# Lecture 8
# Types of Studies

# Review From Monday:

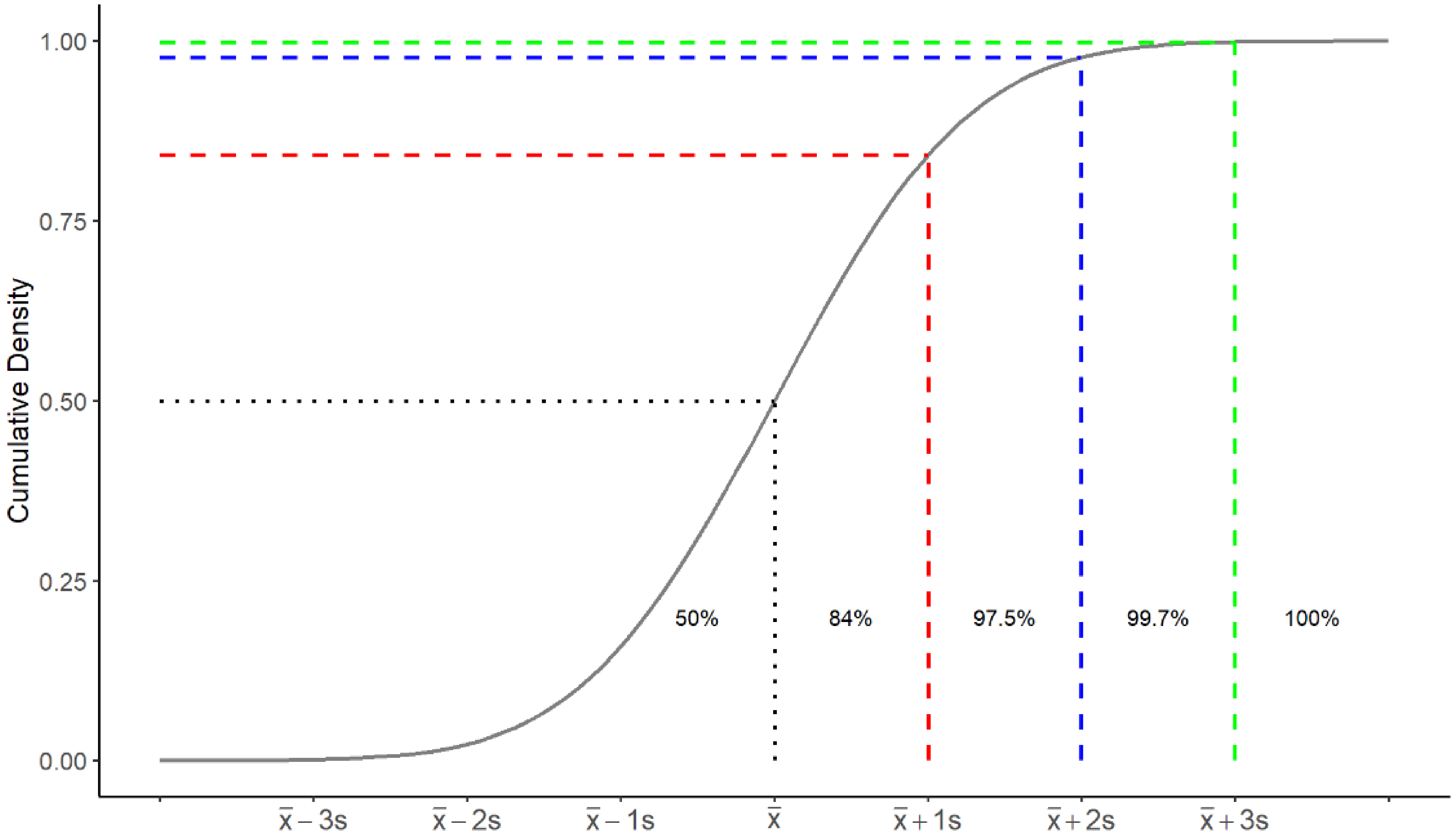
- Density Curve
  - What is a density curve?
  - What types of variables have density curves?

- Normal Distribution
  - What is the shape of normal distribution?

- Empirical Rule
  - Applies to what kind of distribution?
  - What proportion of data will have a value in the interval $\bar{x} \pm 1\sigma$?
  - What proportion of data will have a value in the interval $\bar{x} \pm 2\sigma$?
  - What proportion of data will have a value in the interval $\bar{x} \pm 3\sigma$?

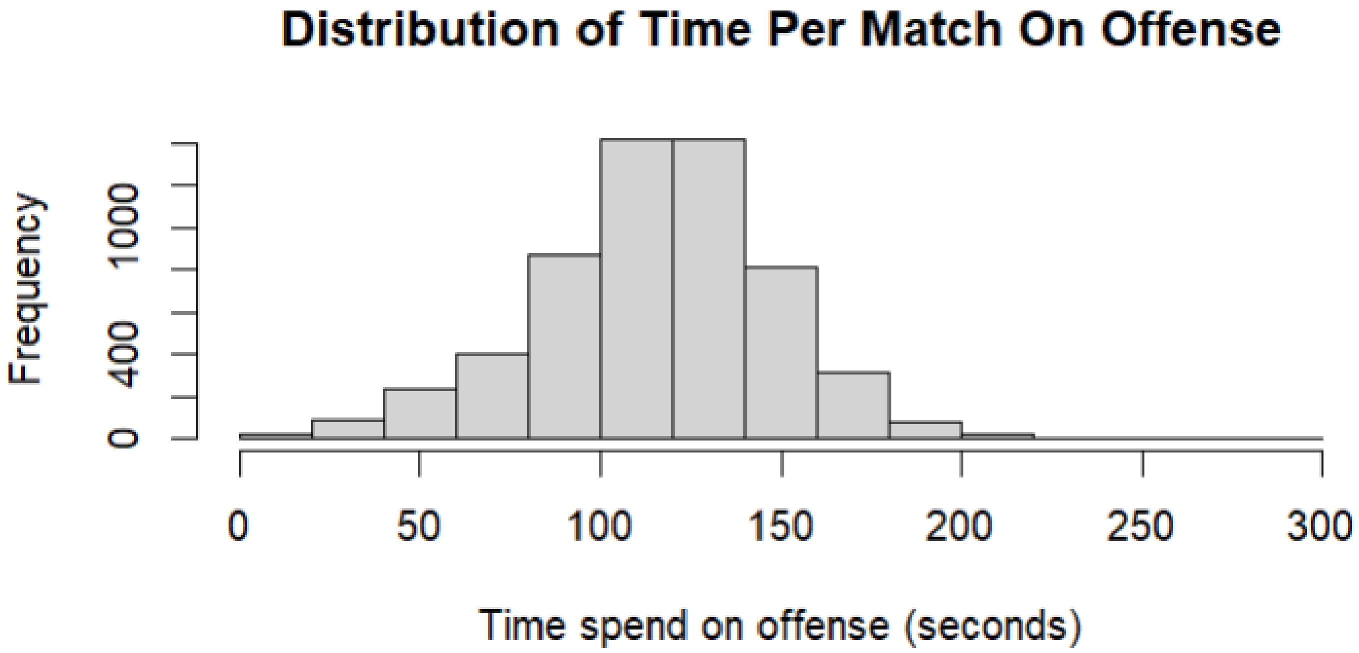
- Z-scores
  - What is a z-score?
  - What does a z-score tell us about an observation?
  - Does z-score convert a variables distribution to the normal distribution?

# Cumulative Density Function: Normal Distributions



# Practice: This Is Rocket League!!!

- Rocket League is a popular online video game and E-sport that emerged in 2015. The game enjoys a healthy following of around 93 million players per month. The game features players from around the world who compete in sports like soccer (football), basketball and hockey while controlling RC-like vehicles.

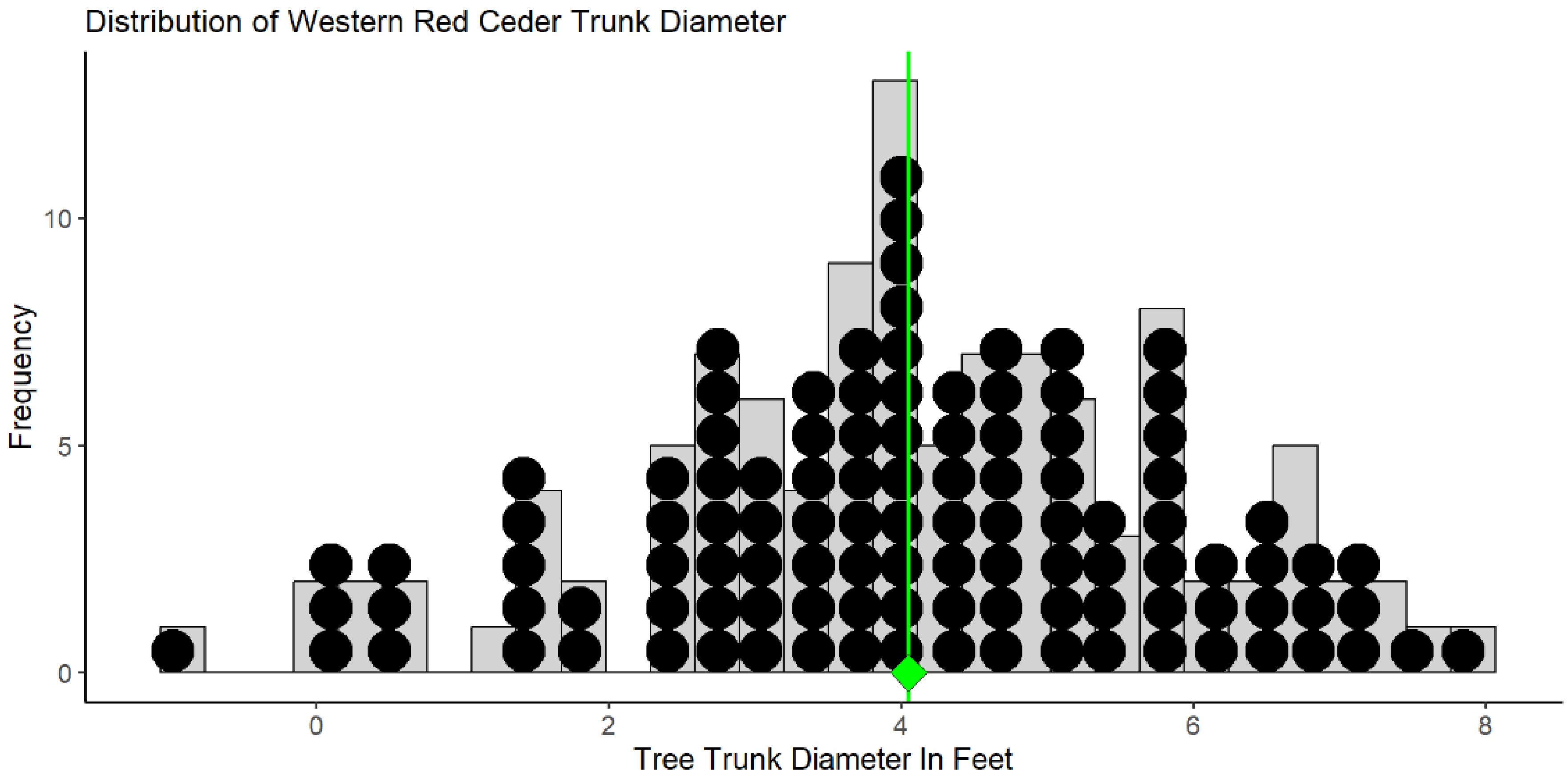


$\bar{x} = 115, \qquad s = 33$

# Estimating Diameter of Western Red Ceders

Consider a small study to estimate the average trunk diameter (in feet) of Western Red Ceders at the idlers rest nature preserve just outside of Moscow.

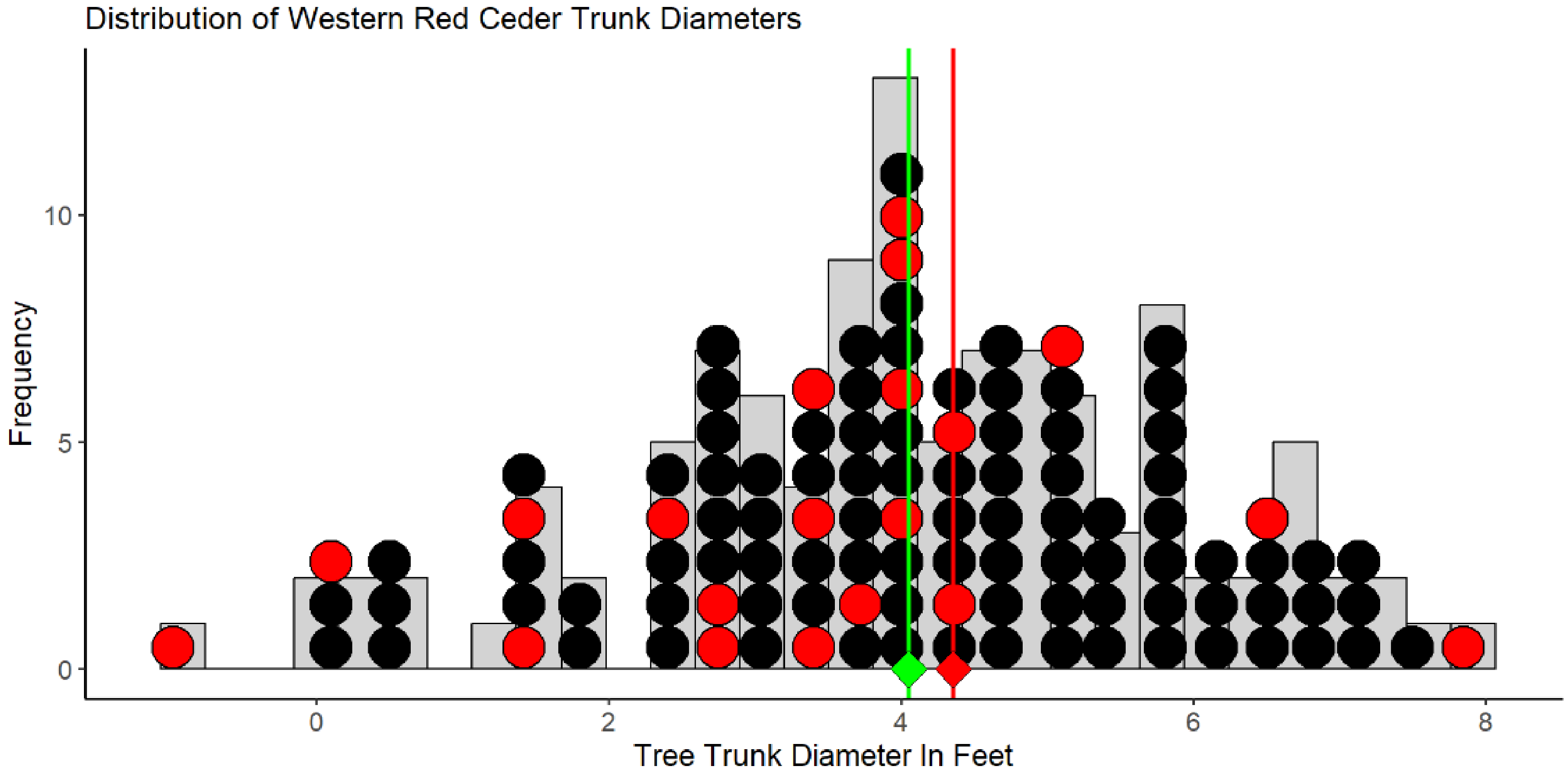
$$N = 109$$
$$\mu = 4.05$$
$$\sigma = 1.81$$



# Estimating Diameter of Western Red Ceders

Consider a sample of $n = 20$ trees to estimate the mean trunk diameter…

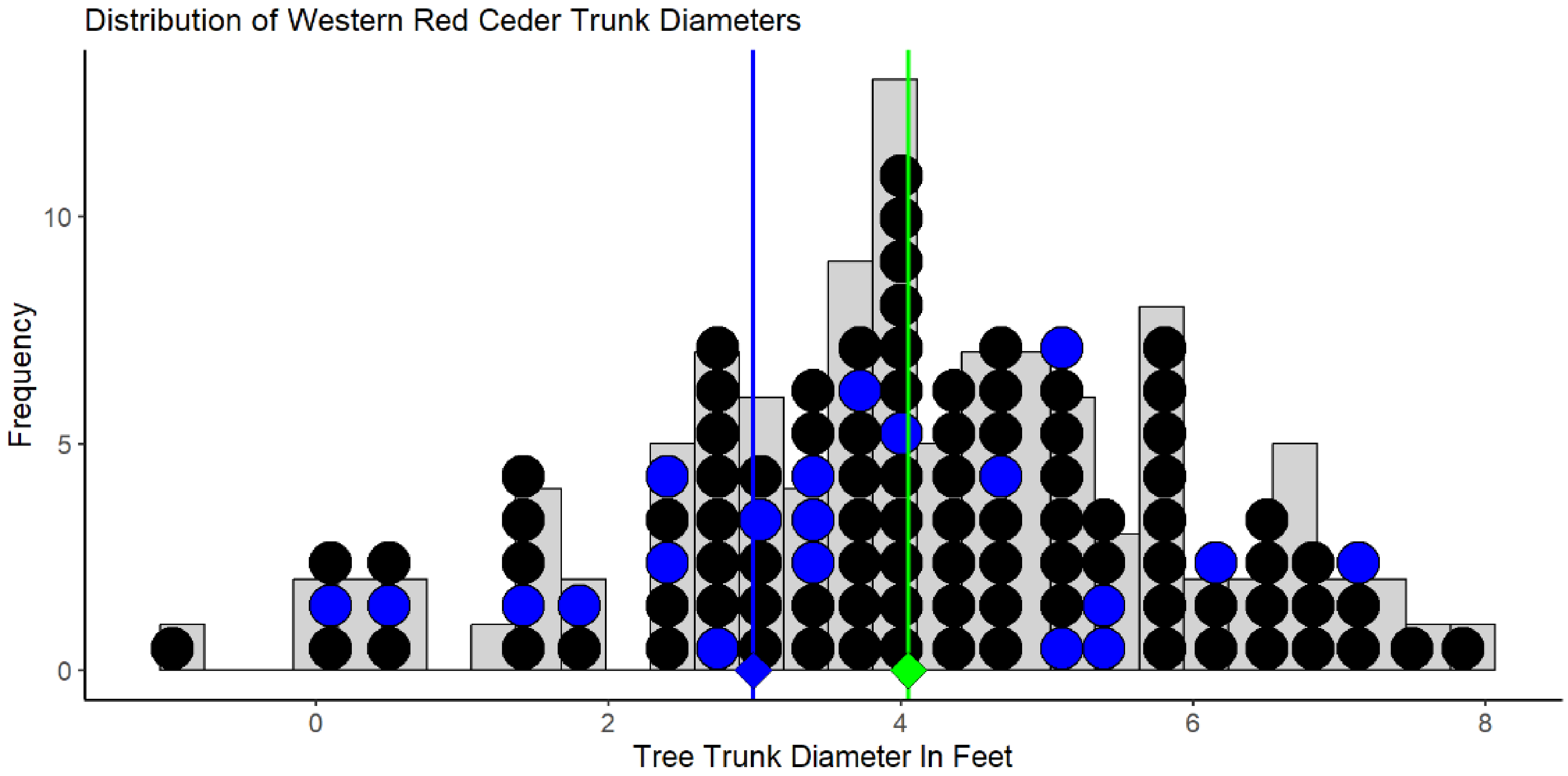
$\bar{x}_1 = 4.35$



# Estimating Diameter of Western Red Ceders

Consider another sample of $n = 20$ trees to estimate the mean trunk diameter…

$\bar{x}_2 = 2.98$



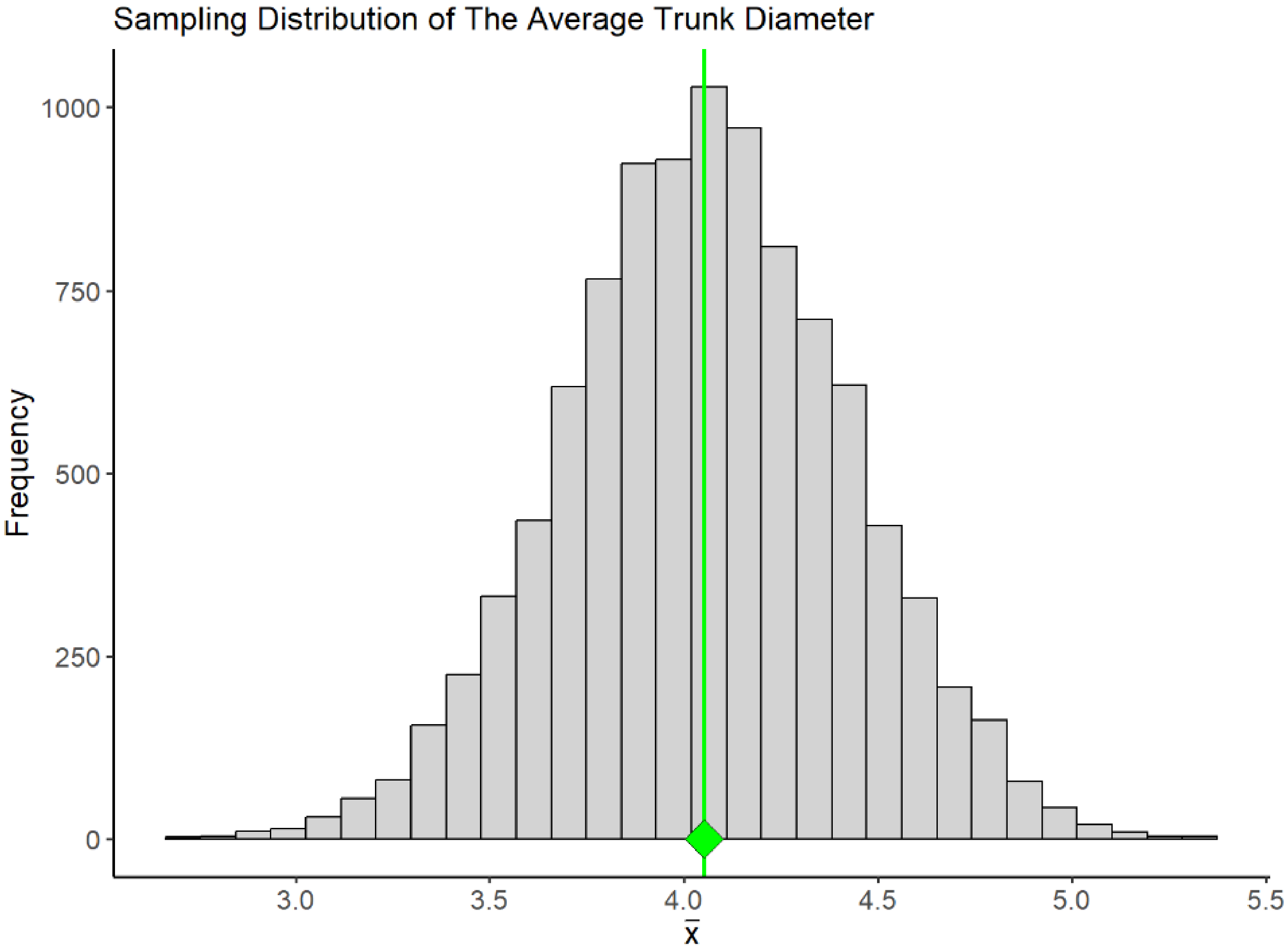
# Sampling Distributions

A **sampling distribution** is the distribution of a statistic.

- It arises from repeatedly sampling and estimation

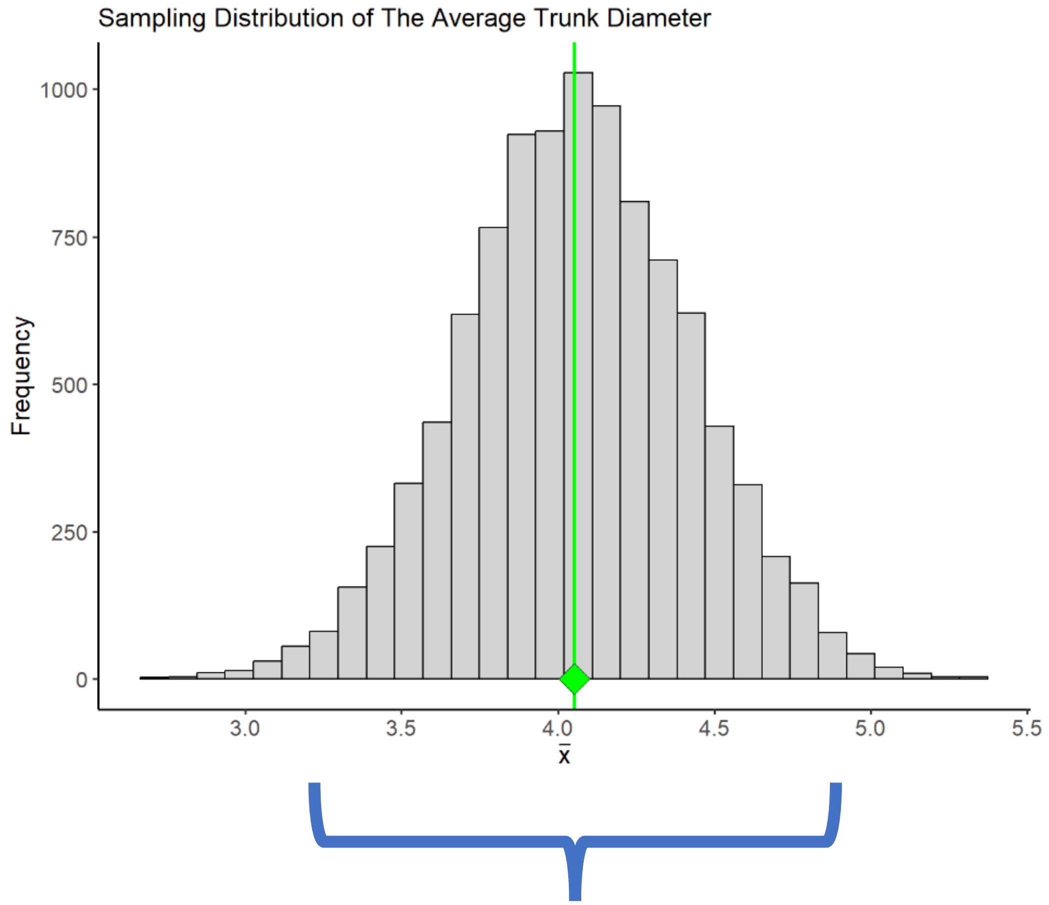
- it is usually a theoretical distribution

The variation between different samples leads to variation in the estimates that are calculated



# Margin of Error

- The **margin of error** of an estimate measures how far we expect an estimate to fall from the true value of a population parameter
- It is a measure of the between sample variability in our estimate
- It is the largest distance between the true population parameter and an estimate that is <u>not an outlier</u>



Between sample variability

# Statistical Significance

- A **Statistically significant** result is one that is decidedly *not* due to ordinary variation - this means a result that is not due to chance or coincidence.

- statistically significant results is which falls outside the margin of error.

# Sampling and Experimental Designs

# Anecdotal evidence

- Descriptive statistics alone is not enough evidence to make conclusive decisions about a variable or patterns

- Patterns in data can arise from many different sources

- **Anecdotal evidence** - evidence from information or testimony that is based on personal observations, individual experiences, or isolated examples, rather than on systematic and rigorous scientific analysis.
  - Anecdotal evidence often starts with phrases like “In my experience” or “it seems to me”

# Case Study: Cell Phones and Health

Cell phones emit electromagnetic radiation, and a cell phone's antenna is the main source of this energy. The closer the antenna is to a person's head, the greater the exposure to radiation. When cell phones started to become popular in the early 2000's , there was concern over the potential health risks they posed to users. Several studies explored the possibility of such risks

# The German Study (Stang et al., 2001)

- This study compared 118 patients with a rare form of eye cancer called uveal melanoma to 475 healthy patients who did not have this eye cancer.
- Cell phone use was measured using a questionnaire
- Findings: on average, the eye cancer patients used cell phones more often

# Study 1: The German Study (Stang et al., 2001)

- What is question the authors are trying to answer with data?

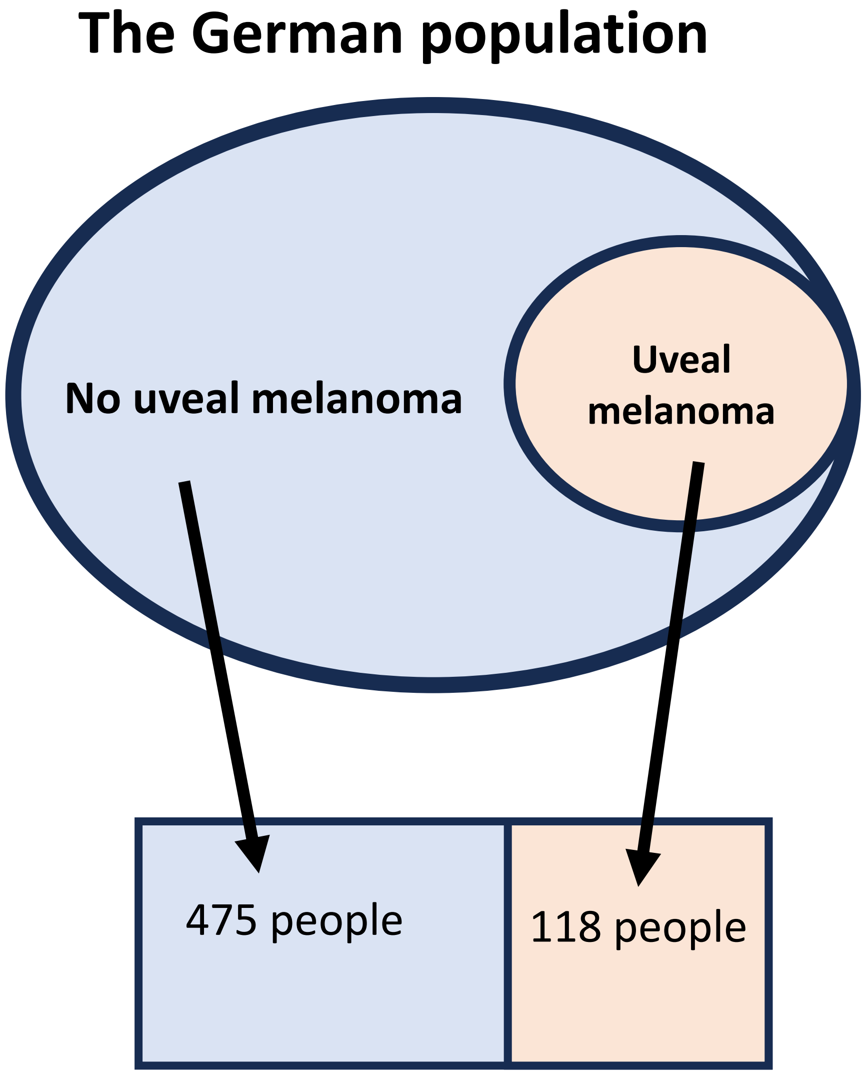
Is cell phone use associated with uveal melanoma?

- What is the population?

All citizens living in Germany

- What is the sample?

Two samples are taken: one taken from the subpopulation of people with uveal melanoma, another taken from the subpopulation of people without uveal melanoma



# Study 2: The British Study (Hepworth et al., 2006)

- This study compared 966 patients with brain cancer to 1,716 healthy patients who did not have brain cancer.
- Cell phone use was measured using a questionnaire
- Findings: cell phone use for the two groups was similar

# The British Study (Hepworth et al., 2006)

- What is question the authors are trying to answer with data?

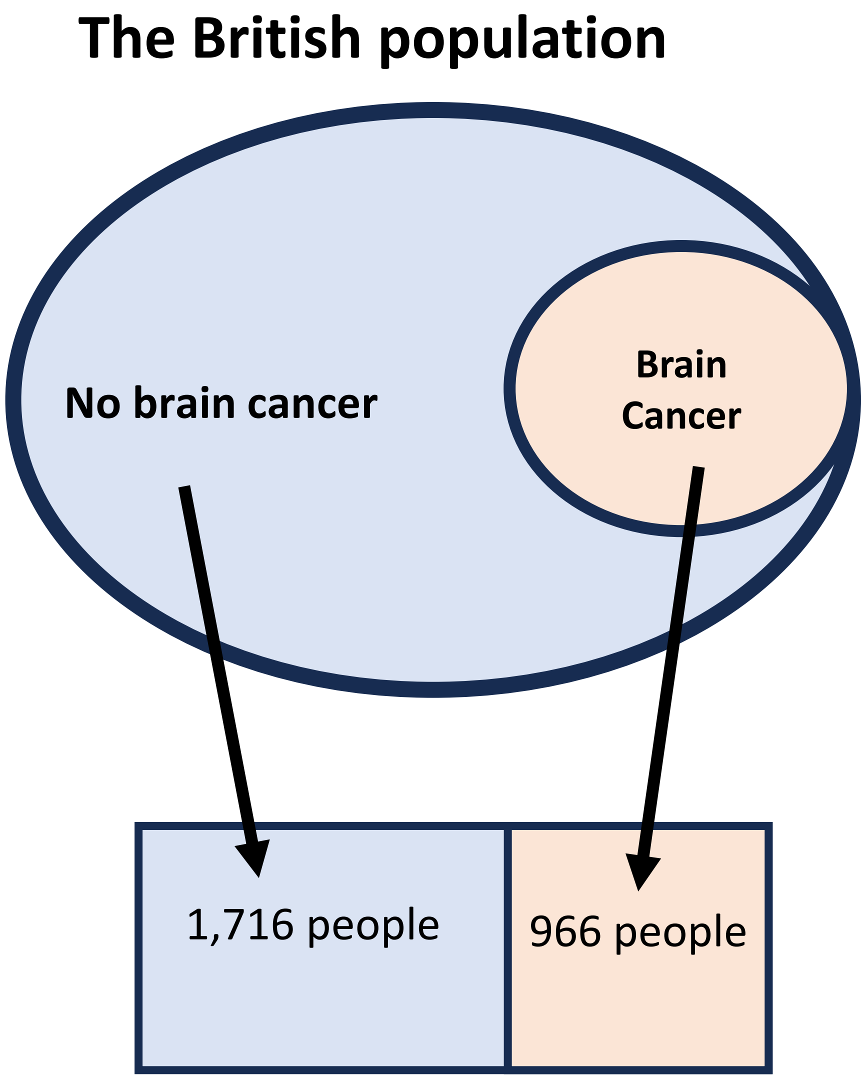
Is cell phone use associated with brain cancer?

- What is the population?

All citizens living in Great Britain

- What is the sample?

Two samples are taken: one taken from the subpopulation of people with brain cancer and another taken from the subpopulation of people without brain cancer



# Study 3: (Volkow et al., 2011) – *Journal of the American Medical Association*

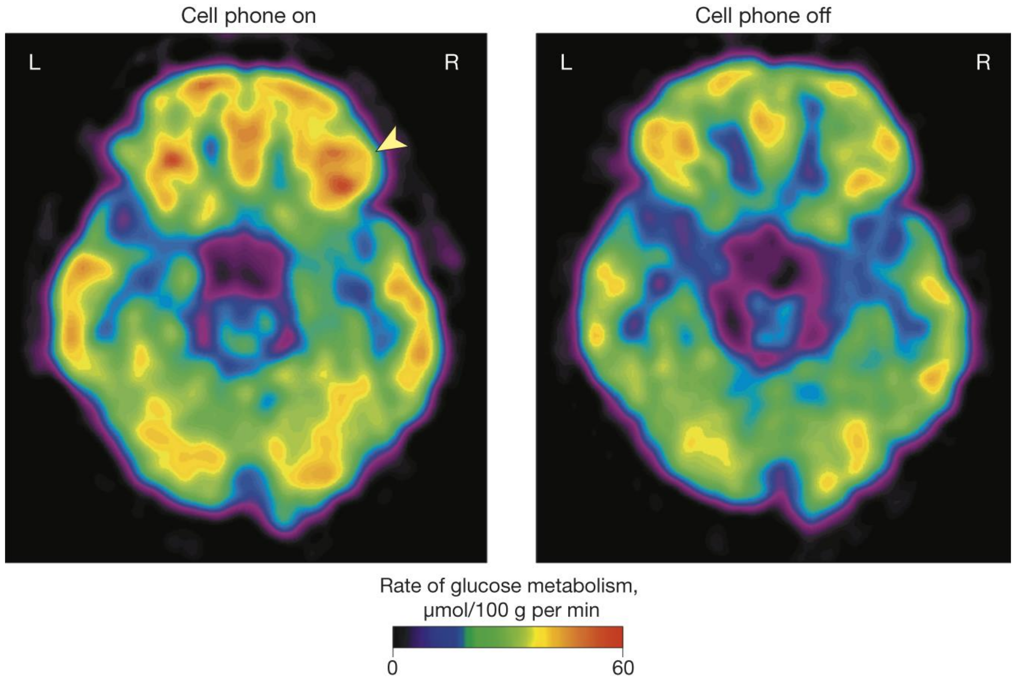
- Authors used a randomized crossover study to compare the effects of cell phone use on brain glucose metabolism in 47 individuals.
- Patients were fitted with two cell phone devices, one for each ear

  - Each patient was given two positron emission topography (PET) scans

  - The first was applied when both phones were off

  - The second was applied during a 50-minute muted call to one of the two cell phones (the call was randomly assigned to either the right or left phone)

- Comparison of the PET scans showed significantly increased activity in the part of the brain closest to the phone.



# Study 3 (Volkow et al., 2011) – *Journal of the American Medical Association*

- What is question the authors are trying to answer with data?

Does cell phone use cause increased brain activity (measured as glucose metabolism)?

- What is the population?

Any healthy individual

- What is the sample?

47 individuals selected as participants

# What is “good” data?

Studies 1 and 3 conflict with the results of study 2 – which should we trust?

Study and sampling designs can have a major impact on results…

Knowledge of different study designs helps guide us in deciding what research we should trust and when we should be skeptical.

# Response and Explanatory Variables

- In all 3 studies investigating cell phone use and physiological activity in the brain, the researchers were interested in two variables, a response variable and an explanatory variable

- **Explanatory variable** – this is the variable we manipulate or observe changes in

  - what was the explanatory variable in the German study?

- **Response variable** – this variable measures the outcome of interest. Studies focus on how the outcome "responds" to changes in the explanatory variable

  - what was the response variable in the German study?

- A study can have multiple response variables and multiple explanatory variables

# Experimental vs Observational Studies

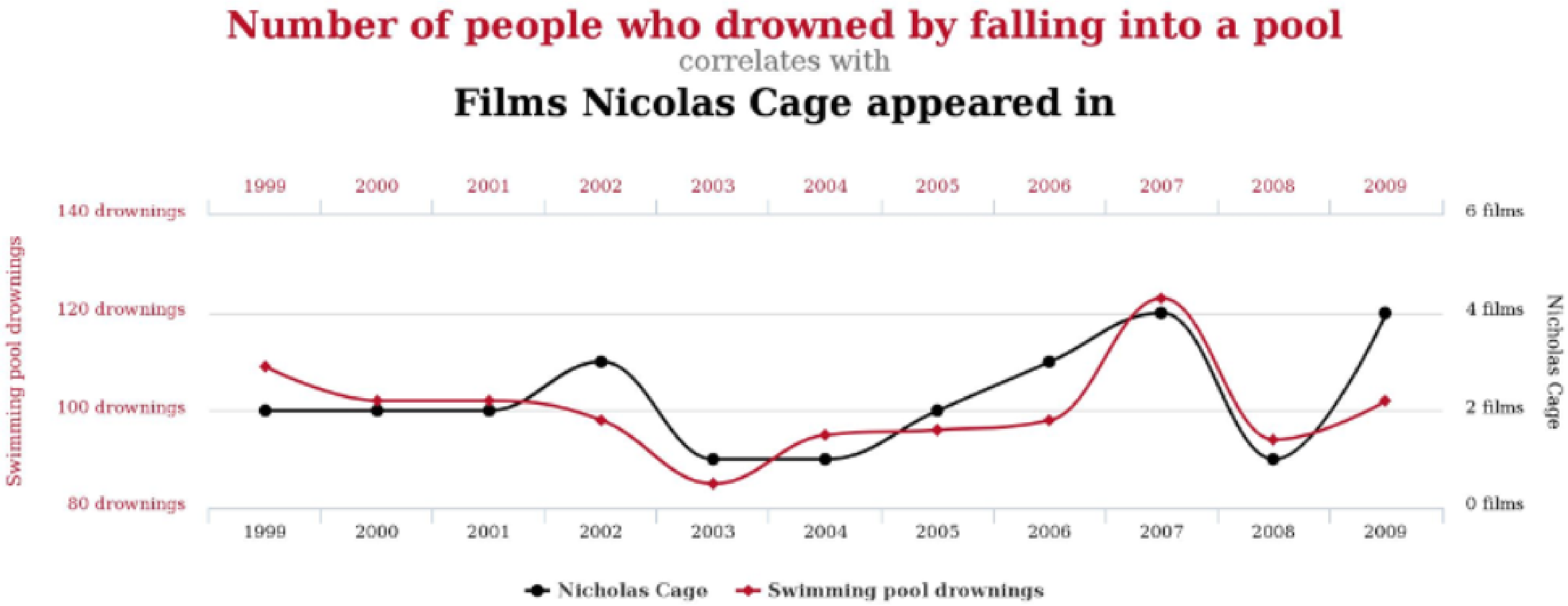
In an **experimental study**, researchers assign subjects to experimental conditions called **treatments** and then observe outcomes of the response variable(s).

In an **observational study** the researcher observes values of the response and explanatory variables in different subjects without any manipulation of the subjects in the study

Which (if any) of the three studies we examined are experimental? Which are observational?

# Association vs Causation

- A **statistical association** between two variables is a numerical measure of their relatedness.
- A **causal** relationship between two variables means that changing one variable causes a proportional change in the other variable  (also called a **cause and effect** relationship)
- An association between two variables **<u>does not</u>** imply causal relationship between them
- Example: there is a statistical association between the number of people who drowned by falling into a pool and the number of films Nicolas Cage appeared in a given year. However, there is obviously no causal relationship.

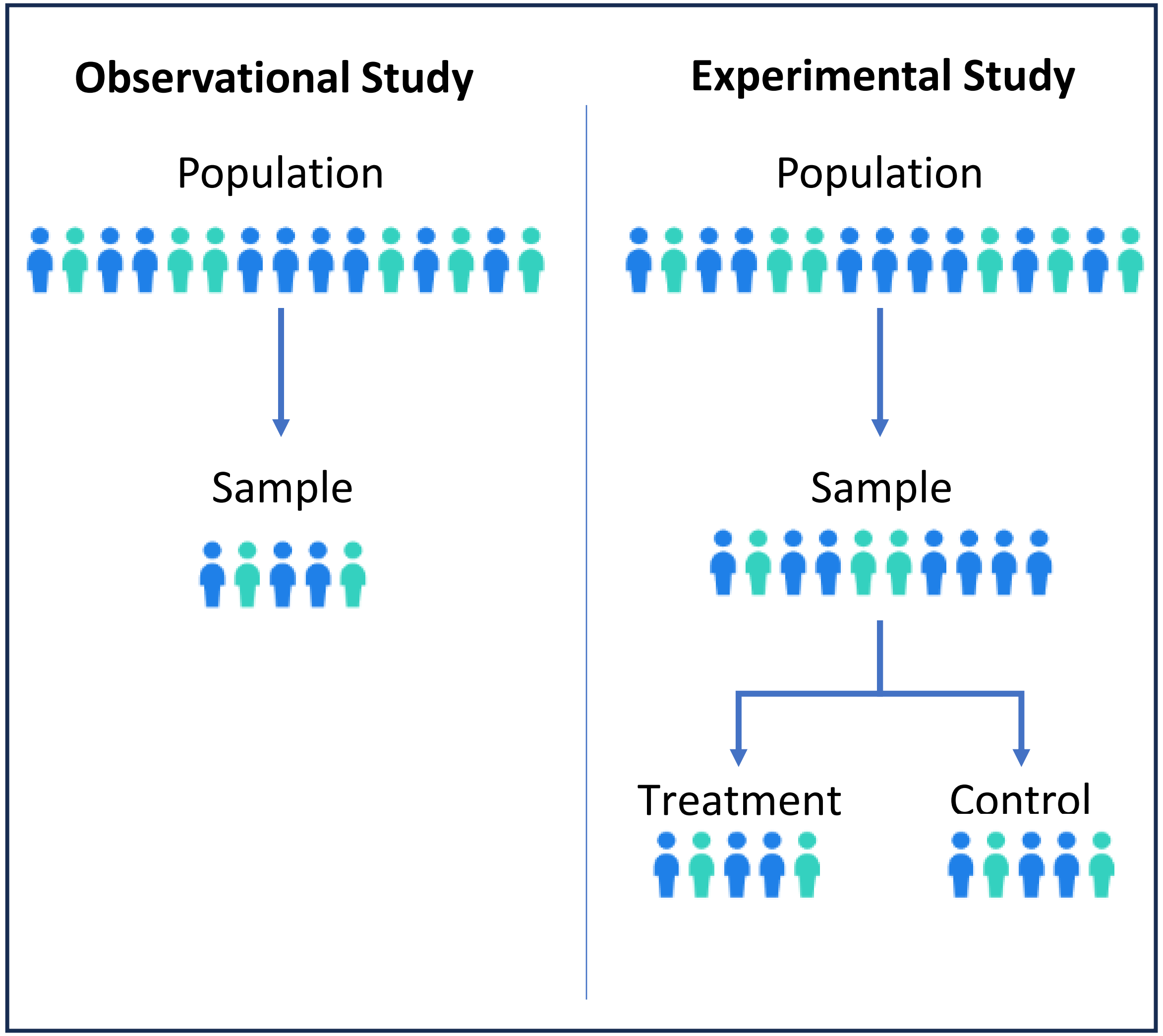


# Advantages of Experimental Studies

- Observational studies cannot definitively establish causation
- Observational studies are prone to **lurking variables** – a variable unknown to the researchers that is not included in the study and has an association with both the response and explanatory variables
- Lurking variables can induce false associations between response and explanatory variables.
- In experimental studies subjects (observations) are randomly assigned to treatment groups.

  this randomization balances the effect of lurking variables between the treatment groups and removes their influence on the association between the response and explanatory variables

**Study Design**



# Hallmarks of a good experiment

- Control group – a group of subjects in the experiment who do not receive the treatment
  - reduces bias in the experiment because by design the only difference between the two groups is the treatment

- Blinding – designing the experiment to ensure the subjects are unaware if they are in the treatment or control group.

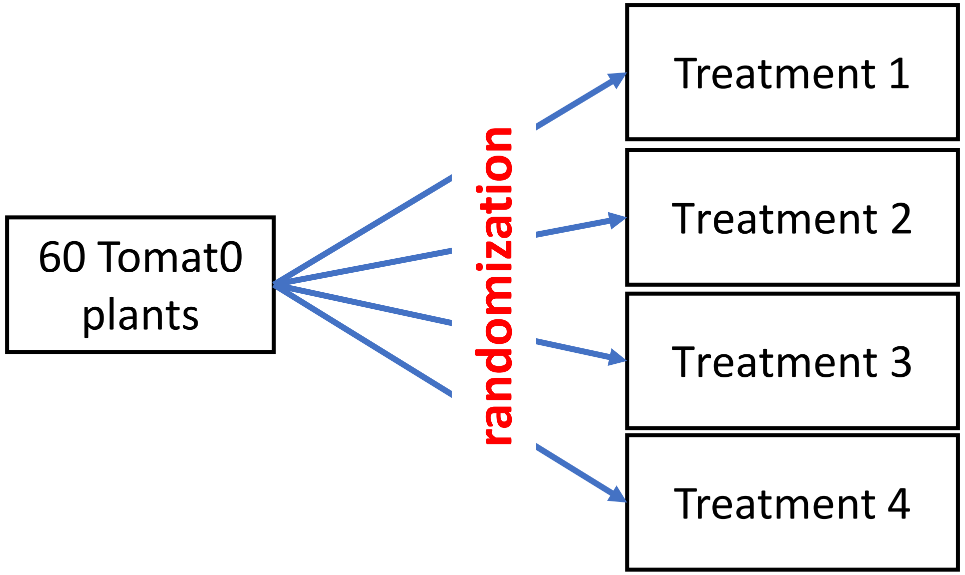
- Double blinding – When subjects as well as the researchers are unaware of who is assigned to the treatment group and who is assigned to the control group.

# Some Experimental Designs

- **Completely Randomize Design** – Subjects are randomly assigned to treatment groups.
  - compares response to a single factor
  - each unit has the same chance of being in the treatment or control groups

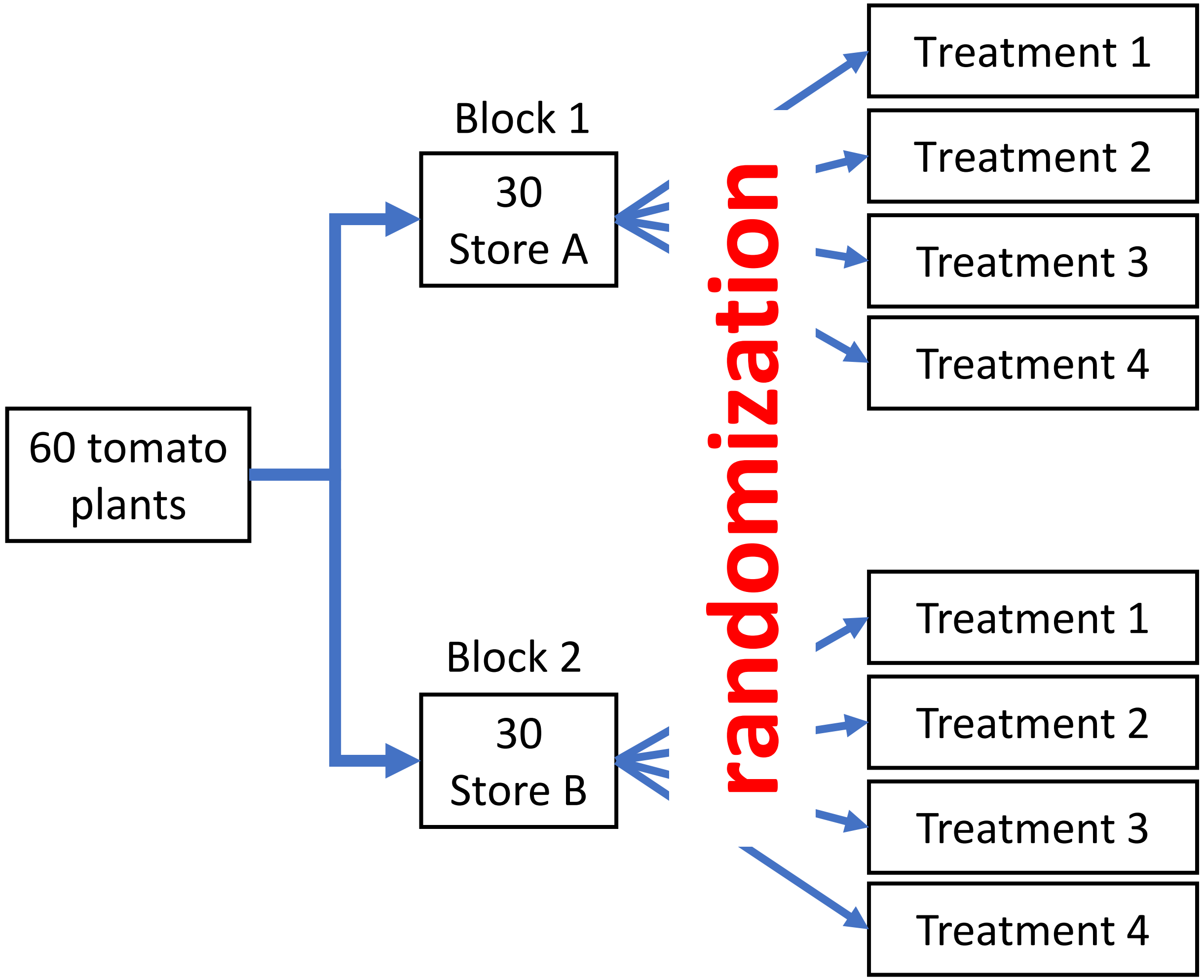
- **Multifactor experiments** – An experiment which compares multiple factors simultaneously
  - cheaper than conducting an experiment for each factor separately
  - we can learn more from a multifactor experiment

|  |  | Fertilizer A |  |
|---|---|---|---|
|  |  | YES | NO |
| Fertilizer B | YES | **Treatment 1**<br>Fertilizer A + Fertilizer B | **Treatment 2**<br>Fertilizer B + Placebo |
|  | NO | **Treatment 3**<br>Fertilizer A + Placebo | **Treatment 4**<br>Placebo only |



- **Randomized Complete Block Design** – When subjects are not similar enough, detecting differences among the treatment groups can be difficult. We instead create groups called **blocks.** Blocks are organized so that units inside a block are more similar. Each block sees all treatments in random order

- **Matched Pair Designs** – a design which takes measurements on each subject, usually once before the treatment and once after the treatment producing a set of paired measurements



# Surveys

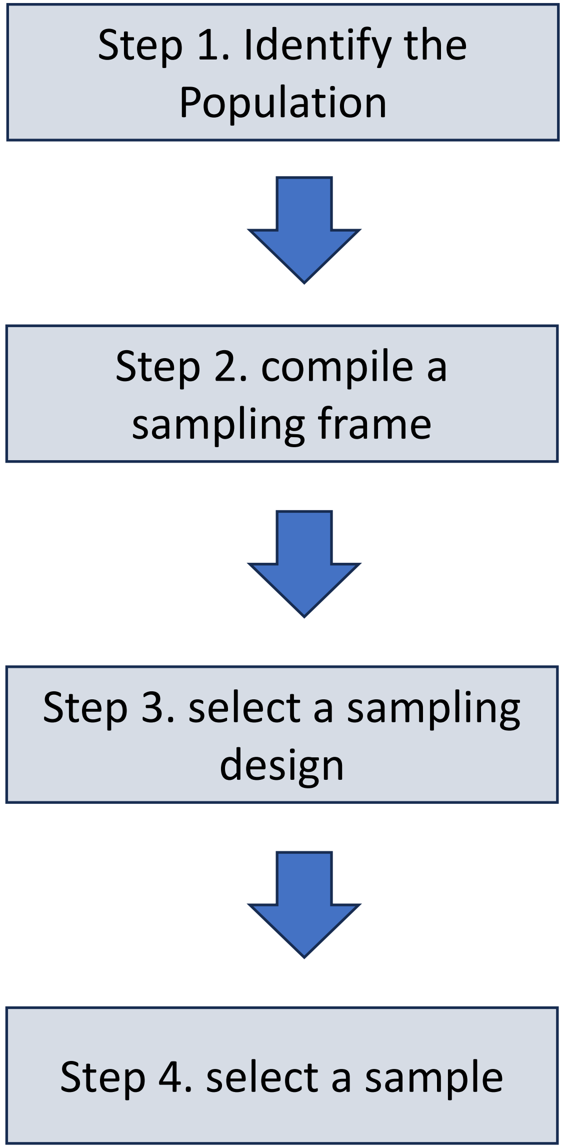
- A **sample survey** selects a sample of subjects from a population and collects data
  - ➢ In statistics, a survey is any information gathered from a sample of subjects.
  - ➢ It is a type of non-experimental study

- A **census** attempts to gather data for all (or nearly all) subjects in a population

- A **sampling frame** is a list of subjects in the population from which the sample will be collected

- A **sampling design** is the method that will be used to select subjects from the sampling frame

- We seek a sampling design that will lead to a sample that is representative of the entire population we are trying to estimate



Step 1. Identify the Population

Step 2. compile a sampling frame

Step 3. select a sampling design

Step 4. select a sample

# Example of a Survey

- Suppose I want to see what proportion of people in Moscow Idaho liked the Star Wars sequel trilogy. So, I acquire a phone book for Latah county, ID and call the first 100 people with an address in Moscow ID and ask them to rate the sequels on a scale of 1-10.
- What is the sampling frame?

The phone book for Latah County

- Do you think such a sample will be representative of the population we are trying to estimate?

**NO,** the sampling frame does not cover all possible people in the population of interest

# Simple Sampling Designs:

- **Convenience sampling** is a type of non-probability sampling that involves the sample being drawn from that part of the population that is close to hand**.**

  **Volunteer sample –** a type of sampling where participants self elect to be part of the study because they volunteer when asked or respond to advertisement

  - the most common type of convenience sampling

  - often required when we don't have a sampling frame for the population

  - this is the type of sampling used for most medical experiments

- A sample is more likely to be representative of the population if we let *chance,* rather than *convenience*, determine which subjects are sampled.

- In **simple random sampling** (also just called random sampling) each subject in the sampling frame to has an equal probability of being selected for the sample.

# Sampling Designs: Simple Random Sampling

- If we **sample with replacement,** then each time we sample a subject from the population we put the subject back i

In the college student height example, this would be the same as allowing each student a $\frac{1}{11,780}$ chance of being selected.

In general, for a population of size $N$ each subject will have a $\frac{1}{N}$ chance of being included in the sample.

- If we **sample without replacement,** then each time we sample a subject from the population we remove that subject from the sampling frame so that we cannot select them again.

    - This means that first subject will have a $\frac{1}{N}$ chance of being selected, the second a $\frac{1}{N-1}$, the third a $\frac{1}{N-2}$ … and so on
    - Sampling without replacement is common in most surveys because the sample size is usually small in comparison to the population size (i.e $n \ll N$) it is approximately the same as sampling with replacement

# Sources of Bias In Surveys

- **Bias** – when a sample is not representative of the population of interest.

- **Undercoverage** – Bias introduced by having a sampling frame that lacks representation from parts of the population

  - non-random sampling designs are prone to undercoverage

- **Nonresponse Bias** – When some of the sampled subjects cannot be reached or refuse to participate

  - most surveys suffer from this kind of bias

  - Current population survey of the U.S Census Bureau has a nonresponse rate of about 7%

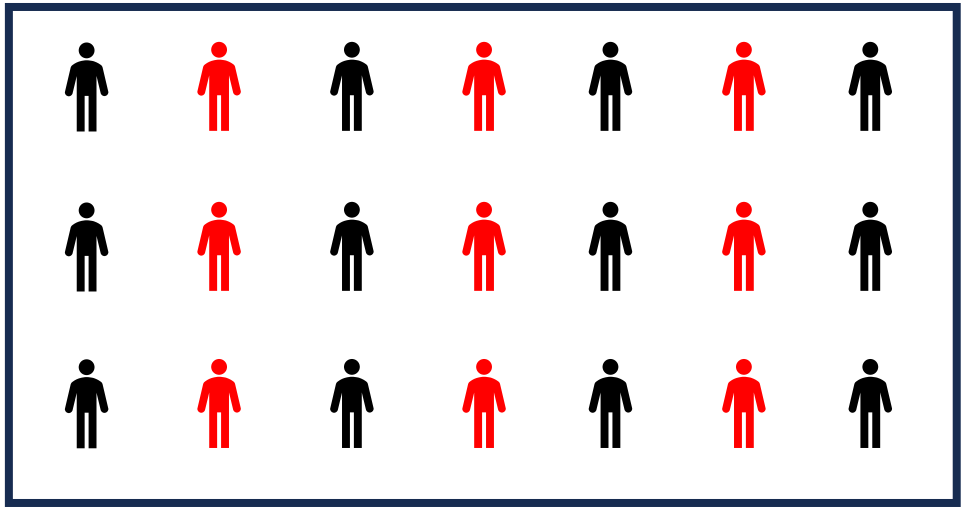
- **Response Bias** – When survey question is asked in a leading way or a subjects emotions affect how they respond

- A large sample size does NOT guarantee an unbiased sample!

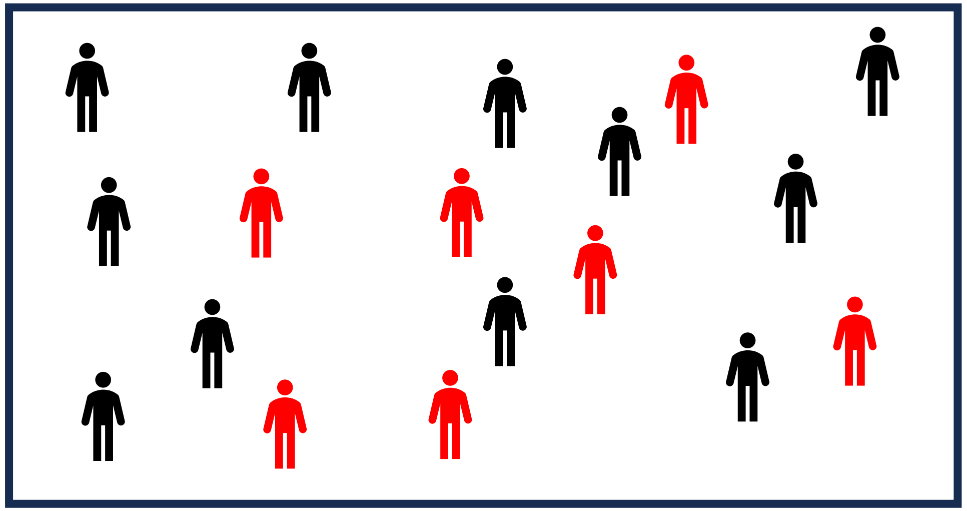
# More complex methods of sampling

- **Systematic Sampling** – A sampling method in which the researcher selects every $k^{th}$ subject from an ordered sampling frame
- **Cluster sampling** – A type of sampling method in which the population is divided in a set of clusters and the researcher selects a simple random sample of the clusters. The sample then comprises all subjects in the selected clusters.
- **Stratified Random Sampling** – A type of sampling method in which the population is separated into groups, call **strata,** based on some characteristic about the subjects. A simple random sample is then taken from <u>each</u> stratum.

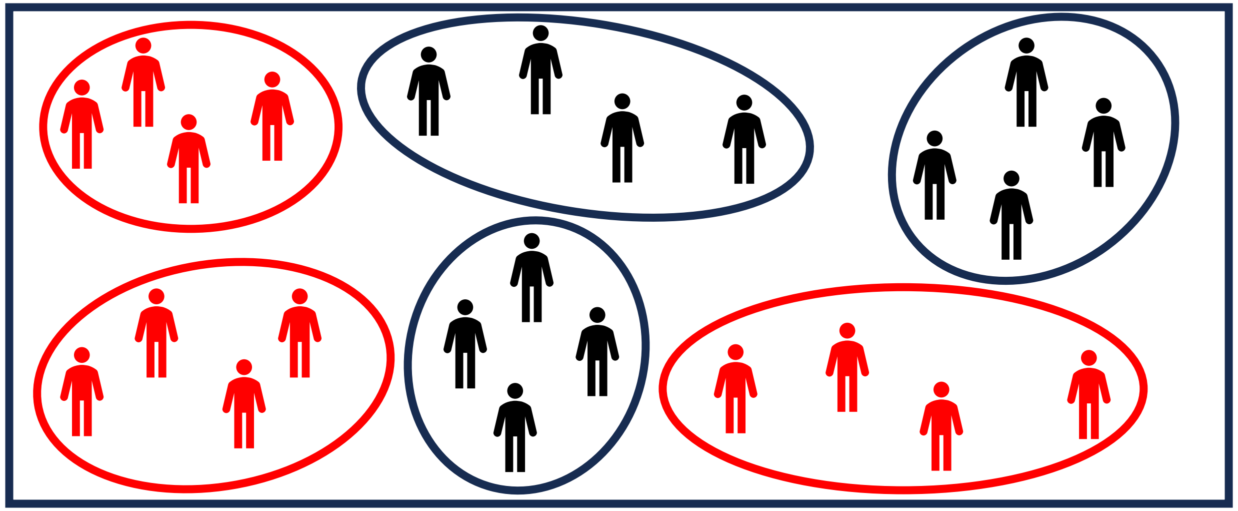
## Systematic Sampling



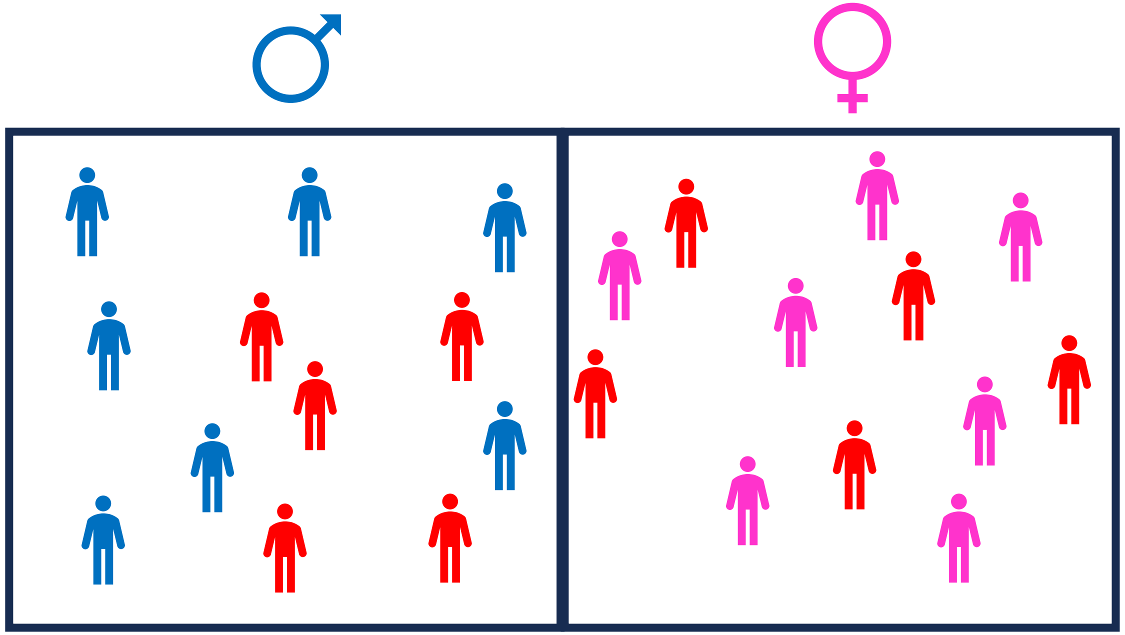
## Simple Random Sampling



## Cluster Sampling



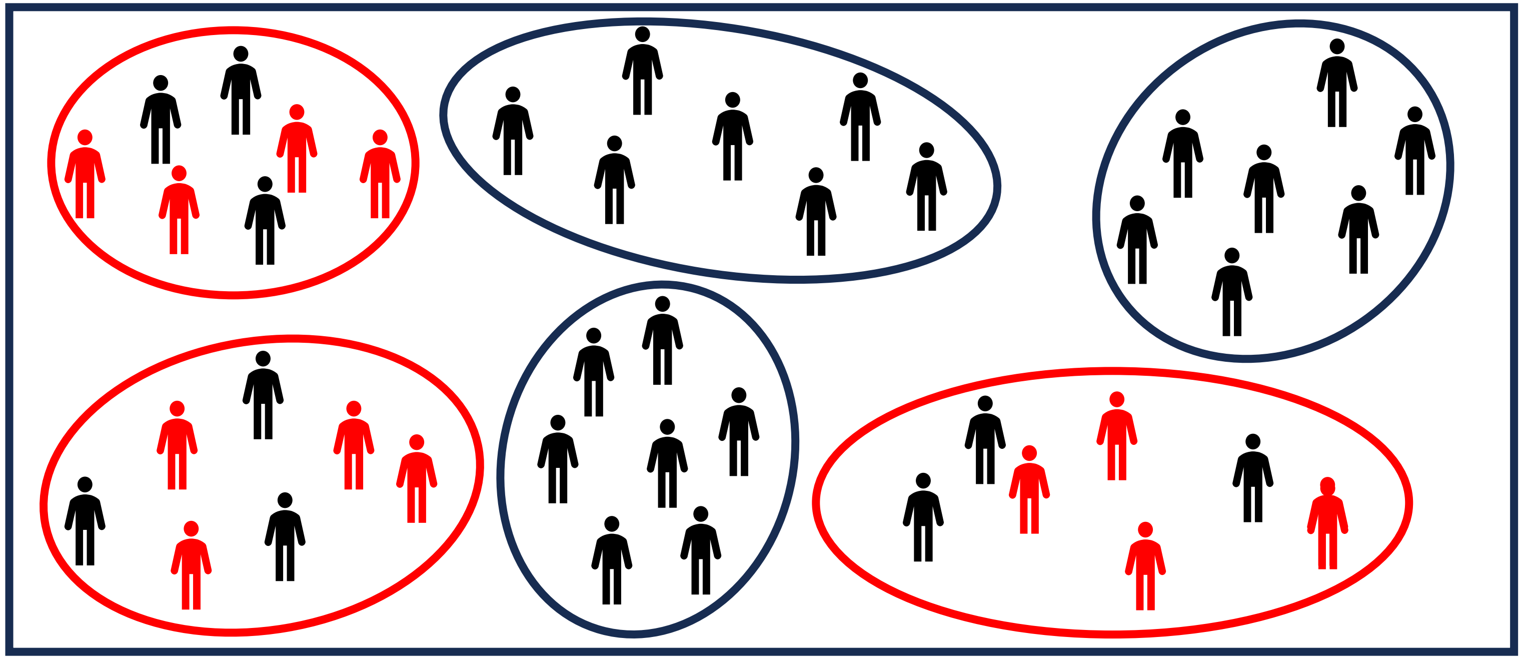
## Stratified Random Sampling



# More complex methods of sampling

- **Two – Stage cluster Sampling** - A type of sampling method in which the population is divided into a set of clusters and the researcher selects a simple random sample of the clusters. A simple random sample is then applied to each cluster

**Two – Stage Cluster Sampling**



# Advantages and Disadvantages of Sampling Designs

**Simple Random Sampling**

- Mathematically simple to compute estimates such as $\bar{x}$ and $s^2$
- Samples tend to be a good representation of the population

**Systematic Sampling:**

- Sometimes useful when there is no sampling frame available.
- Lower margin of error than simple random sampling and some cluster sampling designs.

# Advantages and Disadvantages of Sampling Designs

**Stratified Random Sampling:**

- **Administrative convenience** - It may be easier to conduct several smaller simple random sampling designs than coordinate one larger simple random sampling design.

- **Interest in individual strata** - The design ensures samples from all strata. A simple random sampling design might sample few or no elements from a stratum of interest.

- **Smaller margin of error** - By assuring samples from each strata, the combined sample tends to be more representative of the population, resulting in a smaller margin of error

# Advantages and Disadvantages of Sampling Designs

**Cluster Sampling:**

- The advantages of cluster sampling are that (a) **it can be less expensive than simple or stratified random sampling** and (b) it **can be used when a sampling frame is unavailable**
- A disadvantage of cluster sampling is that **the margin of error is often larger** than what it would be for simple random sampling or stratified random sampling

**Two-Stage Cluster Sampling:**

- Same advantages as above
- Usually has a smaller margin of error, because we can control two sample sizes: the number of clusters to sample, and the number of elements to sample from each sampled cluster

# Practice: Identify the Design

- Suppose I want to estimate the average height of my students in STAT 251 section 01. To do so, I use the registrars list to get the names of the students registered for my section. I number the students from 1 to N and select students 4, 8, 12, 16… to be my sample.
- What is the sampling frame?

The list of students in the class

- What is the sampling design?

Systematic sampling

# Practice: Identify the Design

- Suppose I want to see what proportion of people in Moscow Idaho liked the Star Wars sequel trilogy. So, I acquire a cadastral map (a map that shows the boundaries and ownership of land parcels) for Moscow, Idaho and group the the houses into city blocks. I take a random sample of city blocks and for each city block I selected and put a questionnaire in the mailboxes of all houses on that block.

- What is the sampling frame?

The cadastral map of Moscow

- What is the sampling design?

Cluster Sampling