Week 3 - Sampling and Study Design

Ryan Miller

Week #3 Outline

- Video #1
  - Sampling from a Population
- Video #2
  - Sampling Examples
- Video #3
  - Study Design
- Video #4
  - Randomized Experiments

Introduction

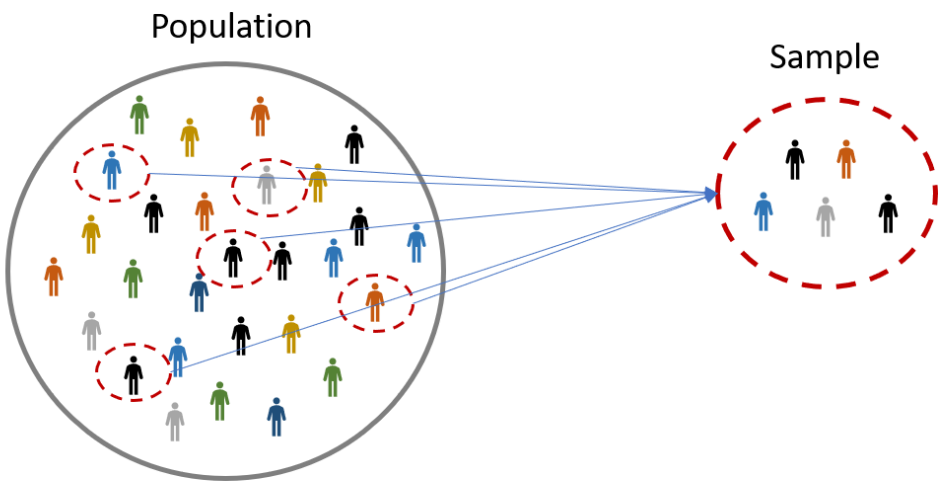
- So far, we've focused on methods for *describing* data
  - Descriptive statistics and graphs allow us to understand *the trends within our data*

Introduction

- So far, we've focused on methods for *describing* data
  - Descriptive statistics and graphs allow us to understand *the trends within our data*
- While what's happening in our data is interesting, we really would like *generalize* these findings to a broader set of cases
  - For example, we might use the fact that 14 of 16 infants chose the "helper" character as evidence that *all* infants can identify friendly behavior

Populations

- Statisticians call this broader, comprehensive set of cases a **population**
  - It is extremely unusual for researchers to have data on the entire population, it is much more common to have access to a **sample** (a subset of cases from the population)



Sampling

When analyzing sample data, statisticians have two primary concerns:

1. **Sampling Bias**: If some cases are preferentially selected into the sample, the data may no longer represent the entire population

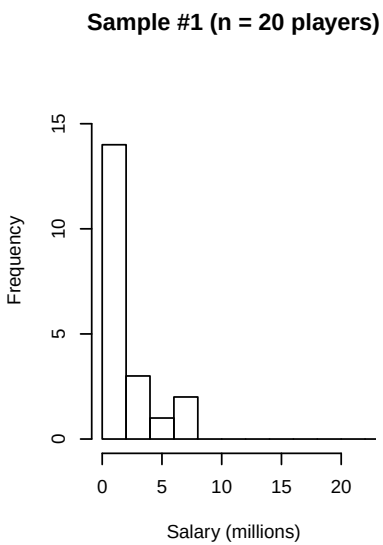
Sampling

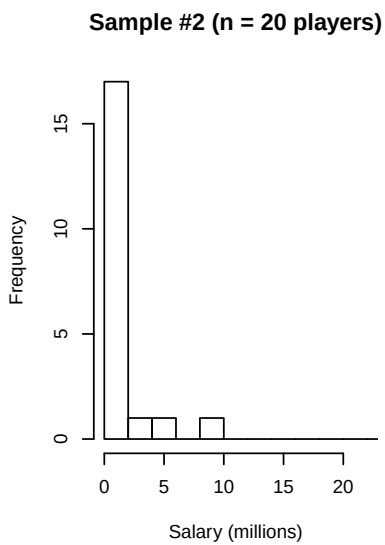
When analyzing sample data, statisticians have two primary concerns:

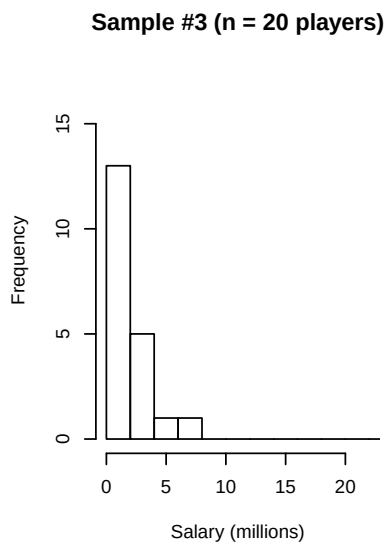
1. **Sampling Bias**: If some cases are preferentially selected into the sample, the data may no longer represent the entire population
2. **Sampling Variability**: Even if all cases had an equal chance of being selected, many different subsets are possible. So, any trends seen in a single sample could possibly be explained by the randomness involved in which cases ended up in that sample.

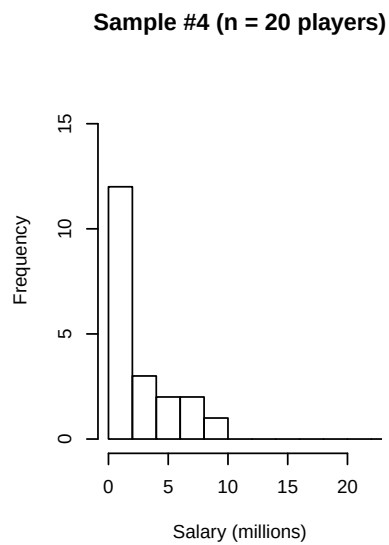
Example - Sampling Variability

If every NFL player has an equal chance of being sampled:



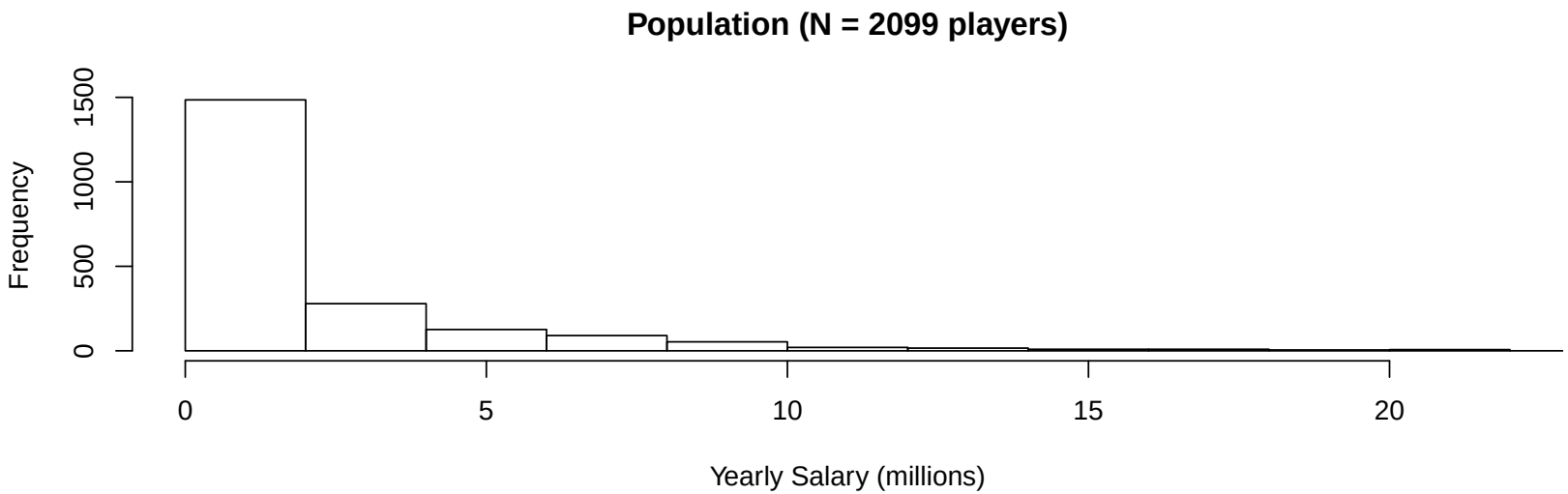


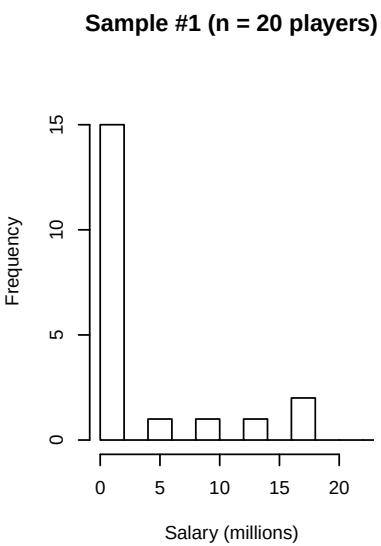


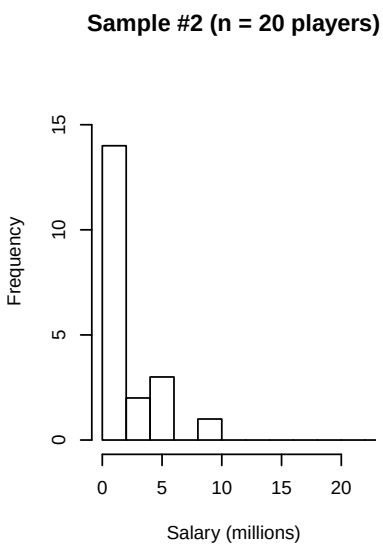


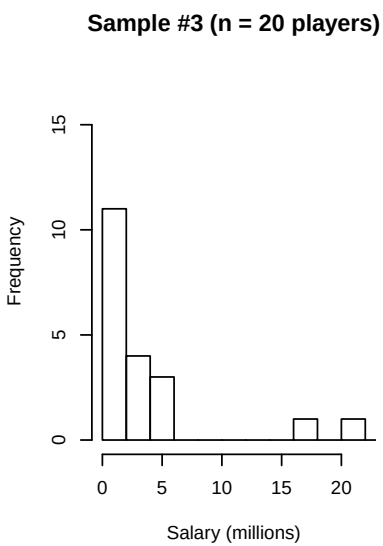
Example - Sampling Bias

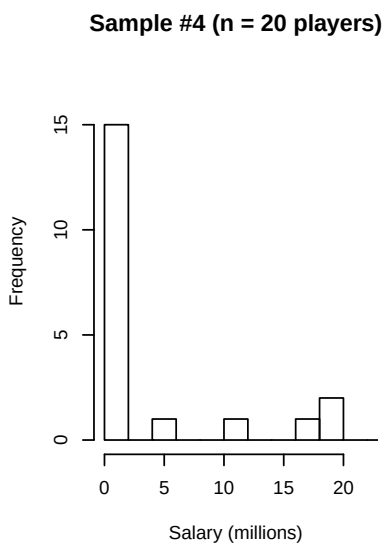
If Quarterbacks are 10x more likely to be sampled:





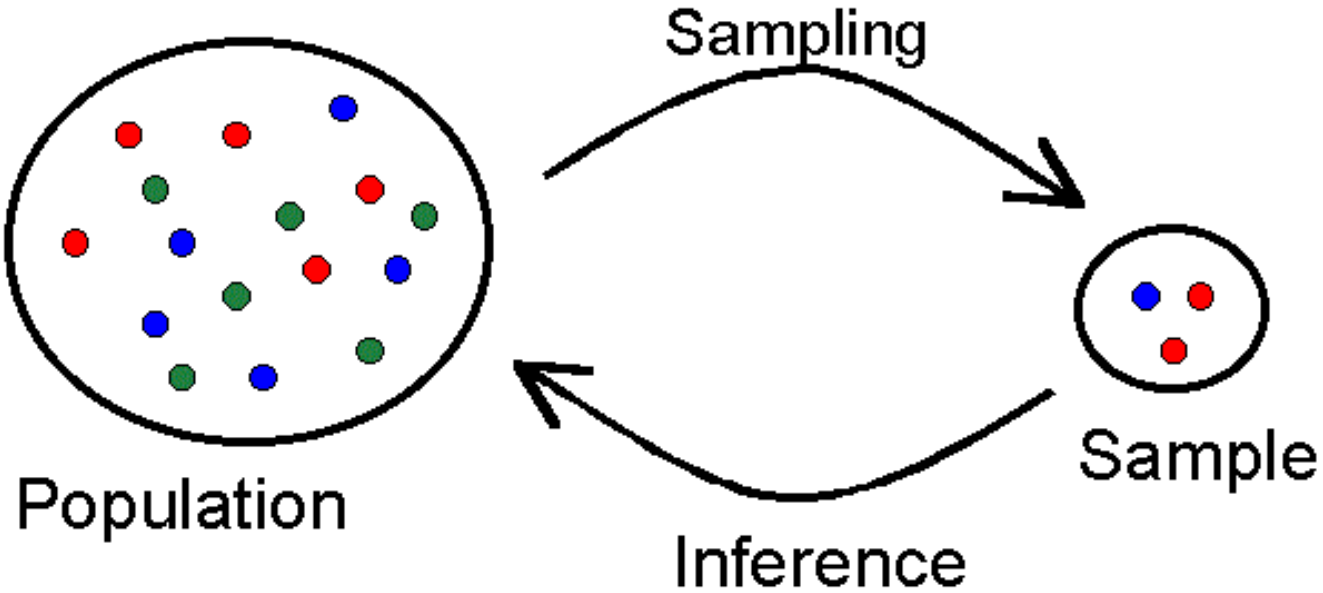






Statistical Inference

- A *fundamental goal* of statisticians is to use information from a sample to make *reliable* statements about a population
  - This idea is called **statistical inference**



- Sampling bias and sampling variability are two obstacles to statistical inference (as are confounding variables)

Statistical Inference - Notation

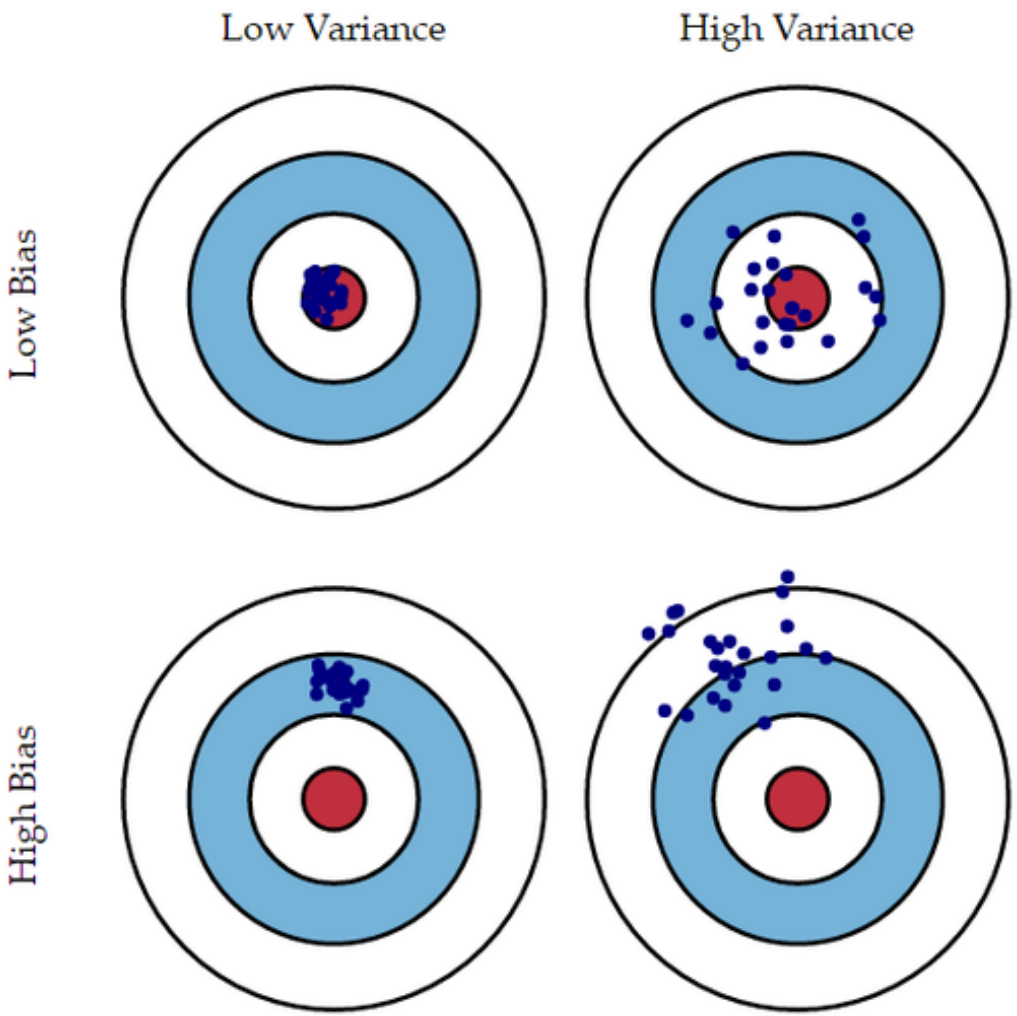
Statisticians use different mathematical *notation* to distinguish *population parameters* (things we want to know) from *estimates* (things derived from a sample). The table below provides a summary:

	Population Parameter	Estimate (from sample)
Mean	μ	$\bar{x}$
Standard Deviation	σ	s
Proportion	p	$\hat{p}$
Correlation	ρ	r
Regression	β_0, β_1	b_0, b_1

For example, μ is the mean of the target population, while $\bar{x}$ is the mean of the cases that ended up in the sample.

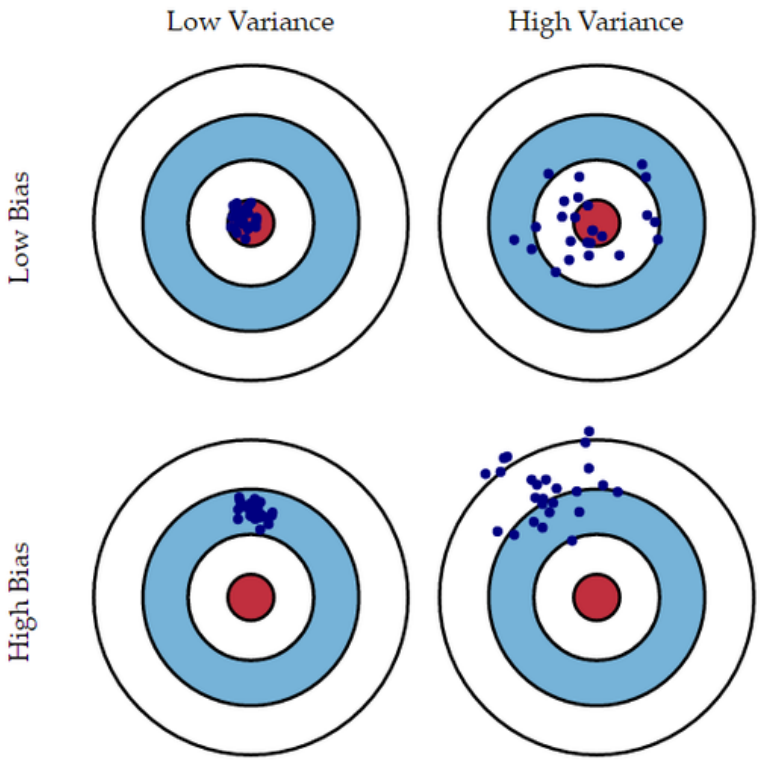
Bias and Variability

In summary, we've now discussed *two reasons* why an estimate might not accurately reflect a population parameter, **bias** and **variability**:



The Role of Sample Size

- Taking a larger sample will *reduce sampling variability*, but it will *not reduce sampling bias*



Sampling Words from The Gettysburg Address

- This fall, I showed by MATH-256 students the text of the Gettysburg Address (target population), asking each of them to come up with a *representative sample* of 5 words
  - I then had them summarize their sample using the *sample mean* (ie: the average word length of their 5 words)

Sampling Words from The Gettysburg Address

- This fall, I showed by MATH-256 students the text of the Gettysburg Address (target population), asking each of them to come up with a *representative sample* of 5 words
  - I then had them summarize their sample using the *sample mean* (ie: the average word length of their 5 words)
- If their samples were actually *representative*, their sample means would cluster randomly around the population mean

The Gettysburg Address

Four score and seven years ago, our fathers brought forth upon this continent a new nation: conceived in liberty, and dedicated to the proposition that all men are created equal.

Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battlefield of that war.

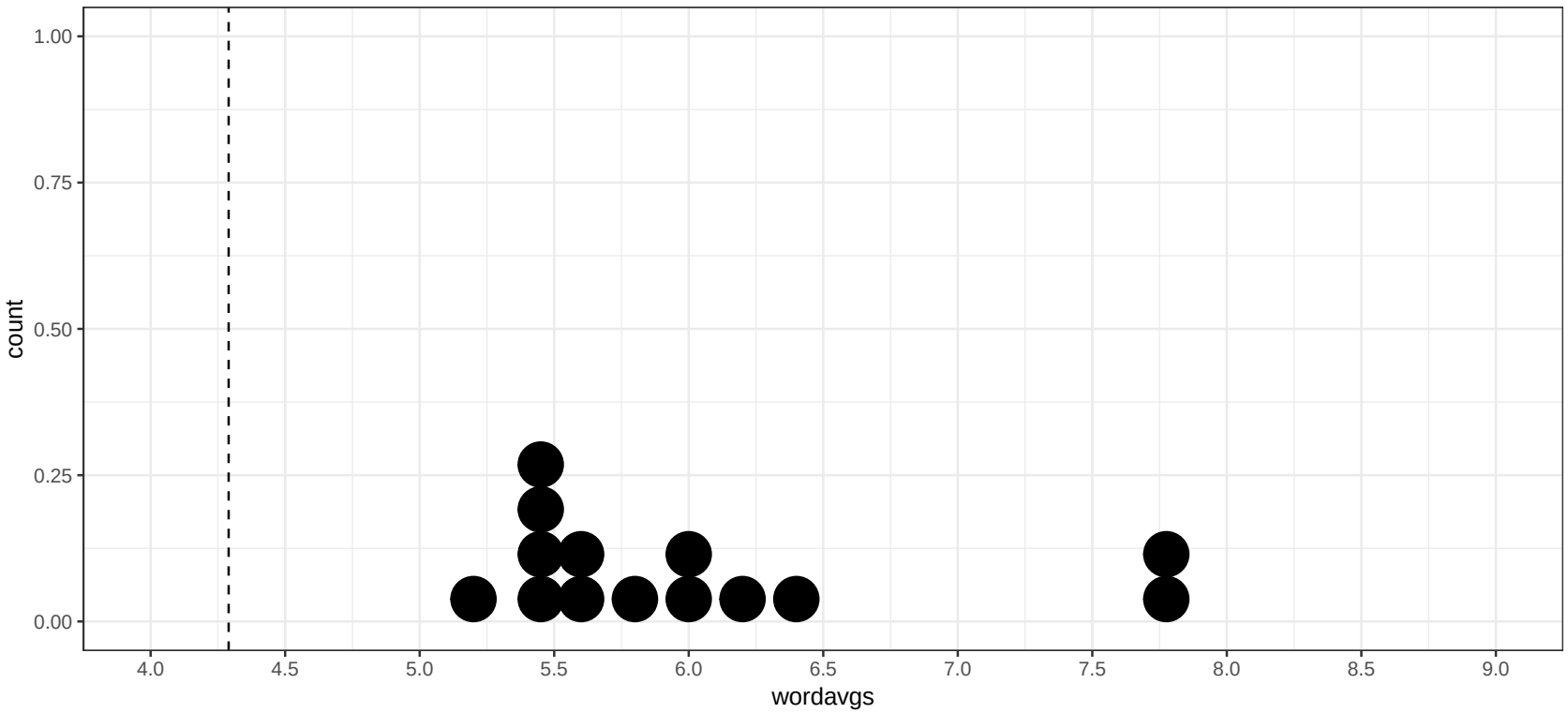
We have come to dedicate a portion of that field as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this.

But, in a larger sense, we cannot dedicate, we cannot consecrate, we cannot hallow this ground. The brave men, living and dead, who struggled here have consecrated it, far above our poor power to add or detract. The world will little note, nor long remember, what we say here, but it can never forget what they did here.

It is for us the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us, that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion, that we here highly resolve that these dead shall not have died in vain, that this nation, under God, shall have a new birth of freedom, and that government of the people, by the people, for the people, shall not perish from the earth.

Results

- The population's mean word length, in *statistical notation*, is $\mu = 4.295$
  - Why were the MATH-256 students' samples so far off?



Sampling Bias

- The answer is **sampling bias**, or a *systematic tendency* for some cases from the population to be more likely to make it into a sample than others
  - In the Gettysburg Address example, students tended to pick longer words
  - For example, perhaps the longer words stood out more prominently, or the students felt awkward selecting multiple two-letter words in their sample

Simple Random Sampling

- The *ideal sampling procedure* is **simple random sampling**, a protocol where each case in the target population has an identical chance of ending up in the sample
- Do you think it would be easy or hard to collect a simple random sample of Xavier students?

Simple Random Sampling

- The *ideal sampling procedure* is **simple random sampling**, a protocol where each case in the target population has an identical chance of ending up in the sample
- Do you think it would be easy or hard to collect a simple random sample of Xavier students?
  - It would actually be quite hard since the University is unlikely to give you a list of all enrolled students
  - Often statisticians will attempt to get representative samples in other ways

Other Sampling Protocols

- A **convenience sample** is exactly what the name suggests, a sample that is easily collected (ie: low monetary or time costs)
  - Convenience samples are *not random*, but they can be representative if carefully selected
  - You might be able to get a representative sample by standing near the center of campus on a typical day and stopping everyone who walks by

Other Sampling Protocols

- A **convenience sample** is exactly what the name suggests, a sample that is easily collected (ie: low monetary or time costs)
  - Convenience samples are *not random*, but they can be representative if carefully selected
  - You might be able to get a representative sample by standing near the center of campus on a typical day and stopping everyone who walks by
- A **stratified sample** is a more complex scheme where the population is broken into similar subcategories, which are sampled separately (typically simple random sampling)
  - The statistical methods we cover in this course won't be sufficient for this sampling scheme
  - Nevertheless, we should be able to recognize stratified sampling for precisely that reason

Practice

For each scenario, determine whether it describes a *population* or a *sample*, as well as whether or not the sample is biased.

1. To estimate the size of trout in a lake, an angler records the weight of the 12 trout he catches over a weekend
2. A subscription-based music website tracks the listening history of its active users
3. The Department of Transportation announces that of the 250 million registered cars in the US, 2.1% are hybrids
4. A car rental company installs an experimental data collection device on the first 20 vehicles from an alphabetized list of license plates

Practice (solution)

1. This is a sample and it's biased, there is no way that the angler has an equal chance of catching every trout in the lake
2. This is a population because it includes all of the website's users
3. This is a population because it's safe to assume the DOT has registration on the overwhelming majority of cars in the US
4. This is a sample and it's unbiased, there is no reason to believe that license plate numbers are related to anything meaningful about each vehicle

Introduction (Study Design)

- So far, we've seen how *sampling* can influence the trends seen in our data, but there are other aspects of data collection that we need to consider
- Suppose researchers want to test a new COVID-19 treatment, how would you design a study to determine if it is effective or not?

Introduction (Study Design)

- So far, we've seen how *sampling* can influence the trends seen in our data, but there are other aspects of data collection that we need to consider
- Suppose researchers want to test a new COVID-19 treatment, how would you design a study to determine if it is effective or not?
- A meaningful study must compare the new treatment with something else
  - Therefore, we'll either need to collect two samples, or proactively split a single sample into two

Two Types of Studies

These two possibilities lead us to distinguish between two types of studies:

- **Observational studies**: the explanatory and response variables are *observed* by the researchers (separate samples)
- **Experimental studies**: the explanatory variable is *assigned* by the researchers (the researchers split up a single sample)

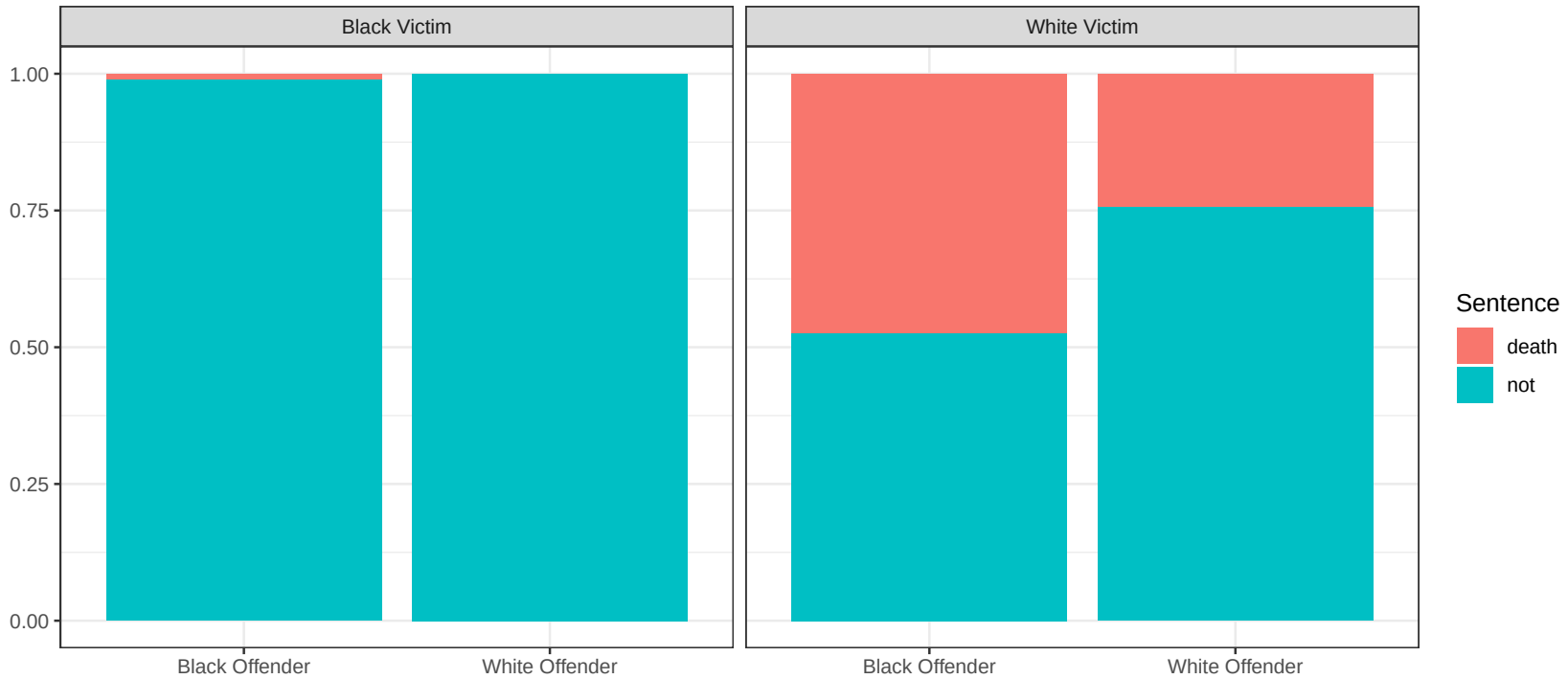
Observational Studies

- We've already seen an example observational study in the Florida Death Penalty Sentencing case study
- Recall the researchers recorded the race of the offender, as well as whether the offender was sentenced to the death penalty or not
  - Did the offender's race appear associated with their sentence?

	death	not
black	38	142
white	46	152

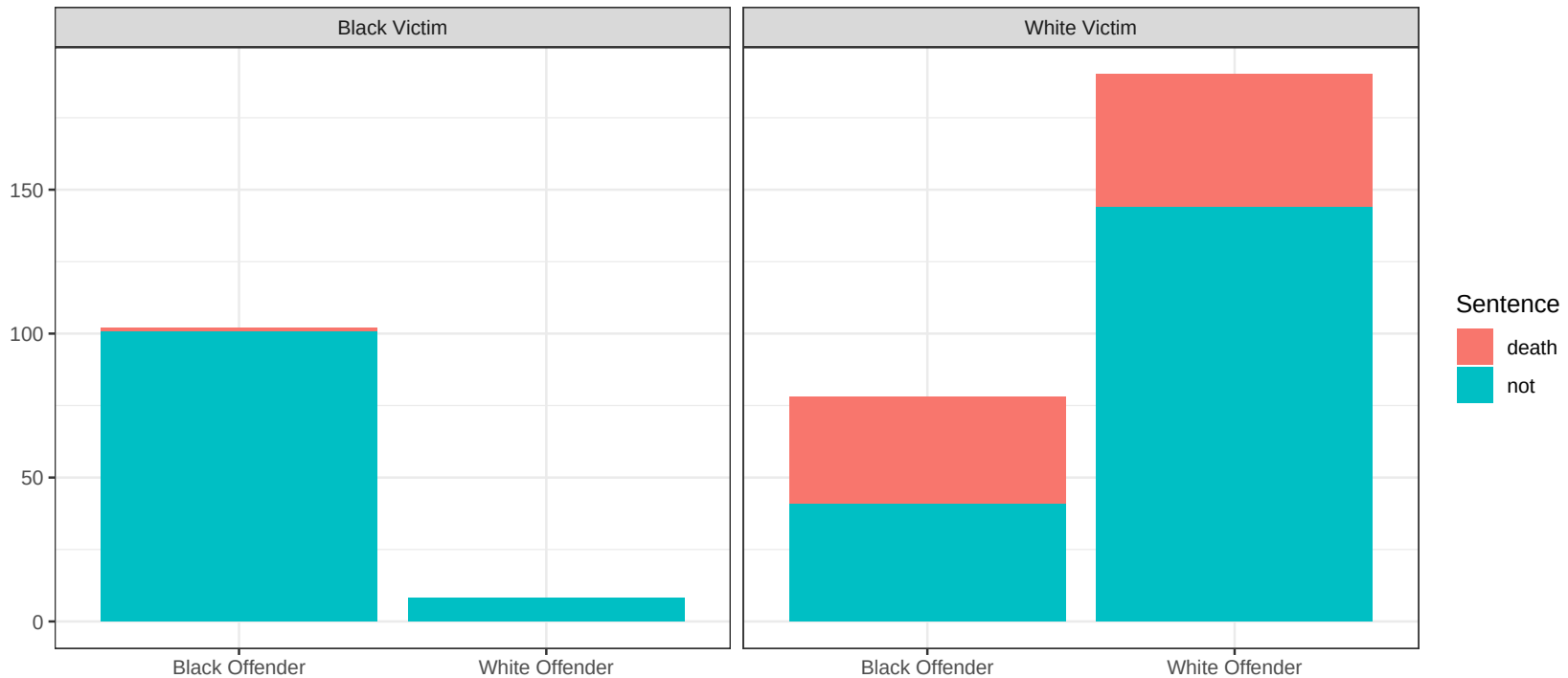
Confounding Variables

Overall, white offenders received the death penalty slightly more often, but this ignored the influence of the victim's race:



Confounding Variables

Because offenders *disproportionately* committed crimes against victims of their own race, the overall death penalty rates were skewed in a way that obscured the racially biased sentencing:



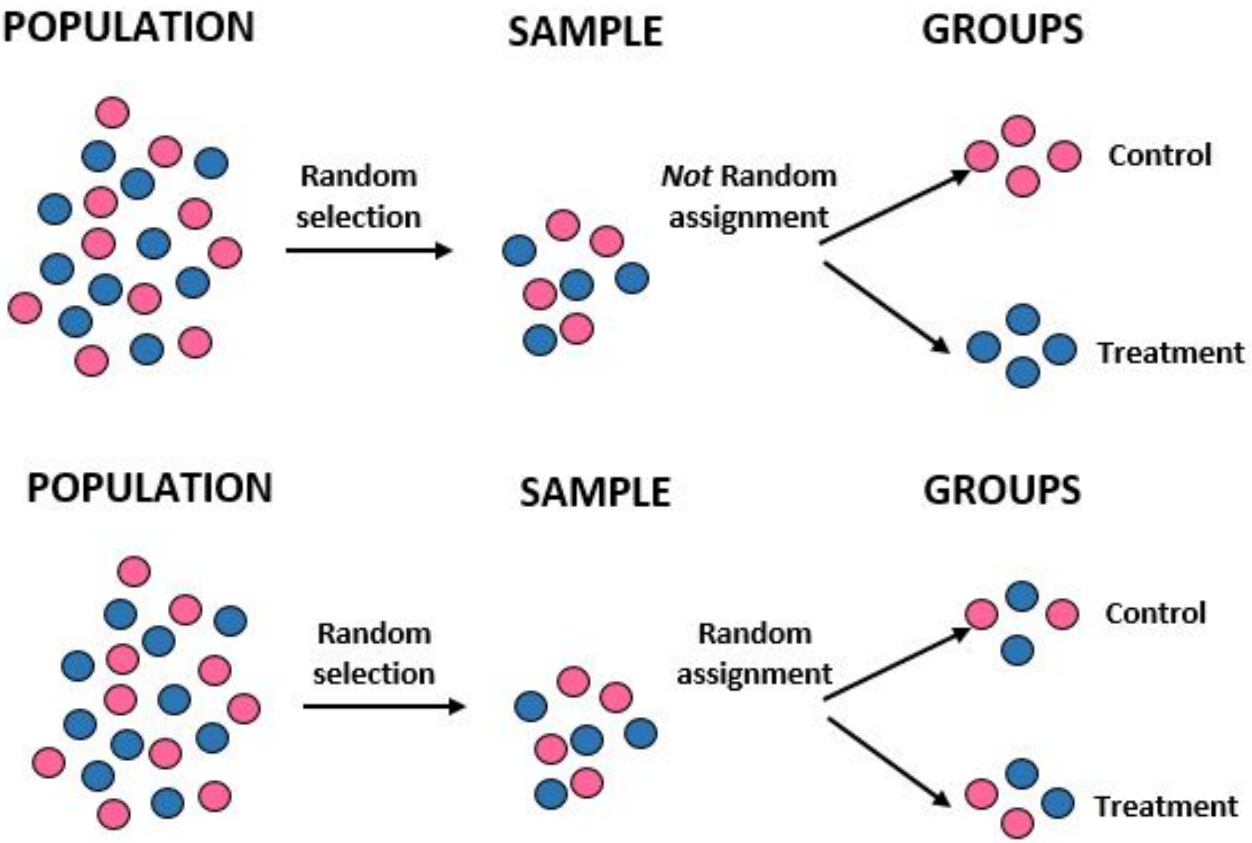
Balance

- We can view the problems caused by confounding variables as an issue of **imbalanced groups**
  - Offenders were more likely to victimize their own race, and crimes against whites tended to be punished more severely
  - The groups white offenders and black offenders were systematically different in an important way (victims race)

Balance

- We can view the problems caused by confounding variables as an issue of **imbalanced groups**
  - Offenders were more likely to victimize their own race, and crimes against whites tended to be punished more severely
  - The groups white offenders and black offenders were systematically different in an important way (victims race)
- Going back to the COVID-19 example, if study participants can choose whether they receive the vaccine, then the vaccine group will likely be disproportionately older, sicker, working riskier jobs, etc.
  - However, these factors would all occur in equal proportions in the vaccinated and control groups if we **randomly assigned** which participants received the vaccine

Random Assignment



Barriers

- Obviously random assignment isn't always feasible, some explanatory variables are too unethical or costly to randomly assign
  - For example, we couldn't assign cases to consume toxic chemicals or expose themselves to harm
  - We also cannot randomly assign explanatory variables that universally pre-date the study like genetics, etc.

Barriers

- Obviously random assignment isn't always feasible, some explanatory variables are too unethical or costly to randomly assign
  - For example, we couldn't assign cases to consume toxic chemicals or expose themselves to harm
  - We also cannot randomly assign explanatory variables that universally pre-date the study like genetics, etc.
- Despite their flaws, observational studies are still very valuable
  - But they will always fall short of *randomized experiments*

Example - Study Design

- Suppose we want to know: "Is arthroscopic surgery is effective in treating arthritis of the knee?" Describe both an *observational study* and a *randomized experiment* that you could conduct to answer this question. Be sure to address the following during your discussion:

1. How costly will it be for the researchers to collect data with each design?
2. Are there any feasibility problems or ethical issues with each design?

Possible Responses

- An observational study might use medical records to find patients with knee arthritis and identify whether they ever received surgery. It then might compare various measures of recovery (recurrent visits, etc.) among those who had or didn't have surgery.
- A randomized experiment would take a sample of patients with arthritis and randomly assign half to have surgery and the other half to not have surgery. It then might compare various measures of recovery.

1. The observational study costs almost nothing, while the randomized experiment likely will cost quite a bit
2. The observational study doesn't present any ethical barriers (assuming sufficient data privacy), but withholding knee surgery seems like it could problematic

Sham Knee Surgery

In the 1990s a study was conducted in 10 men with arthritic knees that were scheduled for surgery. They were all treated identically expect for one key distinction: only half of them actually got surgery! Once each subject was in the operating room and anesthetized, the surgeon looked at a randomly generated code indicating whether he should do the full surgery or just make three small incisions in the knee and stitch up the patient to leave a scar. All patients received the same post-operative care, rehabilitation, and were later evaluated by staff who didn't know whether they had actually received the surgery or not. The result? Both the sham knee surgery and the real knee surgery showed indistinguishable levels of improvement

Source: https://www.nytimes.com/2000/01/09/magazine/the-placebo-prescription.html

Control Groups, Placebos, and Blinding

The Sham Knee Surgery example illustrates several important aspects of a well-designed experiment that we've yet to discuss:

- **Control Group** - Some patients were randomly assigned not to receive the knee surgery, providing a comparison group that is, on average, balanced with surgery group in all baseline characteristics
- **Placebo** - Patients in the control group received a fake surgery
- **Blinding** - Using a placebo is not helpful if patients know the group they're in. Similarly, the staff interacting with the patients might treat them differently if they knew the patient's group
  - **Single-blind** - the participants don't know the treatment assignments
  - **Double-blind** - the participants *and* everyone interacting with the participants don't know the treatment assignments

Control Groups, Placebos, and Blinding

- The goal of each of these design elements is to prevent biasing the measurement of the outcome variable in a particular direction

Control Groups, Placebos, and Blinding

- The goal of each of these design elements is to prevent biasing the measurement of the outcome variable in a particular direction
- Other types of studies are susceptible to other types of biases, which we should also carefully consider

1. Social Desirability Bias - Respondents tend to answer questions in ways that portray themselves in a positive light Link
2. Habituation Bias - Respondents tend to provide similar answers for similarly worded or structured questions (the brain going on autopilot) Link
3. Leading Questions - The wording of a question impacts how people respond, great examples in the textbook
4. Cultural Bias - Questions are often to be constructed with one's own culture in mind, they might not even make sense to people from other cultures.

Discussion - Can Randomization Fail?

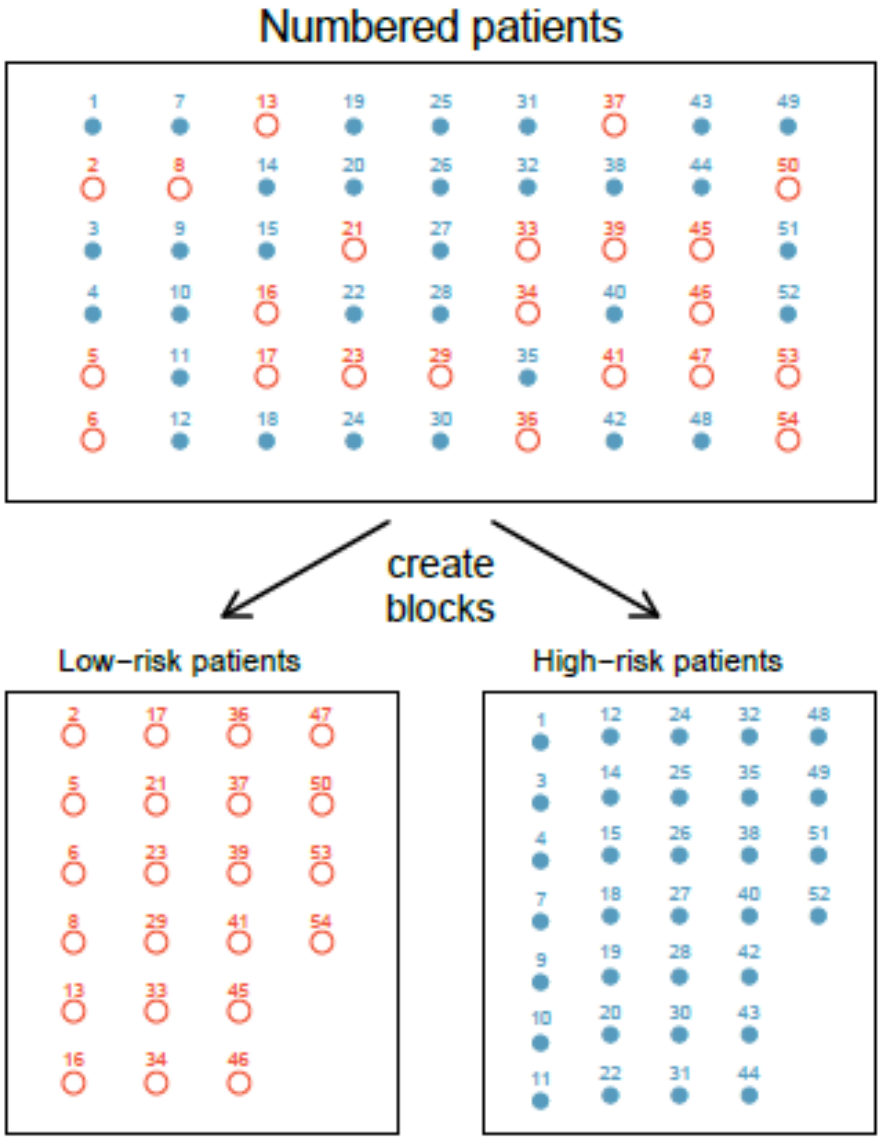
- A University of Iowa researcher was conducting an experiment on lab monkeys
- Lab monkeys are expensive, so his experiment had $n = 8$
- Having taken a statistics course, he randomly assigned treatment/control groups
- After conducting the experiment and seeing surprising results, the researcher recognizes that the 4 monkeys in the control group were also the oldest 4 monkeys
- The researcher knew that the age of the monkey had an important on the outcome variable, but he expected randomization to handle that

Should he report his results? What could he have done differently?

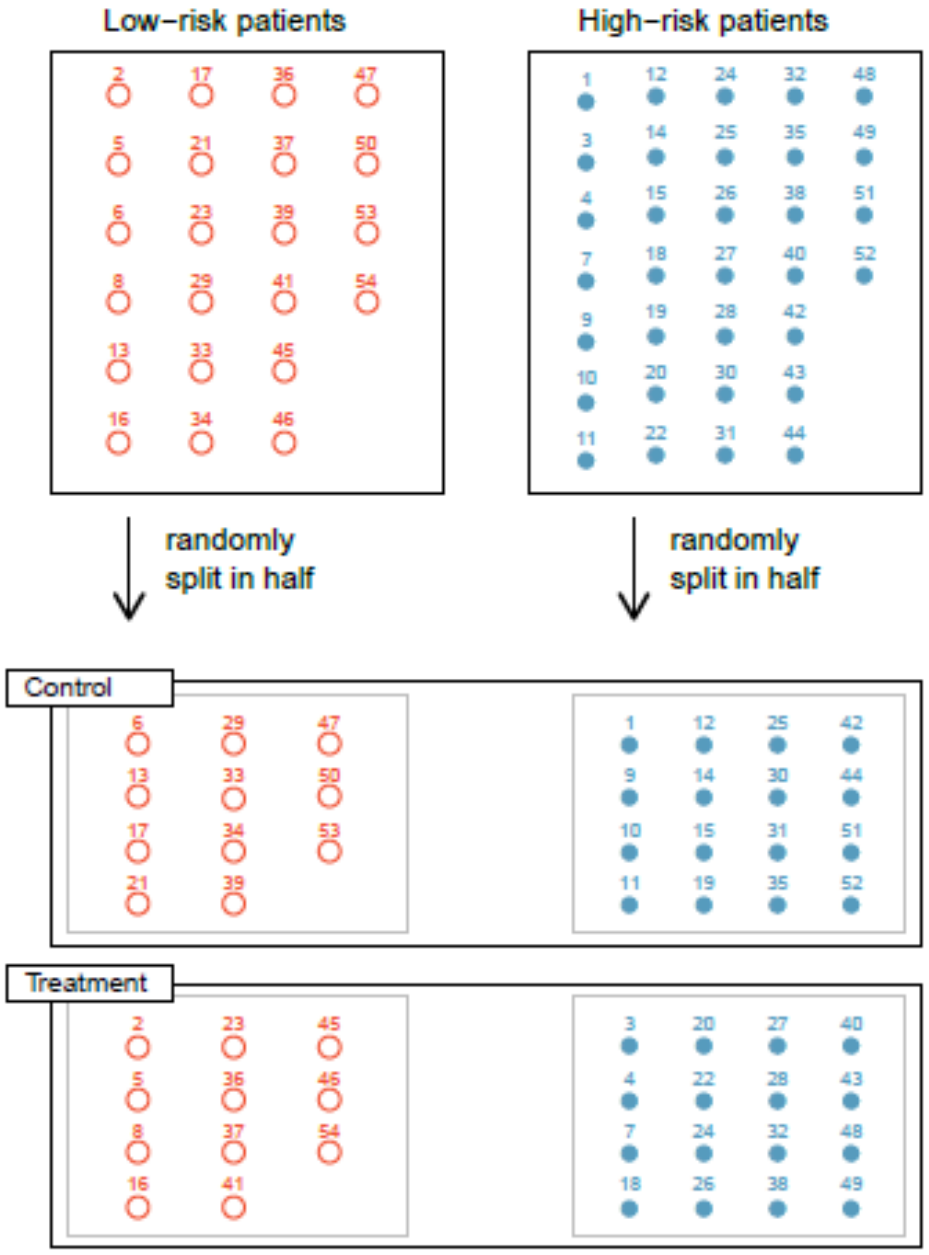
Can Randomization Fail?

- Randomization is not guaranteed to properly balance the treatment and control groups unless the sample size is relatively large
- At smaller sample sizes, strategies such as **blocking** can be used
  - In this design, cases are first split using a **blocking variable**, then random assignment is done within each block
  - This ensures the blocking variable is balanced in each group

Blocking


Numbered patients
1 7 13 19 25 31 37 43 49
2 8 14 20 26 32 38 44 50
3 9 15 21 27 33 39 45 51
4 10 16 22 28 34 40 46 52
5 11 17 23 29 35 41 47 53
6 12 18 24 30 36 42 48 54
create blocks
Low−risk patients
2 17 36 47
5 21 37 50
6 23 39 53
8 29 41 54
13 33 45
16 34 46
High−risk patients
1 12 24 32 48
3 14 25 35 49
4 15 26 38 51
7 18 27 40 52
9 19 28 42
10 20 30 43
11 22 31 44

Blocking


Low−risk patients
2 17 36 47
5 21 37 50
6 23 39 53
8 29 41 54
13 33 45
16 34 46
High−risk patients
1 12 24 32 48
3 14 25 35 49
4 15 26 38 51
7 18 27 40 52
9 19 28 42
10 20 30 43
11 22 31 44
randomly split in half
randomly split in half
Control
6 29 47
13 33 50
17 34 53
21 39
1 12 25 42
9 14 30 44
10 15 31 51
11 19 35 52
Treatment
2 23 45
5 36 46
8 37 54
16 41
3 20 27 40
4 22 28 43
7 24 32 48
18 26 38 49

Study Design and Statistical Inference

- The overarching goal of a statistician is to *rule out* as many possible explanations for an observed association as possible

Study Design and Statistical Inference

- The overarching goal of a statistician is to *rule out* as many possible explanations for an observed association as possible
- So far we've considered the following design-related explanations, as well as methods for addressing for them
  - Sampling bias - Simple random sampling
  - Confounding variables - Random assignment of the explanatory variable, or stratification
  - Other biases - Using placebo, double-blinding, proper instruments, etc.

Study Design and Statistical Inference

- As we discussed in Week 1 (babies choosing the helper or hinderer toys), when a study is well-designed the only viable explanations for trends that are observed are:
  - Random chance
  - The association is real

Study Design and Statistical Inference

- As we discussed in Week 1 (babies choosing the helper or hinderer toys), when a study is well-designed the only viable explanations for trends that are observed are:
  - Random chance
  - The association is real
- We'll soon learn how statisticians use *probability theory* to decide between these explanations