# Sampling and Representation

# Table of contents

# 📝 Instructor Note 📝

The following module focuses on ethical issues that stem from sampling techniques and lack of representative samples. This module contains several real-world examples that aim to promote discussions around the importance and the consequences of sampling techniques and non-representative samples. Additionally, ethical principles from the American Statistical Association, along with recommended practices, are delineated. The module concludes with an acknowledgement that sampling is complex and that data scientists may have the tools to deal with and discern natural, nuanced patterns.

# 1 Amazon’s Hiring Algorithm

A Real World Example and Discussion on the Impact of Poorly Represented Populations

# 1. Amazon’s Automated Hiring Tool

In 2014, Amazon engineers developed an algorithm using the resumes of current employees from the past decade to expedite the hiring process. The algorithm reviewed applicants’ resumes and provided job candidate scores (1-5), with higher scores indicating a more well-qualified candidate.

# 1. Amazon’s Automated Hiring Tool: Reflection

Consider the following points -

How might:

- a small vs. a large number of observations influence an algorithm?
- quality of resumes play a role in the development of a fair hiring algorithm?
- the algorithm generalize to a population of applicants?

# 1. Amazon’s Automated Hiring Tool: Discussion

Take 5 minutes to discuss in groups any of your observations, posed questions, and thoughts related to ethics.

Take an additional 5 minutes to discuss with the whole class your group’s conversation.

# 📝 Instructor Note 📝

There are several points related to ethical or social justice practices that can be discussed:

- what and how much additional information would be important to collect from current employees to develop an accurate classification tool?
- Responsibility: Amazon has data scientists who reviewed the algorithm for bias after it was employed, which is key to upholding ethical principles.
- Amazon did not provide information related to the number of and quality of resumes that were used to create the hiring algorithm or any additional features. However, the number and the quality of the sample of resumes is key.
- You could spend time discussing how data scientists could mitigate bias, besides using female resumes.

# 1. Amazon’s Automated Hiring Tool: Final Comments

- Findings: Amazon’s researchers quickly discovered that fewer qualified women were being considered for more technical positions, such as software engineering and data science roles. A review of the training dataset and Amazon’s employee resumes revealed that male employees held the majority of technical positions.
- Amazon’s employees resumes are a sample of the general population of resumes for various employment positions.
- Amazon took a non-probabilistic sampling approach to collect data.

# 💬 Possible Discussion Questions 💬

- Consider posing the question: “If researchers determined that more women were hired overall in comparison to men, would you consider the hiring algorithm to be fair?”

# 1. Amazon’s Automated Hiring Tool: Final Comments

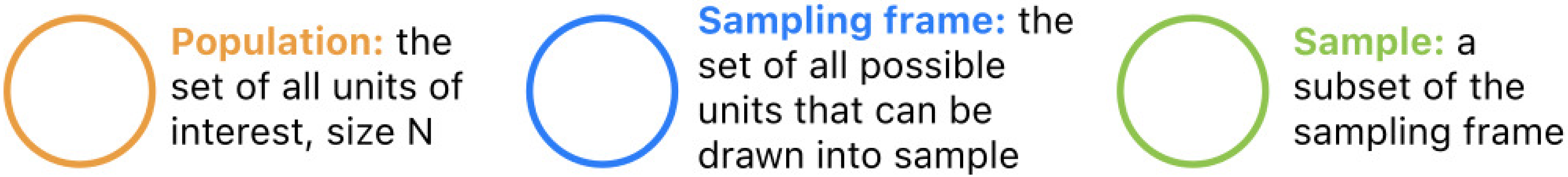
- Using probabilistic sampling approaches would have led to a more representative sample of resumes, which staff could have used to infer the qualifications of the population of applicants. A substantial body of research has demonstrated that non-probabilistic data collection can lead to biased analysis, and statisticians generally do not recommend making inferences about the population.

# 2 Defining Sampling

# 2. Defining Sampling

All data are collected somehow. A **sampling design** is a *way of selecting observational units for measurement*. It can be construed as a particular relationship between:

- a **population** (all entities of interest);
- a **sampling frame** (all entities that are possible to measure); and
- a **sample** (a specific collection of entities).



**Population:** the set of all units of interest, size N

**Sampling frame:** the set of all possible units that can be drawn into sample

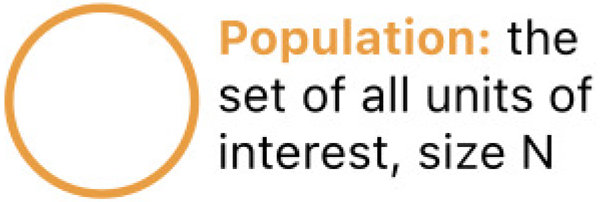
**Sample:** a subset of the sampling frame

# 2. Defining Sampling: Population

The **observational unit** is the ***the entity measured for a study*** – datasets consist of observations made on observational units.

In less technical terms, all data are data *on* some kind of thing, such as countries, species, locations, and the like.

A statistical **population** is the ***collection of all units of interest***. For example:



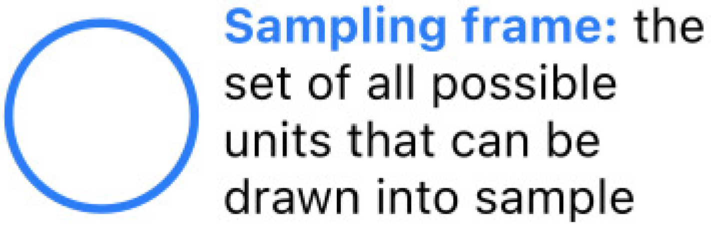
Population: the set of all units of interest, size N

- all countries (GDP data)
- all mammal species (Allison 1976)
- all babies born in the US (babynames data)
- all locations in a region (SB weather data)
- all adult U.S. residents (BRFSS data)

# 2. Defining Sampling: Sampling Frame

There are usually some units in a population that can’t be measured due to practical constraints – for instance, many adult U.S. residents don’t have phones or addresses.

For this reason, it is useful to introduce the concept of a **sampling frame**, which refers to ***the collection of all units in a population that can be observed for a study***. For example:



**Sampling frame:** the set of all possible units that can be drawn into sample

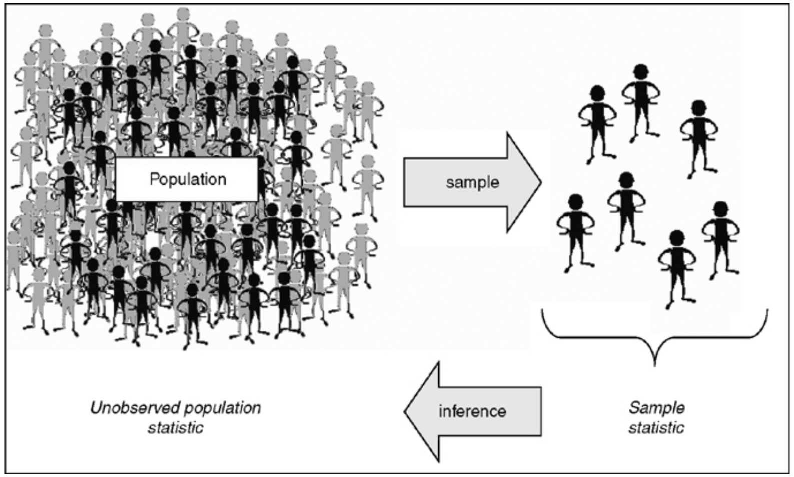
- all countries reporting economic output between 1961 and 2019
- all babies with birth certificates from U.S. hospitals born between 1990 and 2018
- all adult U.S. residents with phone numbers in 2019

# 2. Defining Sampling: Sample

Finally, it's rarely feasible to measure every observable unit due to limited data collection resources – for instance, states don't have the time or money to call every phone number every year.

A **sample** is ***a subcollection of units in the sampling frame actually selected for study***. For instance:

- 234 countries;
- 62 mammal species;
- 13,684,689 babies born in CA;
- 1 weather station location at SB airport;
- 418,268 adult U.S. residents.

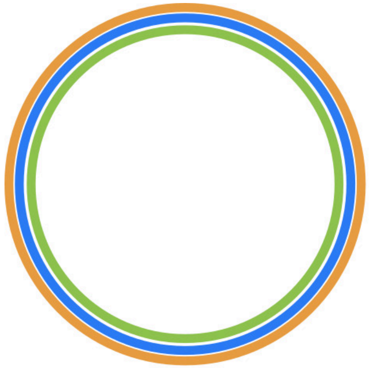


# 2. Defining Sampling: Census

The simplest scenario is a **population census**, where the entire population is observed.

For a census: $S = F = P$

*All properties of the population are definitively* ***known*** *in a census.* So there is no need to sample.



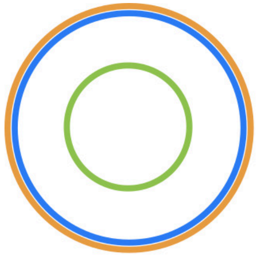
**Census:** sample covers entire population

# 2. Defining Sampling: Simple Random Sample

The statistical gold standard for inference, modeling, and prediction is the **simple random sample** in which units are selected at random from the entire population.

For a simple random sample, the frame equals the population, and the sample is a subset of the population.

*Sample properties are reflective of population properties in simple random samples.* Population inference is straightforward.



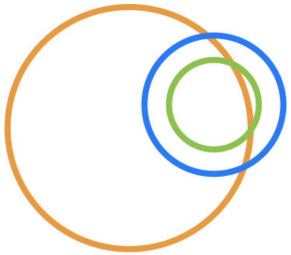
**Ideal random sample:** sampling frame is the population and the sample is drawn at random

# 2. Defining Sampling:‘Typical’ Sample

More common in practice is a random sample from a sampling frame that overlaps but does not cover the population.

For a ‘typical’ sample: $S \subset F \quad \text{and} \quad F \cap P \neq \emptyset$

*Sample properties are **reflective of the frame** but not necessarily the study population.* Population inference gets more complicated and may not be possible.

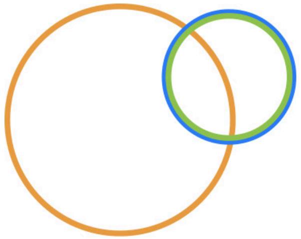


**Typical sample in practice:** sampling frame partially overlaps with population

# 2. Defining Sampling: ‘Administrative’ Data

Also common is **administrative data** in which all units are selected from a convenient frame that partly covers the population.

*Administrative data are not really proper samples; they cannot be replicated and they do not represent any broader group.* No inference is possible.



**Administrative data:** sample covers entire sampling frame, but sampling frame does not cover population

# 2. Defining Sampling: Sampling Design

What is a sampling design?

- **A sampling design** is a way of selecting observational units for measurement (individuals, groups, organizations, events, or other relevant elements).
- *General Goal*: Select a set of individuals (or any unit of measurement of interest) from a population in a way that can be used to learn about characteristics of the population.
- *Secondary Goal*: Sampling design is to yield maximize the accuracy of inferences about a population (i.e. not biased, minimum variance).

# 2. Defining Sampling: Scope of Inference

The relationships among the population, frame, and sample determine the **scope of inference**: the ***extent to which conclusions based on the sample are generalizable***.

A good sampling design can ensure that the statistical properties of the sample are expected to match those of the population. If so, it is sound to generalize:

- the sample is said to be *representative* of the population
- the scope of inference is *broad*

A poor sampling design will produce samples that distort the statistical properties of the population. If so, it is not sound to generalize:

- sample statistics are subject to bias
- the scope of inference is *narrow*

# 2. Defining Sampling: Characterizing Sampling Designs

The sampling scenarios above can be differentiated along two key attributes:

1. The overlap between the sampling frame and the population.
   - frame $=$ population
   - frame $\subset$ population
   - frame $\cap$ population $\neq \emptyset$
2. The mechanism of obtaining a sample from the sampling frame.
   - random sampling
   - convenience sampling

*If you can articulate these two points, you have fully characterized the sampling design.*

# 2. Defining Sampling: Sampling Mechanisms

In order to describe sampling mechanisms precisely, we need a little terminology.

Each unit has some **inclusion probability** – ***the probability of being included in the sample***.

Let's suppose that the frame $F$ comprises $N$ units, and denote the inclusion probabilities by:

$$p_i = P(\text{unit } i \text{ is included in the sample}) \quad i = 1, \ldots, N$$

The inclusion probability of each unit depends on the physical procedure of collecting data.

# 2. Defining Sampling: Sampling Mechanisms

**Sampling mechanisms** are *methods of drawing samples* and are categorized into four types based on inclusion probabilities.

- in a **census** every unit is included
  - $p_i = 1$ for every unit $i = 1, \dots, N$
- in a **random sample** every unit is equally likely to be included
  - $p_i = p_j$ for every pair of units $i, j$
- in a **probability sample** units have different inclusion probabilities
  - $p_i \neq p_j$ for at least one $i \neq j$
- in a **nonrandom sample** there is no random mechanism
  - $p_i = 1$ for $i \in S$

# 2. Defining Sampling: Example

Annual observations of GDP growth for 234 countries from 1961 - 2018.

- Population: all countries in existence between 1961-2019.
- Frame: all countries reporting economic output for at least one year between 1961 and 2019.
- Sample: equal to frame.

So:

1. Overlap: frame partly overlaps population.
2. Mechanism: sample is every country in the sampling frame.

*This is administrative data* with no scope of inference.

# 2. Defining Sampling: Example

Phone surveys of 418K U.S. residents in 2019.

- Population: all U.S. residents.
- Frame: all adult U.S. residents with phone numbers.
- Sample: 418K adult U.S. residents with phone numbers.

So:

1. Overlap: frame is a subset of the population.
2. Mechanism: probability sample.
    - Randomly selected phone numbers were dialed in each state, so individuals in less populous states or with multiple numbers are more likely to be included

*This is a typical sample* with narrow inference to adult residents with phone numbers.

Consider posing the following questions:

- Can you use and trust phone surveys to make generalization about the U.S. population?
- What **types** of people are excluded from phone surveys? What are the ethical implications of excluding these individuals? Some concrete examples of common areas of research that use phone surveying include public health surveys (health behaviors like smoking, exercise, and diet, access to healthcare or mental health indicators), quality of life study or survey related to market research.

# 3 Swain’s Jury Panel Case Study

A Real World Example and Activity that Demonstrates Representation in Samples

# 3. Swain’s Jury Panel Case Study

The following case study highlights the ethical importance and implications surrounding sampling. This case study and discussion is an extension of Data8’s example.

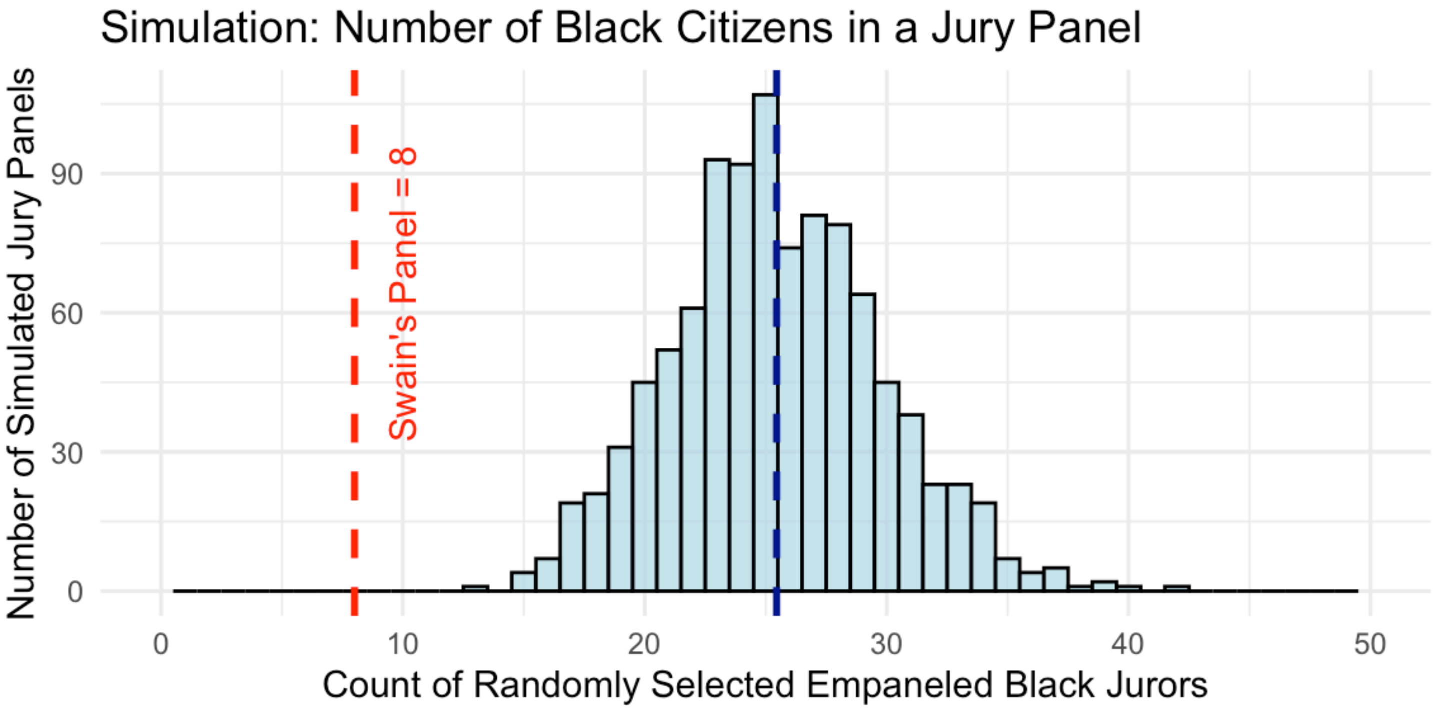
- In 1962, Robert Swain was charged and convicted with committing a crime in Alabama’s Talladega County, which had the possibility of a death sentence.
- Prior to his trial and after his conviction, his lawyers argued that Black citizens of Talladega were systematically excluded from grand juries such as Swain’s case (Carrington 2023), which violated the right to an impartial jury guaranteed in the U.S. Constitution.
- Note, only men 21 years and older were allowed to serve as jury members at the time.
- Jurors are selected from a jury panel of 100 citizens which should be representative of a region’s eligible population.

# 3. Swain’s Jury Panel Case Study

- During Swain’s trial, **26%** of males in Talladega County were Black, however the jury panel consisted of 8 Black males out of 100, in other words Black males made up 8% of the jury panel. None actually served on the final jury.
- Swain’s lawyers appealed his conviction arguing that he was not given an opportunity to **fair trial** due to lack of representation of Black males on the jury panel.
- His appeal went all the way up to the Supreme Court, who decided that the percent difference was small and that there was no attempt by the State of Alabama to exclude Black citizens.
- Using simulation, we shall evaluate the number of Black citizens in jury panels if they were randomly selected from a population that consisted of 26% Black and 74% White males. We can also evaluate a distribution in which 8% of the population, on average, were Black males.

# 3. Swain’s Jury Panel Case Study: Jury Pool

The statistic of interest in these simulations is the count of Black males on a panel of 100 if randomly selected from a population similar to Talladega Countym and a population with an 8% Black male proportion. Each randomly selected sample consists of 100 individuals, representing the number of citizens on a jury panel. We will do 1000 iterations of such random selection.



# 3. Swain’s Jury Panel Case Study: Jury Pool Discussion

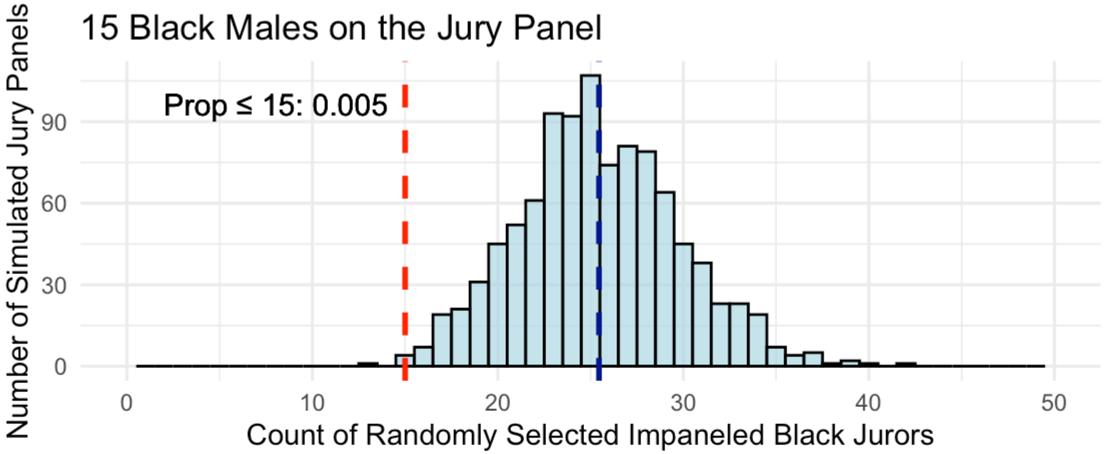
Discuss in a peer group (5 minutes):

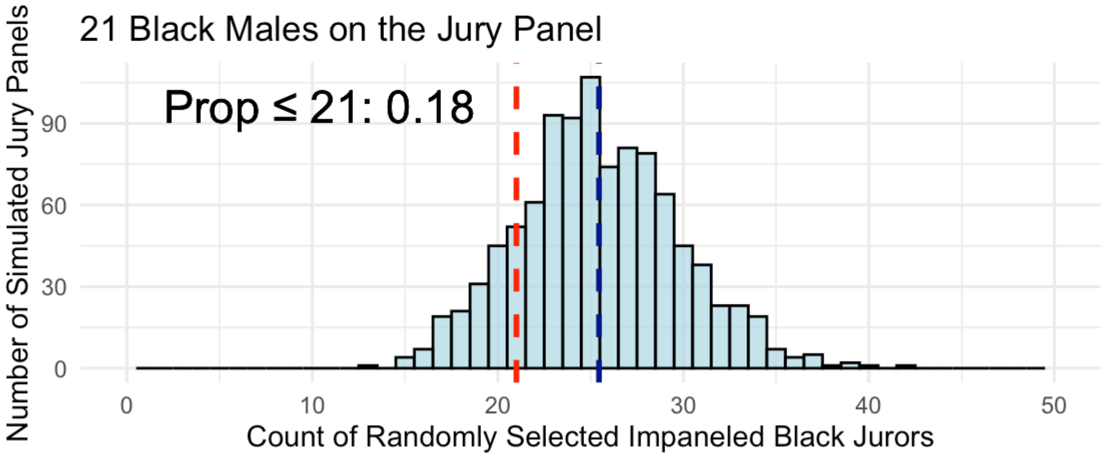
- What do you observe in the plot? Evaluate the range and the mean of possible number of Black males that could have been selected for the jury panel.
- What does the **distribution** of randomly selected samples and actual count of jurors suggest about the jury panel in Swain’s case? Is the difference small?

# 3. Swain’s Jury Panel Case Study: Jury Pool

Consider the following hypothetical scenarios:

- What if the panel of jurors consisted of 15 Black males?
- Would a higher count of Black panelists automatically lead to fair panel or impartial jury?
- Would you consider a panel of 21 Black males fair?
- Take 5 minutes to discuss your thoughts on above.





# 💬 Possible Discussion Questions 💬

Some discussion questions:

- What are some advantages of random sampling in this context? How might random sampling promote ethical practices?
- You can use use the jury panel examples to discuss how random sampling only leads to racial balance **on average** and how chance variability can lead to imbalance.
- How many Black males need to be on the panel for you to be convinced that the sampling process was followed fairly? Can use this question to motivate the gray area in defining notions of significance, especially as pertaining to legal issues.
- Is random sampling the fairest way to select the panel? One alternative is to use some kind of conditional random sampling to explicitly balance along certain demographic characteristics.

# 📝 Instructor Note 📝

There were 52 simulated jury panels which had 15 or less, and 1,599 simulated jury panels with 21 or less impaneled Black jurors.

# 4 Defining Sampling Bias

# 4. Sampling Bias: Defined

What is sampling bias?

- The outcome of a sampling technique that suggests that not every member of the population has an equal opportunity of being in a sample.
- Consequently, statistical bias in estimates may be observed. In other words, estimates cannot be trusted as reflecting or representing the population.
- **Not investigating and addressing sampling issues raises ethical concerns**

# 5 The Relationship between Sampling and Representativeness

# 5. Sampling and Representativeness

# 5. Sampling and Representativeness: Flint, Michigan Example

- In late 2014, the residents of Flint, Michigan had noticed and reported that children and adults were experiencing lead poisoning after the city had switched to Flint River for the city's water supply.
- An investigation into the drinking water supply identified that some of the water pipes were **leaded brass**. The city resumed services to the water supply after it noted that they would start a task force.
- The same year the Flint Drinking Water Task Force was established to help monitor and assess the quality of the city's water. The goal of the group was to sample drinking water from homes in areas where lead was expected, to prepare a drinking water quality report.

# 5. Sampling and Representativeness: Flint, Michigan Example

- In 2015, the Flint Drinking Water Task Force went door-to-door distributing water and asked for volunteer homeowners to submit water samples.
- The task force did not specify if random selection was used to select homes from those who volunteered to participate in water collection. The site chosen by the Task Force were labeled *sentinel* sites. All other volunteers were labeled non-sentinel/volunteers.
- The initial water water collection and household information came bi-weekly from 156 out of 1951 volunteer households.
- More households were included in the sentinel samples across time with a total 759 sentinel homes and 4041 non-sentinel.
    - The Sentinel Sites sampling technique: The members of the Task Force
    - Volunteer Sample: concerned homeowners throughout the city also volunteered to provide samples and information.

# 📝 Instructor Note 📝

The following real-world example can be used to highlight several important points.

- Ask students whether they have enough information to make sense of the sampling technique.
- Note that this is a form of convenience sampling (non-random), which can lead to bias in samples and real world consequences.
- This is an opportunity to engage in a conversation about the representation and under-representation using this example. See notes below. Ask students if results from the task force can be trusted.

Notes: - There are 51,045 residential properties in the city of Flint. Therefore, the task force sampled 0.30% (156) of residents in the initial sample. The sample size was eventually increased to 759 which is 1.48% of the residents. Including volunteer houses, 4 percent of households were sampled from Flint residents in the initial sampling. Later sampling include 9 percent of the Flint households providing water samples. - Ask students whether they have enough information to make sense of the sampling technique. - Note that this is a form of convenience sampling (non-random), which can lead to bias in samples and real

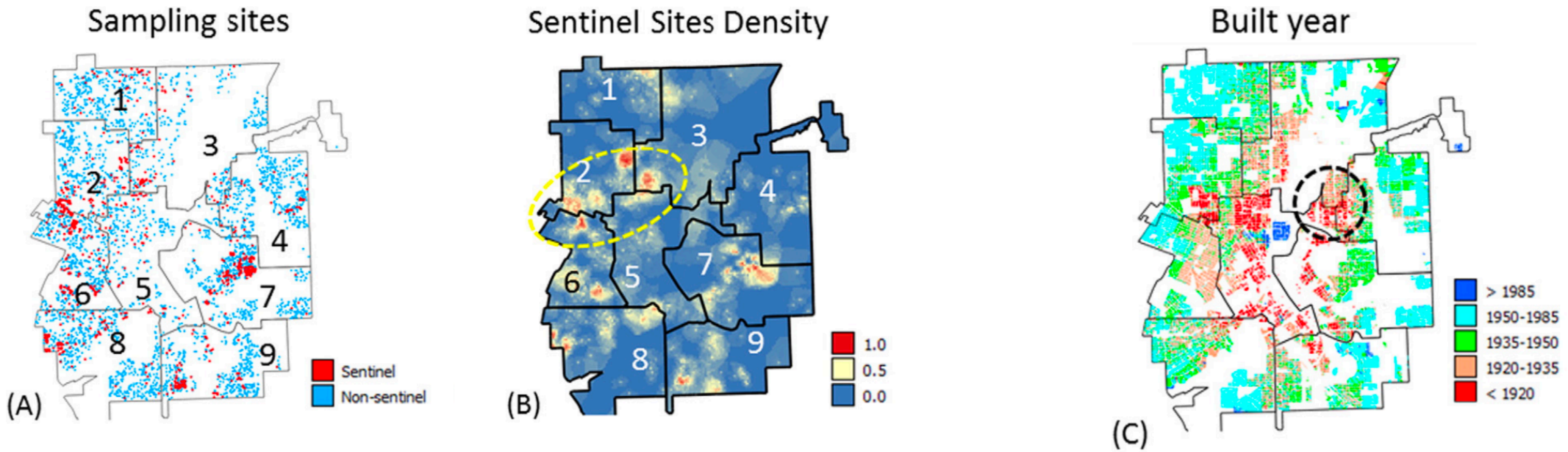
# 📝 Instructor Note 📝

- There are 51,045 residential properties in the city of flint. Therefore, the task force sampled .30% (156) of residents at the initial sample. The sample size has increased to 759 which is representative of 1.48% of the residents. Including volunteer houses 4 percent of household were sampled from flint residents in the initial sampling. Later sampling include 9 percent of the Flint household residents providing water samples.

# 5. Sampling and Representativeness: Example

- In 2017, Goovaerts (Goovaerts 2017) compared the housing characteristics and geographical distribution of households that were sampled (Sentinel and Volunteer/Non-sentinel) by the Flint Safe Drinking Water Task Force over two months in 2016.
- Visualizations illustrated sample differences between Sentinel Sites and volunteering sites (Non-sentinel), such as:
    - location/spacial reference in the city (Figure A)
    - density of sentinel sampled sites (Figure B)
    - year the households were built (Figure C)
    - material of water pipes (Figure D)
    - percent of households in the city above the poverty line (Figure E)
    - percent of elevated blood lead level (Figure F)

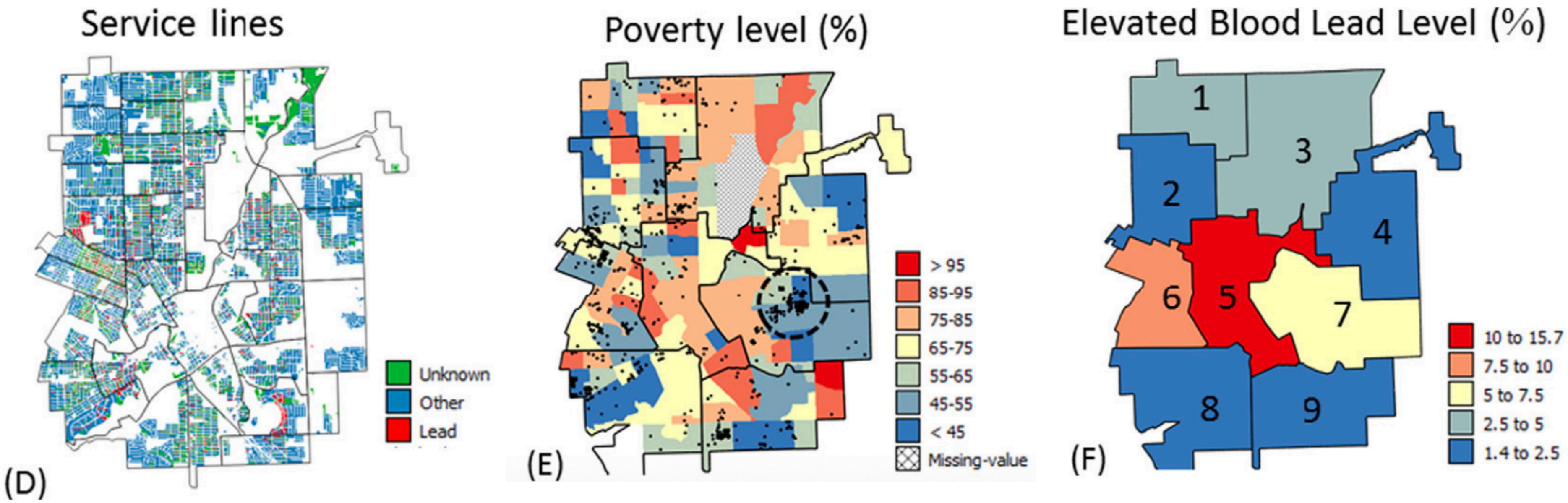
# 5. Sampling and Representativeness: Example


Sampling sites

Sentinel Sites Density

Built year

# 5. Sampling and Representativeness: Example



# 📝 Instructor Note 📝

Figure A can be used as a comparison visualization. Figure A is a map of Flint divided into 9 parts, which can be also be seen through out the other maps. Red points are sampled sites by the task force also labeled sentinel. The blue points are volunteer residents providing water samples. Comparing Figure A to Figure C, shows the year residences were built. Comparing Figure A to Figure B students can observe that most sentinel water sampling sites were built between 1935 to 1985. Homes built before 1935 were mostly identified as volunteer households. Consider also pointing out to students that both figure A and C show that there is a pattern of missing water sample information starting from the right hand corner of section 3 through the left hand corner of section 8.

Comparing Figure A to D, sentinel sites primarily are identified as having other or unknown water pipes. Some volunteer households were identified has having water serviced from lead pipes, but mostly were unknown and other pipes. In comparison to figure F, most sites sampled by the Flint task force experienced low levels of elevated lead in blood. However, volunteer household residents were more likely to have elevated lead in their blood samples.

# 📝 Instructor Note 📝

Figure E can be compared to figure A, C, D and F, which would help compare the percent poverty levels to sampling site, year the dwelling was built, service lines, and lead in blood levels.

# 5. Sampling and Representativeness: Discussion

Discuss in your groups (5 minutes):

- Figure B is a density/heat map of the sampled water collected by task force. Does the map suggest that sampled water could be representative of the whole city? Are there subsections of the city that are more represented?
- What characteristics are representative of those heavy sampled areas (year pipes were built, types of service line, the percent of residents who are at poverty level, and elevated blood lead levels)?
- Does there appear to be any relationships between the households that were sampled versus those that were volunteered?
- In Figure A, the white regions suggests that no samples were collected by the Task Force nor voluntarily. How do these areas compare to those that had been sampled?

# 6 Beyond Random Sampling

# 6. Beyond Simple Random Sampling: Stratified

- **Stratified Random Sampling**: If the population of interest consists of smaller groups, arrange the population into those groups and then randomly sample from each group. The sampling probability may be proportional to the size of another group to ensure equal representation across all groups (Som 2014).
    - Ex. Dividing Los Angeles house sales by geographical sub-divisions or by city socioeconomic classes.

# 6. Beyond Simple Random Sampling: Clustering

- Cluster Random Sampling: Divide a population into clusters based on some natural characteristic. Then randomly select the number of clusters (predetermined by researcher) to be included in the sample.
- Ex. Cluster students based on ACT scores and then randomly select cluster to evaluate different cluster characteristics.

# 6. Beyond Simple Random Sampling: Different Types of Random Sampling Techniques

- Systematic Random Sampling: A variant of random sampling. First determine the percent of the population to sample. Then, randomly select an observation within an interval (1 and 1/percent). Starting from the randomly selected unit add the interval upper bound to that observation to determine the next observation to include in the sample (Shahzad et al. 2023).
    - Ex. Suppose we are interested in sampling five percent of the population. A random number is first chosen between 1 and (1/0.05 =) 20 ; Then randomly select number is 12. Therefore, the first observation of the random sample is the 12th observation. Then the (12 + 20 =) 32nd, (32 + 20 =) 52nd etc (Som 2014).

# 6. Beyond Simple Random Sampling: Disadvantages of Random Sampling

- There are disadvantages to random sampling from a population (Emerson 2015) :
    - Time consuming and expensive
    - Not easy to implement with populations that have low incidence
    - Used to give an average representation of the population, when a data scientist is more concerned with the nuances.

# 6. Beyond Simple Random Sampling: Disadvantages of Random Sampling

- Alternative non-random sampling (Convenience and Snowballing) mechanisms may help gather data from a smaller population.
    - Example, collecting data from children with visual impairments who attend the California School for the Blind.
- However, these mechanism will limit the generalization of findings (Sauro and Lewis 2012).
- To help data scientists engage in ethical practices when analyzing data and reporting findings, even in cases where sampling information is unknown or highly complex, institutions and organizations have developed guiding principles.

# 7 Ethics and Sampling

# 7. Ethics and Sampling: ASA principles

Ethical Principles related to Sampling by the American Statistical Association (ASA, 2022) :

**Principle B**

The ethical statistical practitioner seeks to understand and mitigate known or suspected limitations, defects, or biases in the data or methods and communicates potential impacts on the interpretation, conclusions, recommendations, decisions, or other results of statistical practices.

**Principle D**

The ethical statistical practitioner does not misuse or condone the misuse of data. They protect and respect the rights and interests of human and animal subjects. These responsibilities extend to those who will be directly affected by statistical practices.

# 7. Ethics and Sampling: ASA principles

The following are recommended practices that may help meet Principle B:

- Use various research methods and/or statistical techniques to determine whether the sample is representative of the population of interest.
- Use a random sampling technique and a large enough sample. A small sample size can lead to statistical bias.
- Thoroughly explore data to identify patterns, clusters, and outliers that can confound results, which suggests that clusters represent different perspectives in the variable of interest.
- When available, evaluate the same research question in a new sample to ensure that the findings are replicated.
- Do not extrapolate findings, conclusions, and recommendations beyond statistical results.
- Be transparent about the extent to which findings can be generalized to population of interest and other groups.

# 7. Ethics and Sampling: ASA principles

The following are recommended practices that may help meet Principle D:

- Generalize findings to the population of interest only after confirming that the sample represents the population.
- Critically reflect on the research question and consider how the sampled population of interest and those excluded may be impacted by findings and future policy.

# 7. Ethics and Sampling: ASA principles

Data scientists are often tasked with analyzing data with little information about sampling methods. To engage in ethical data science practices, scientists should ask questions about how the data were collected.

Consider answering the following questions when working with a new dataset:

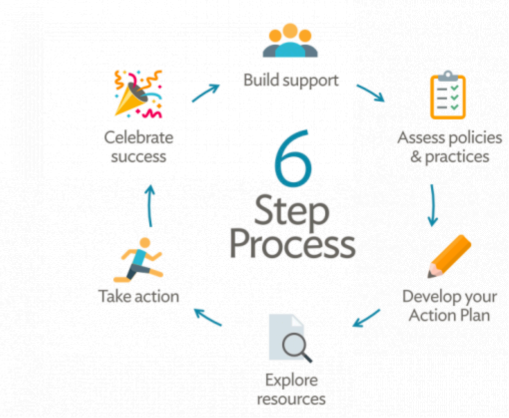
- Was the data collection technique under the control of investigators or was the process pre-determined by nature or happenstance?
- What are the apparent and unseen consequences related to the sampling method used?
- How should a researcher or analyst use such data to make inferences about the larger population of interest? In other words, could there be issues with generalizing findings?

# 7. Ethics and Sampling: Discussion

The American Heart Association and William J. Clinton Foundation, created the Healthy Schools Program (HSP) in 2006. The HSP aims to help administrators and teachers create an encouraging physically healthy and nutritious school environment for all.

To determine the efficacy for future implementation of HSP at other schools and to support nationwide policy change, researchers collected data from participating schools from 2007 through 2014.

The 6 Step Process



# 7. Ethics and Sampling: Discussion

Read through HSP_Codebook (pg. 3) to determine sampling procedures used to collect data from participating schools for the “Healthy Schools Program Participation and Inventory Data”. Take 5 minutes to critically reflect on the following questions within your group.

- Was the data collection technique under the control of investigators or was the process pre-determined by nature or happenstance? Was there enough information about the sampling technique to ensure that data are representative of the population of interest?
- What are the apparent and unseen consequences related to the sampling method used? How might the sampling technique influence responses? How should a researcher or analyst use such data to make inferences about the larger population of interest? In other words, could there be issues with generalizing findings?

# 7. Ethics and Sampling: Practices Related to Sampling

Unethical practices related to sampling:

- Under- or over-sampling a part of a population to confirm beliefs:
    - Purposefully or unintentionally excluding a minority group or oversampling a majority group which results in diminishing the perspective the minority group.
    - E.g., A meta analysis revealed that the majority of skin cancer studies and datasets lack dark skin images which has contributed to the under-diagnosis of skin cancer in men and women of color (Wen et al. 2022).
- Using sampling techniques known to produce bias:
    - E.g., (Rourke and Lakner 1989) determined that men are considerably under-sampled, fostering gender bias, using a sampling technique that relies on a random selection of telephone numbers and surveying the last birthday in the household.

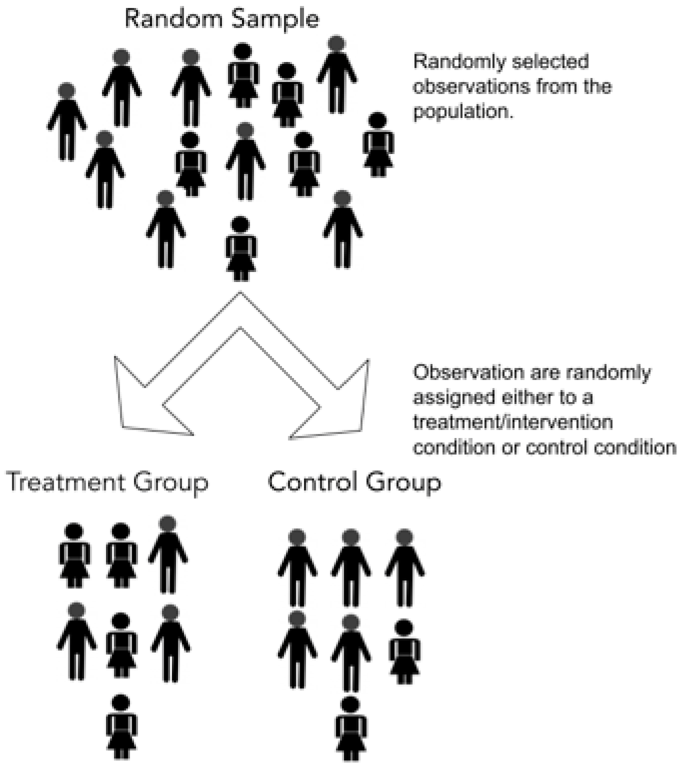
# 7. Ethics and sampling: Practices Related to Sampling

Unethical practices related to sampling:

- Purposefully altering data of one or more observations to meet sampling expectations:
  - After implementing a sampling technique, manipulating or fabricating data.
  - E.g., In 2010, a journalist had discovered that Andrew Wakefield had falsified data by picking and choosing data reported by children to find an association between the MMR vaccine and autism. Articles published in the British Medical Journal suggested that the data were manipulated for financial gain (Sathyanarayana Rao and Andrade 2011).
- Knowingly failing to disclose observed patterns beyond aggregated data to stakeholders:
  - By focusing on a large population, data scientists can fail to recognize patterns or heterogeneity across smaller groups.

# 8 Randomized Controlled Trials

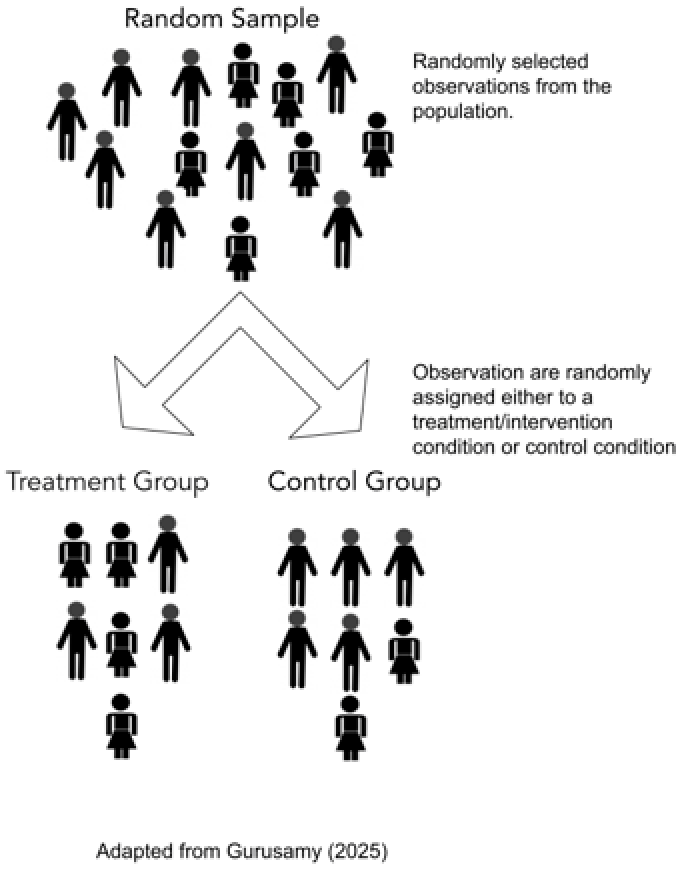
# 8. Randomized Controlled Trials


Random Sample
Randomly selected observations from the population.
Observation are randomly assigned either to a treatment/intervention condition or control condition
Treatment Group
Control Group

Adapted from Gurusamy (2025)

- Random sampling plays a crucial role in true experiments, also known as randomized controlled trials (RCTs).
  - In the simplest RCT, researchers randomly assign observations to either receive a treatment/intervention (Tx) or no treatment (control group) (Gurusamy 2025).
  - Goals of RCTs using data science techniques:
    1. Determine if a Tx had an effect on the observations that were assigned to the group.
    2. If so, assess how large of an effect the Tx had on an outcome that can be measured and compared in both the Tx and control group.

# 8. Randomized Controlled Trials


Random Sample
Randomly selected observations from the population.
Observation are randomly assigned either to a treatment/intervention condition or control condition
Treatment Group
Control Group

Adapted from Gurusamy (2025)

- Randomization is a crucial tool: *on average*, treatment group and control group should be similar along important characters (age, race, gender, health etc)
- Without randomization, the treatment and the control group could yield different outcomes because the groups are fundamentally different (e.g. one group is younger than the other) not because the Tx had any effect.
- This is why we say “correlation does not imply causation!”

# 8. Randomized Controlled Trials

Hypothetical example: researchers are interested in the effects of caffeine dosage on cyclists performance (Beedie et al. 2006).

- Researchers took a sample of up-and-coming cyclists
- Cyclists were randomly assigned to three different groups, two Tx groups and one control group
  - Tx group 1: received 4.5 mg of caffeine capsule - moderate caffeine group
  - Tx group 2: received 9 mg of caffeine capsule - high caffeine group
  - Control group (Placebo): received a powder capsule with no caffeine
- After consuming the capsule cyclist were then put in a simulated cycling performance test

# 💬 Possible Discussion Questions 💬

- Assume that the results suggested that cyclists assigned to the moderate caffeine group completed the cycling simulation faster than all others. Can we conclude that moderate caffeine benefits all cyclists in the population?
  - If the sampling frame is the same as the population, and the sample size was large, we could probably conclude that moderate caffeine was beneficial.
  - If the sampling frame was not the same as the population then we can only make statements about the frame.
  - If the sample size was small, then the treatment groups may differ by chance (randomization only guarantees balance on average)
- Assume, the researchers noticed that one randomly assigned cyclist in the moderate group was a winner of the Giro d'Italia Grand Tour and had the fastest completion time across all observations. Can we still conclude that moderate caffeine benefits cyclists?
  - Depends on the sampling frame and the sample size. If the Giro winner was just one of many cyclists in the moderate group, his results might not skew our estimates.
  - If the sampling frame includes only professional cyclists then it may be find that this rider is included.

The example is an adaptation of (Beedie et al. 2006).

# 8. Randomized Controlled Trials

- Statistical differences are valid only if the sampling frame matches the population and if the sample is “balanced” across the treatment groups.
- If one or more observations do not represent the population, confounding variables (variables not apart of the study that influence results) may explain any differences found between the Tx and control group on the outcome of interest.

# 8. Randomized Controlled Trials: Ethics

- Data scientists should inquire about sampling methods because different techniques could influence the finding of statistical differences between the Tx and control groups.
- Consequently, a data scientist may wrongfully, knowingly, or unknowingly report invalid results.
- In addition to sampling techniques, data scientists should also aim to understand additional methodological processes and practices used to construct the RCT.
- Engaging in morally good data science also includes purposefully inquiring about the RCT design, such as participant consent, deception/harmful practices participants had to endure, and benefits/compensation.
- Consider that all steps taken during the research or RCT process may influence findings.

# Randomized Controlled Trials: Ethics

The Helsinki report is a historically unparalleled report that outlined a code of research ethics which inform RCTs. The World Medical Assembly established the document in 1964 (World Medical Assembly 1964) in response to the Nuremberg trails and ethical concerns and documentation that spurred from experimental atrocities that occurred during the Holocaust (Schmidt, Mehring, and McMillan 2010).

# Randomized Controlled Trials: Ethics

The WMA Declaration of Helsinki is now in its tenth revision. There are currently 37 principles which are categorized under 10 broad categories WMA Declaration of Helsinki. The following categories and summarized principles which may closely relate to data, analytics, and sampling:

- Free and Informed Consent theme suggests that:
  - consent to participate in a research study must be voluntary.
  - participants are aware that they may withdraw at any point throughout the study.
  - researchers need to confirm that participants understand the consent forms.

# Randomized Controlled Trials: Ethics

- Risks, Burdens, and Benefits theme suggests that:
  - research including human participants may be conducted if the importance of the research objective outweighs the risk to the participants.
- Minimizing Harm theme suggests that:
  - research should be conducted in a manner that avoids or minimizes harm to the environment and participants.
  - harmed participants due to study's treatment are give appropriate compensation.

# 8. Randomized Controlled Trials: Tuskegee Syphilis Study

The Tuskegee Syphilis study is a primary example that demonstrates the implications of not upholding research ethics (“Tuskegee Files” 1997).

- Conducted by the U.S. Public Health Service (USPHS) from 1932 to 1972.
- 399 Black males with late latent-syphilis and 201 Black males without the disease were recruited to participate in an “experimental study” that would provide treatment for their “bad blood”.
- However, researchers used deception and never provided treatment and prevented participants to receive treatment outside of the study.
- Participants’ compensation included free meals, free medical exams, and burial insurance.
- According to records approximately 28-100 participants died from their syphilis diagnosis.

# 9 Observational Studies

# 9. Observational Studies

> “Excellent methods of analysis will not salvage a poorly designed *observational* study”
>
> – Rosenbaum, 2010

- Simple observational study:
  - is an empirical study of effects caused by treatments when a randomized RCT is unethical or not feasible (Rosenbaum 2010).
  - Observations are not subjected to randomization, but rather natural circumstances divide individuals into groups, those with and/or without the natural circumstance.

# 💬 Possible Discussion Questions 💬

<<<<<<< Updated upstream The following are some examples where observational studies may be needed because randomization is not ethical:

- Classic examples include measuring the effect of smoking or drinking alcohol on health. In general, cannot ethically randomly assign any exposure that may cause harm.
- 10 Randomly denying access to helpful treatments, etc., when it is reliably available would also be unethical. This includes

unethical. This includes denying access to basic needs and rights around education, housing etc.

Stashed changes + Examples that may violate autonomy and personal choice?

The following are questions for student and points to be made

- What are some examples where observational studies may be needed because randomization is not ethical?
  - Classic examples include measuring the effect of smoking or drinking alcohol on health. In general, cannot ethically randomly assign any exposure that may cause harm.

# 9. Observational Study

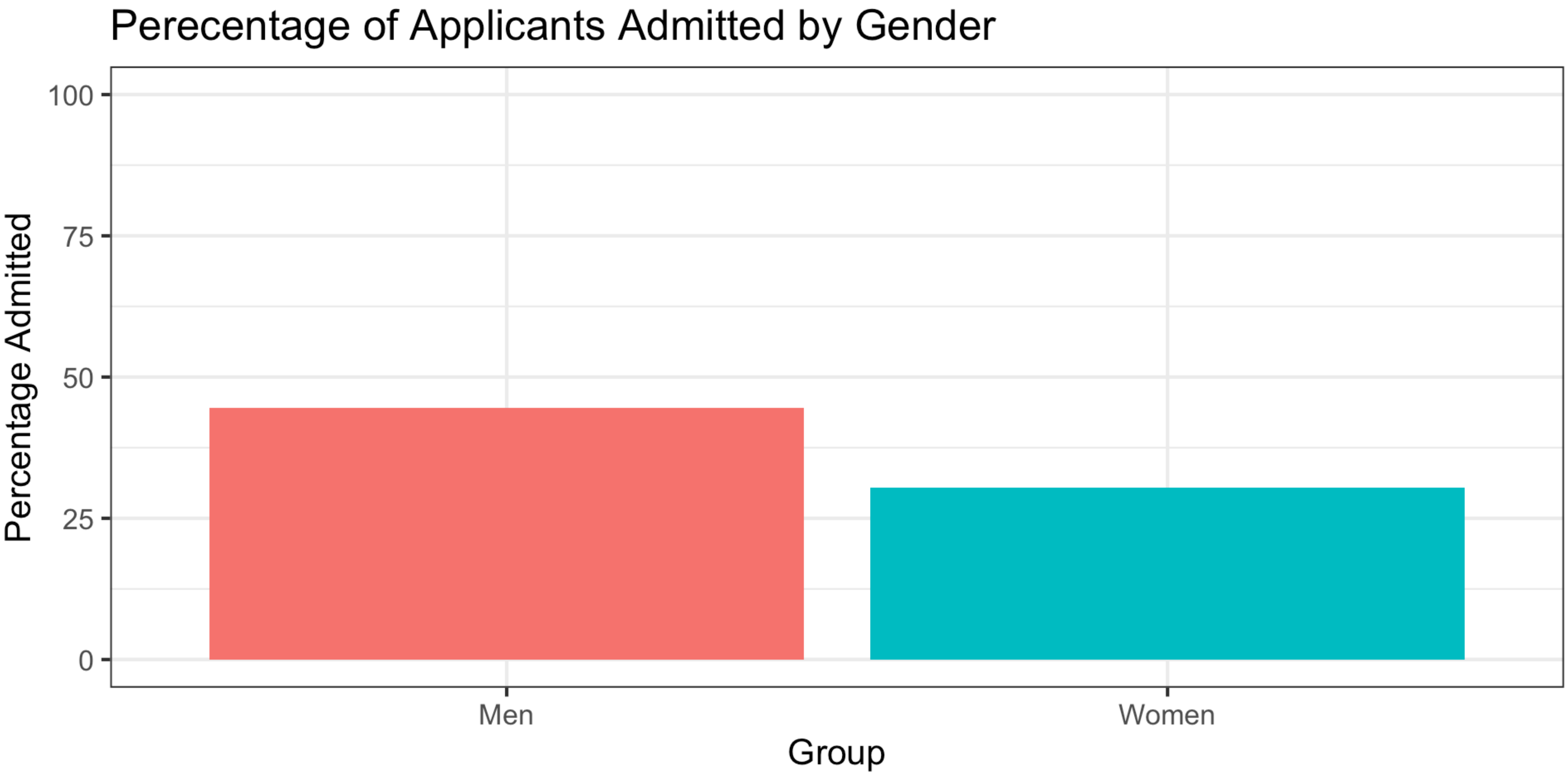
- Two types: natural or structured:
  - **Natural**: observe people or other units in an environment in which behaviors naturally occur.
    - Ex. documenting shoppers behaviors at a mall.
  - **Structured**: carefully document one or more specific behaviors in a particular setting.
    - Most medical observational studies are structured.
- Researchers take extreme caution and use various methodological techniques when using an observational study to evaluate Tx effects.
- Methodological practices include participant matching based on background features of interest, such as age, sex, ethnicity. Creating pairs is key to determine if the Tx truly impacted the participants in the Tx group.

# 9. Observational Study

- Observational studies face many challenges
  - In general, data scientists should pay special attention to observations recruitment, study procedures, and survey/assessment selection.
  - Since, observational studies are not true experiments, researchers and data scientists can expect bias in samples and lower confidence in analytic findings along with other limitations Jhangiani, Cuttler, Leighton, 2019.
  - Ethical practices that data scienists can engage while working with data from observational studies includes:
    - Communicating with stakeholders that findings are not absolute truths.
    - Educating readers that correlation does not equal causation
    - Use statistical techniques, such as weights or conservative statistics, to deal with unequal representation and analytical confidence.

# 11 Simpson’s Paradox

# 10. Simpson's Paradox: UC Berkeley Graduate Admissions



# 💬 Possible Discussion Questions 💬

What are some possible explanations for the discrepancy in the previous plot?

Possibilities:

- The difference is due to random chance (sampling variability)
- Female applicants were less qualified than male applicants
- UC Berkeley was biased against women
- Confounding or lurker variables (the real explanation, described in the following slides)

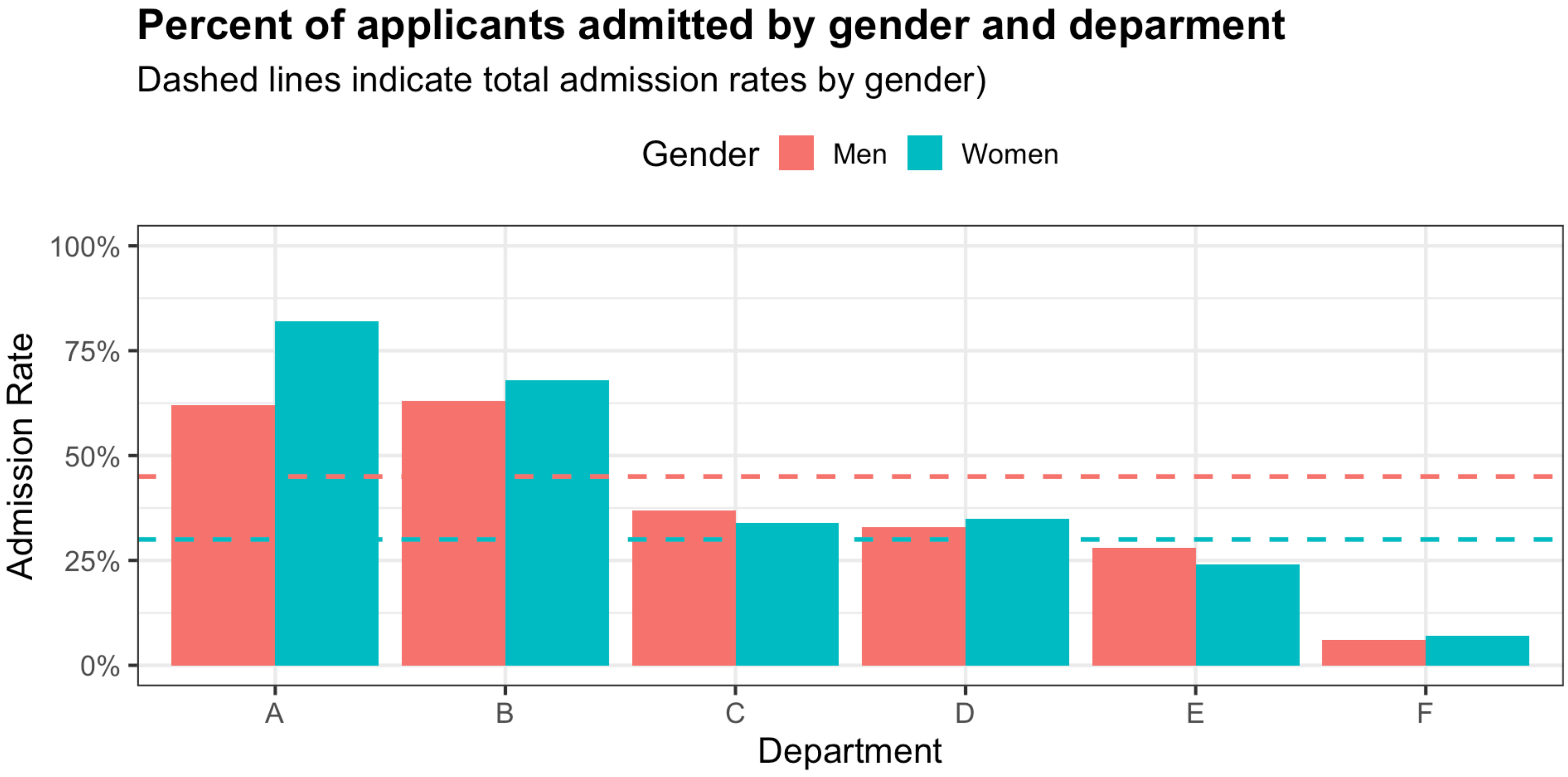
This example can also fit into the storytelling and visualization module to highlight the number of different possible explanations for the same (limited) set of data.

# 📝 Instructor Note 📝

From wikipedia:

The admission figures for the fall of 1973 showed that men applying were more likely than women to be admitted, and the difference was so large that it was unlikely to be due to chance. However, when taking into account the information about departments being applied to, the different rejection percentages reveal the different difficulty of getting into the department, and at the same time it showed that women tended to apply to more competitive departments with lower rates of admission, even among qualified applicants (such as in the English department), whereas men tended to apply to less competitive departments with higher rates of admission (such as in the engineering department). The pooled and corrected data showed a “small but statistically significant bias in favor of women”.

# 10. Simpson’s Paradox: UC Berkeley Graduate Admissions

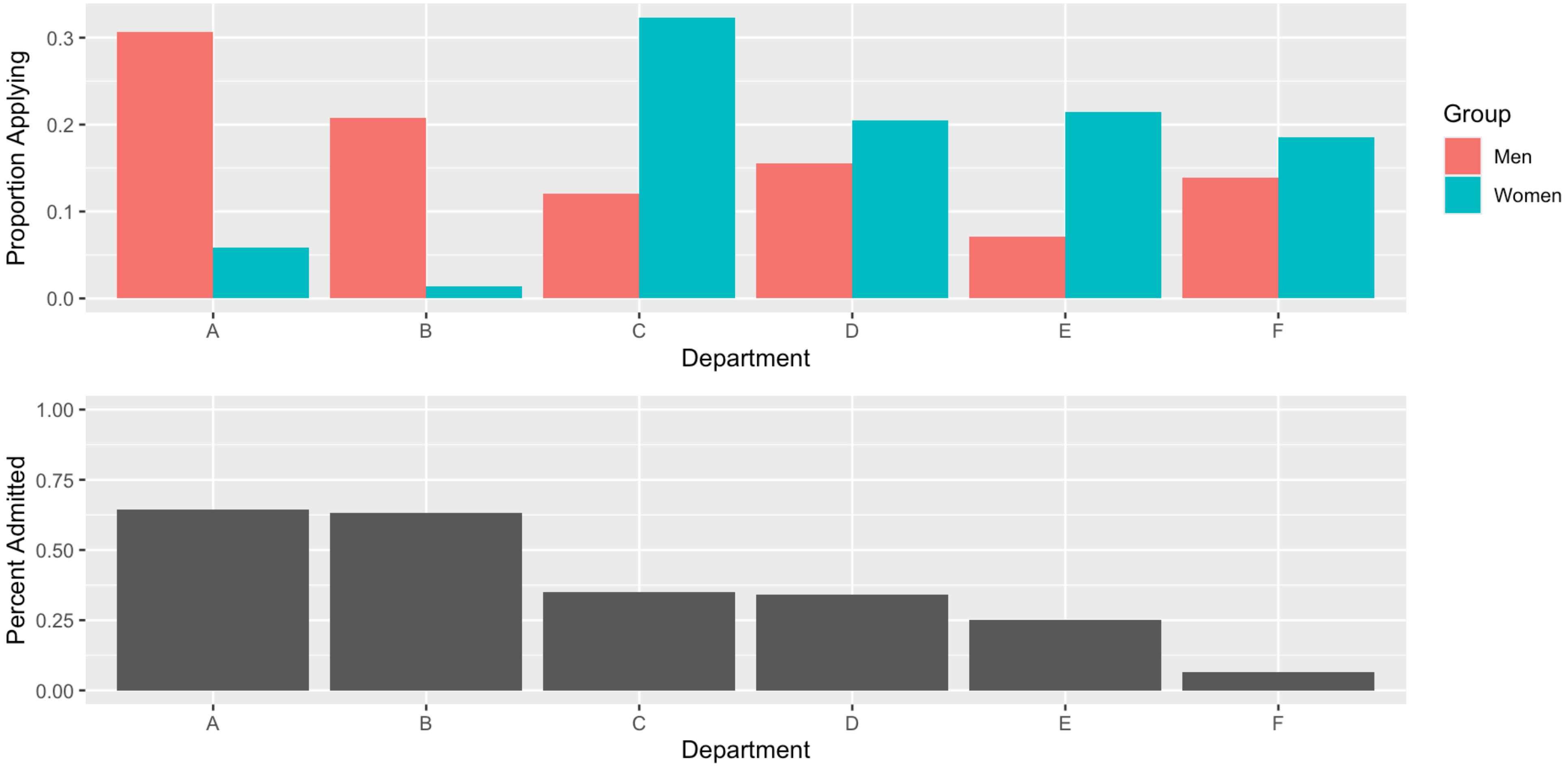


Percent of applicants admitted by gender and deparment

Dashed lines indicate total admission rates by gender)

How is this possible?

# 10. Simpson's Paradox: UC Berkeley Graduate Gender Discrimination



# 📝 Instructor Note 📝

For the previous plots, point out that that women tended to apply to departments that had much lower acceptance rate. There was no major gender differences, once accounting for departmental preference.

# 10. Simpson’s Paradox: Define

- Simpson’s Paradox is a statistical phenomenon in which a trend that appears across multiple subgroups tends to **reverse or disappear** when the groups are combined.
- Correlations between variables can exist for many different reasons. Sometimes correlation is due to some kind of causation, but often there is a simple explanation related to confounding variables or selection effects. Stratifying according to confounding variables can make associations disappear.
- Randomized control trials balance confounders (e.g., departmental preference) across groups, but aren’t always ethically feasible.
- Simpson’s paradox highlights one way in which statistics (from observational data) can be misused or misinterpreted.

# 💬 Possible Discussion Questions 💬

Summary discussion questions

- Discuss some of the tradeoffs between conducting a randomized experiment and using observational data.
- What are the ethical benefits and barriers in running randomized experiments?
  - Pro: Results will reflect the population (on average)
  - Challenges: previously discussed issues with the ethics of assigning treatments
- What are the ethical concerns associated with “natural experiments”?
  - Data collection & privacy concerns
  - Dangers in drawing conclusions with reasoning over relevant confounders. May not know what confounders to include. Expertise needed to reason about results carefully.
  - Can increase prevalence of misinformation or misleading conclusions if not reported carefully.

# References

Beedie, Christopher J., Elizabeth M. Stuart, Damian A. Coleman, and Abigail J. Foad. 2006. “Placebo Effects of Caffeine on Cycling Performance.” *Medicine & Science in Sports & Exercise* 38 (12): 2159–64. https://doi.org/10.1249/01.mss.0000233805.56315.a9.

Carrington, Tucker. 2023. “Annual Review of Criminal Procedure.” *The Georgetown Law Journal* 52.

Dinov, Ivo D., Nicolas Christou, and Robert Gould. 2009. “Law of Large Numbers: The Theory, Applications and Technology-Based Education.” *Journal of Statistics Education* 17 (1): 1. https://doi.org/10.1080/10691898.2009.11889499.

Emerson, Robert Wall. 2015. “Convenience Sampling, Random Sampling, and Snowball Sampling: How Does Sampling Affect the Validity of Research?” *Journal of Visual Impairment & Blindness* 109 (2): 164–68. https://doi.org/10.1177/0145482X1510900215.

Goovaerts, Pierre. 2017. “Monitoring the Aftermath of Flint Drinking Water Contamination Crisis: Another Case of Sampling Bias?” *Science of The Total Environment* 590-591 (July): 139–53. https://doi.org/10.1016/j.scitotenv.2017.02.183.

Gurusamy, Kurinchi. 2025. *A Guide to Performing Systematic Reviews of Health and Disease*. UCL Press. https://doi.org/10.2307/jj.18255586.

Kotu, Vijay, and Bala Deshpande. 2019. "Data Science Process." In, 19–37. Elsevier. https://doi.org/10.1016/B978-0-12-814761-0.00002-2.

Rosenbaum, Paul R. 2010. *Design of Observational Studies*. Springer Series in Statistics. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4419-1213-8.

Rourke, Diane O', and Edward Lakner. 1989. "GENDER BIAS: ANALYSIS OF FACTORS CAUSING MALE UNDERREPRESENTATION IN SURVEYS." *International Journal of Public Opinion Research* 1 (2): 164–76. https://doi.org/10.1093/ijpor/1.2.164.

Sathyanarayana Rao, Ts, and Chittaranjan Andrade. 2011. "The MMR Vaccine and Autism: Sensation, Refutation, Retraction, and Fraud." *Indian Journal of Psychiatry* 53 (2): 95. https://doi.org/10.4103/0019-5545.82529.

Sauro, Jeff, and James R. Lewis. 2012. *Quantifying the User Experience: Practical Statistics for User Research*. Chantilly, UNITED STATES: Elsevier Science & Technology. http://ebookcentral.proquest.com/lib/ucsb-ebooks/detail.action?docID=872581.

Schmidt, H, S Mehring, and Dr J McMillan. 2010. "INTERPRETING THE DECLARATION OF HELSINKI (2008): "MUST", "SHOULD" AND DIFFERENT KINDS OF OBLIGATION." *Medicine and Law* 29.

Shahzad, Usman, Ishfaq Ahmad, Nadia H. Al-Noor, Muhammad Hanif, and Ibrahim Mufrah Almanjahie. 2023. “Robust Estimation of the Population Mean Using Quantile Regression Under Systematic Sampling.” *Mathematical Population Studies* 30 (3): 195–207. https://doi.org/10.1080/08898480.2022.2139072.

Som, Ranjan Kumar. 2014. *Practical Sampling Techniques*. 2nd edition, revised and expanded. Statistics: Textbooks and Monographs, volume 148. Boca Raton London New York: CRC, Taylor & Francis Group.

“Tuskegee Files.” 1997.

Wen, David, Saad M. Khan, Antonio J. Xu, Hussein Ibrahim, Luke Smith, Jose Caballero, Luis Zepeda, et al. 2022. “Characteristics of Publicly Available Skin Cancer Image Datasets: A Systematic Review.” *Lancet Digit Health* 4: e64–74.

World Medical Assembly, 18th. 1964. “DECLARATION OF HELSINKI.”