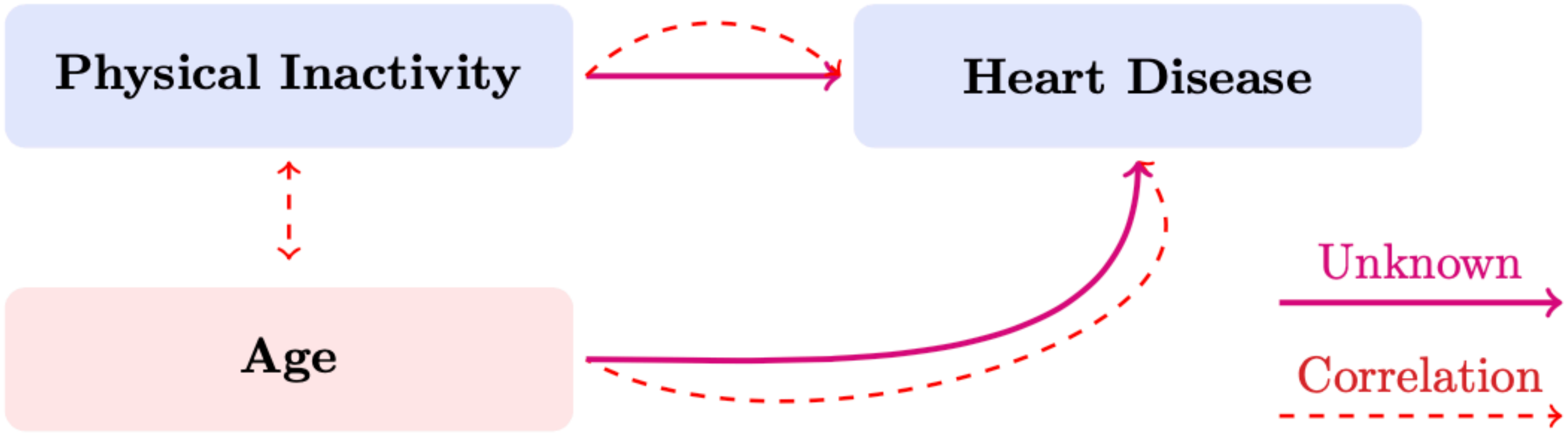
Put the 30 grade 12 students in order from highest to lowest GPA. Take two students with the highest GPA and pair them. Flip a coin to assign one to the online class and one to the teacher class. Repeat this process with the next two highest GPA's until all 30 students are assigned.

Blinding

- **Single Blind Experiments:** This is where the experimental units are not told which treatment they are receiving.
- **Double Blind Experiments:** This is where experimental units, and researchers do not know which treatment group is which. It might be a good idea to blind the researcher to avoid confirmation bias.
- **Confirmation Bias:** This is where a researcher comes to conclusions because it verifies their own beliefs. It is easy for our brains to connect dots using ‘intuition’ when we are trying to make explanations.

Example 3: In an observational study we notice that low physical activity level seems to be associated with heart disease. Our study also included a variety of different variables, could any of them be confounded with physical inactivity?

Example 3: In an observational study we notice that low physical activity level seems to be associated with heart disease. Our study also included a variety of different variables, could any of them be confounded with physical inactivity?

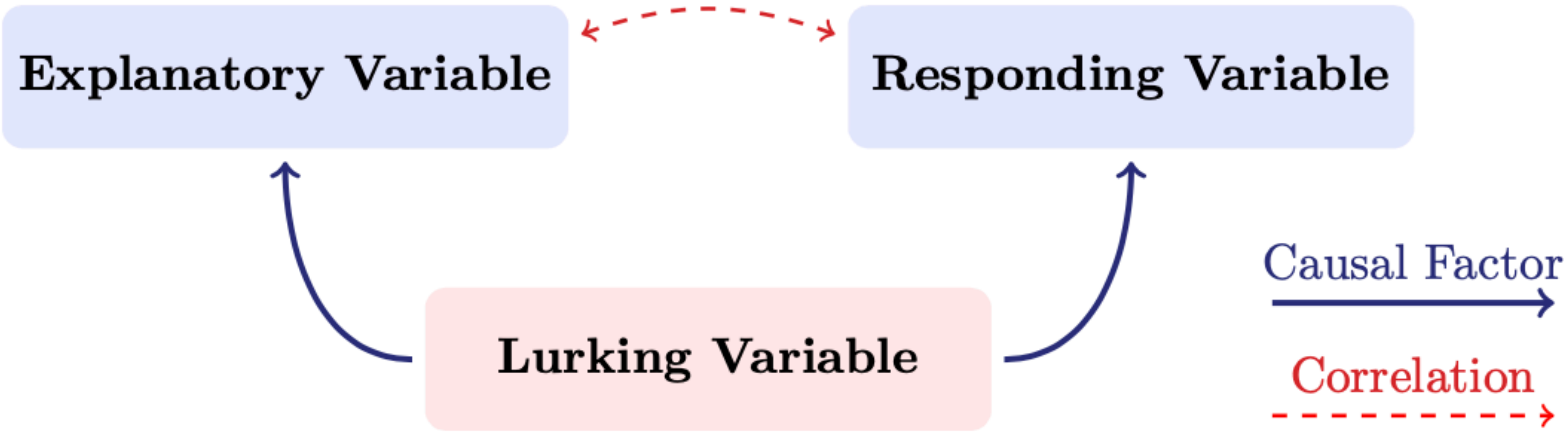


Placebo

A treatment that resembles another but with no active ‘ingredients’ at play. Sometimes placebos can give a psychological effect. This is the idea when Harry Potter gives Ronald Weasley a fake vial of ‘liquid luck’.

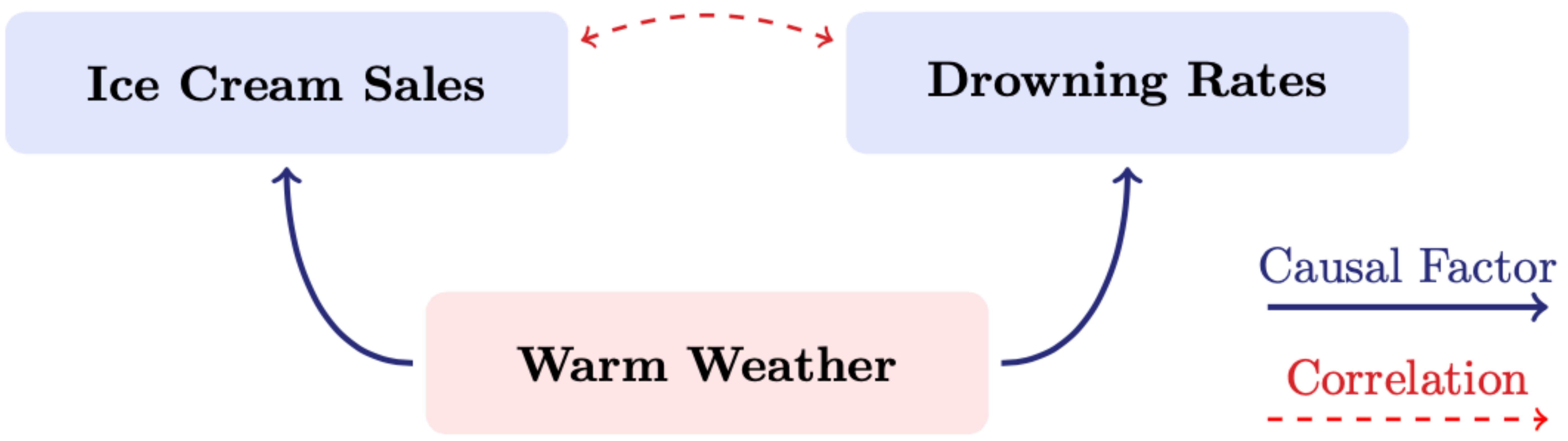
-

Lurking Variables: These variables can affect the response variable, but are not accounted for in the studies design. They 'lurk' without being known by the researcher and cause changes in both the explanatory and response variables.



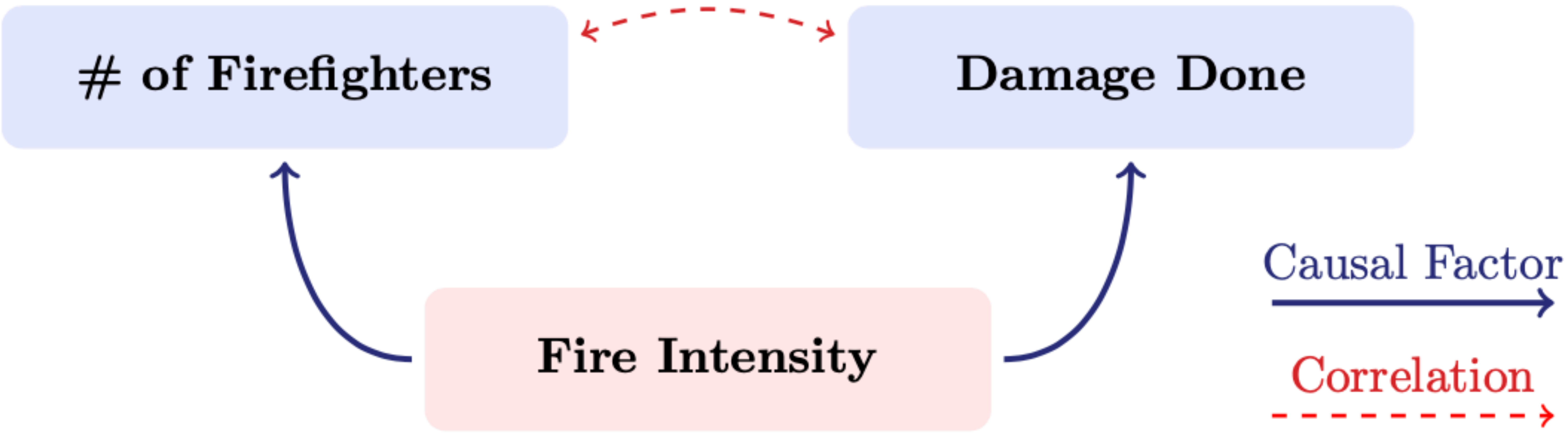
Example 1: Suppose that in a research study you are exploring two variables. Ice cream sales, and rates of drowning. You discover that ice cream sales seems to be associated with drowning rates.

Example 1: Suppose that in a research study you are exploring two variables. Ice cream sales, and rates of drowning. You discover that ice cream sales seems to be associated with drowning rates.



Example 2: The number of firefighters at a fire, and the amount of damage done are highly correlated. Is there a lurking variable at play here?

Example 2: The number of firefighters at a fire, and the amount of damage done are highly correlated. Is there a lurking variable at play here?


# of Firefighters
Damage Done
Fire Intensity
Causal Factor
Correlation