Sampling and Bias

After completing this section, students should be able to:

- List potential sources of bias in a survey or opinion poll.
- Evaluate different methods of picking a sample and specify which are more likely to get a sample that is representative of the population.
Before class

• Read the Pew Research article: “Why 2016 election polls missed their mark”

• Answer the questions:

1. What is meant by non-response bias on a survey?
   (a) Non-response bias happens when do not answer the survey questions honestly.
   (b) Non-response bias happens when the kinds of people who answer the survey are different from the kinds of people who do not answer it.
   (c) Non-response bias occurs when the pollsters are more likely to survey certain kinds of people than others.

2. According to the Pew Research article “Why 2016 election polls missed their mark”, which of the following are possible reasons that the polls might have been wrong? Select all that apply.
   (a) The pollsters didn’t ask enough people.
   (b) Non-response bias.
   (c) Respondents may not have answered honestly.
   (d) Mis-calculations of who was likely to vote.
Introduction

In 2016, many polls predicted that Clinton would win and were wrong! We are now starting another election cycle. Can we trust the polls?
Samples

**Definition.** A *population* is ...

**Definition.** A *sample* is ...

**Definition.** *Statistical inference* is ...
What is the population and what is the sample?
What is the population for the YouGov / Econ poll in the top line? (PollEv)
A. Everyone in the US
B. All US citizens
C. All US likely voters
D. 3677 US likely voters

What is the sample?

Note that the actual vote turned out to be 48% Clinton, 46% Trump, 3% Johnson, 1%
Stein, 2% Others.
What are some reasons why different polls might give different answers? (PollEv)
Variability in Samples

Before we look at real polls (where we don’t have information about the entire population), we are going to look at an example of taking samples of words from speeches (where we do have complete information about the population). This may help us experience some ways that polls can go wrong.
Who do you think uses longer words in their speeches? Obama or Trump? What about in their inaugural speeches? (PollEv)
REMIDS OF PRESIDENT DONALD J. TRUMP – AS PREPARED FOR DELIVERY
INAUGURAL ADDRESS
FRIDAY, JANUARY 20, 2017
WASHINGTON, D.C.

As Prepared for Delivery –

Chief Justice Roberts, President Carter, President Clinton, President Bush, President Obama, fellow Americans, and people of the world: thank you.

We, the citizens of America, are now joined in a great national effort to rebuild our country and to restore its promise for all of our people.

Together, we will determine the course of America and the world for years to come.

We will face challenges. We will confront hardships. But we will get the job done.

Every four years, we gather on these steps to carry out the orderly and peaceful transfer of power, and we are grateful to President Obama and First Lady Michelle Obama for their gracious aid throughout this transition. They have been magnificent.

Transcript: Inaugural Address of Barack Obama

Jan. 20, 2009

SPEAKER: PRESIDENT BARACK OBAMA

[*] OBAMA: Thank you. Thank you.

CROWD: Obama! Obama! Obama! Obama!

My fellow citizens; I stand here today humbled by the task before us, grateful for the trust you have placed in me. We are mindful of the sacrifices borne by our ancestors.

I thank President Bush for his service to our nation...

(APPLAUSE)

... as well as the generosity and cooperation he has shown throughout this transition.

Forty-four Americans have now taken the presidential oath.

OBAMA: The words have been spoken during rising tides of prosperity and the still waters of peace. So often the oath is taken amidst gathering clouds and raging storms. At these moments, America has risen to the weight of our fundamentals: our idealism, our self- reliance; true to our founding documents.

OBAMA: So it has been. So it must be with this generation of Americans.

That we are in the midst of crisis is now well understood. Our nation is at war against a far- reaching struggle for power,霸权ism and hatred. Our economy is badly weakened, a consequence of greed and irresponsibility on the part of some but also our collective failure to make hard choices and prepare the nation for a new age.

Homes have been lost, jobs shed, businesses shuttered. Our health care is too costly, our schools and each day brings further evidence that the ways we use energy strengthen our adversaries planet.
Task 1:

- Select a sample of 10 words that are representative of Trump’s overall inaugural speech. (See transcripts and a numbered list of words in https://tinyurl.com/uncMath115StudentSpreadsheets.)
- Calculate the average number of letters for the words in your sample.
- Repeat for Obama’s 2009 inaugural speech. You’ll need to change the range for the random number generator to match the length of Obama’s speech.
- Enter your averages in GoogleForms - use the FIRST form.

What is the population for this activity? What is the sample?

What method did you use to pick your samples?
Random Samples

Task 2:

- Google a random number generator and pick 10 random numbers between 1 and the total number of words in Trump’s speech (if you get a duplicate, generate another number). Write these numbers down. Note: there are 1455 words total.

- Calculate the average number of letters in the words corresponding to your random numbers, using the numbered list of words from Trump’s 2017 inaugural speech (see link on Sakai > Lessons).

- Repeat for Obama’s 2009 inaugural speech. Note: there are 2393 words total.

- Enter your averages in GoogleForms - use the SECOND form.
Who used longer words in their inaugural speeches? (PollEv)
A. Trump
B. Obama
C. It’s a tie.

What is the point of this exercise?
Random vs. Non-random Samples

- Random samples have averages that are centered around the correct number
- Non-random samples may suffer from sampling bias, and averages may not be centered around the correct number
- Only random samples can truly be trusted when making generalizations to the population!
Soup Analogy

Think of tasting a bowl of soup . . .

Population =

Sample =

If you take bites non-randomly from the soup (if you stab with a fork, or prefer noodles to vegetables), you may not get a very accurate representation of the soup.

If you take bites at random, only a few bites can give you a very good idea for the overall taste of the soup.
Forms of bias

Even with a random sample, data can still be biased, especially when collected on humans

Other forms of bias to watch out for in data collection:

- Question wording
- Context
- Inaccurate responses
- Non-response bias
- Many other possibilities!
Question Wording

“Do you think the US should allow public speeches against democracy?”

21% said speeches should be allowed

“Do you think the US should not forbid public speeches against democracy?”

39% said speeches should be not be forbidden

Context
Ann Landers column asked readers

“If you had it to do over again, would you have children?

The request for data contained a letter from a young couple which listed worries about parenting and various reasons not to have kids

30% said “yes”

Newsday conducted a random sample of all US adults, and asked them the same question, without any additional leading material

91% said “yes”
Inaccurate Responses

In a study on US students, 93% of the sample said they were in the top half of the sample regarding driving skill


From random sample of all US college students, 22.7% reported using illicit drugs. Do you think this number is accurate?

Non-response bias

Why might a survey conducted today in China by random digit dialing of cell phones underestimate the number of people sick with coronavirus?
What forms of bias do you think could have impacted the 2016 presidential election polls?

### Four-way race

<table>
<thead>
<tr>
<th>Poll source</th>
<th>Date</th>
<th>Hillary Clinton Democratic</th>
<th>Donald Trump Republican</th>
<th>Gary Johnson Libertarian</th>
<th>Jill Stein Green</th>
<th>Leading by (points)</th>
<th>Sample size</th>
<th>Margin of error</th>
</tr>
</thead>
<tbody>
<tr>
<td>YouGov/The Economist</td>
<td>November 4–7, 2016</td>
<td>45%</td>
<td>41%</td>
<td>5%</td>
<td>2%</td>
<td>4</td>
<td>3,677</td>
<td>± 1.7%</td>
</tr>
<tr>
<td>Insights West</td>
<td>November 4–7, 2016</td>
<td>49%</td>
<td>45%</td>
<td>4%</td>
<td>1%</td>
<td>4</td>
<td>940</td>
<td>± 3.2%</td>
</tr>
<tr>
<td>Bloomberg News/Seelzer</td>
<td>November 4–6, 2016</td>
<td>44%</td>
<td>41%</td>
<td>4%</td>
<td>2%</td>
<td>3</td>
<td>799</td>
<td>± 3.5%</td>
</tr>
<tr>
<td>Gravis Marketing</td>
<td>November 3–6, 2016</td>
<td>47%</td>
<td>43%</td>
<td>3%</td>
<td>2%</td>
<td>4</td>
<td>16,639</td>
<td>± 0.8%</td>
</tr>
<tr>
<td>ABC News/Washington Post</td>
<td>November 3–6, 2016</td>
<td>47%</td>
<td>43%</td>
<td>4%</td>
<td>1%</td>
<td>4</td>
<td>2,220</td>
<td>± 2.5%</td>
</tr>
<tr>
<td>Fox News</td>
<td>November 3–6, 2016</td>
<td>48%</td>
<td>44%</td>
<td>3%</td>
<td>2%</td>
<td>4</td>
<td>1,295</td>
<td>± 2.5%</td>
</tr>
<tr>
<td>IBD/TIPP</td>
<td>November 3–6, 2016</td>
<td>41%</td>
<td>43%</td>
<td>6%</td>
<td>2%</td>
<td>2</td>
<td>1,026</td>
<td>± 3.1%</td>
</tr>
<tr>
<td>Monmouth University</td>
<td>November 3–6, 2016</td>
<td>50%</td>
<td>44%</td>
<td>4%</td>
<td>1%</td>
<td>6</td>
<td>802</td>
<td>± 3.6%</td>
</tr>
<tr>
<td>Ipsos/Reuters</td>
<td>November 2–6, 2016</td>
<td>42%</td>
<td>39%</td>
<td>8%</td>
<td>3%</td>
<td>3</td>
<td>2,195</td>
<td>± 2.4%</td>
</tr>
<tr>
<td>CBS News/New York Times</td>
<td>November 2–6, 2016</td>
<td>45%</td>
<td>41%</td>
<td>5%</td>
<td>2%</td>
<td>4</td>
<td>1,426</td>
<td>± 3.0%</td>
</tr>
<tr>
<td>Rasmussen Reports</td>
<td>November 2–6, 2016</td>
<td>45%</td>
<td>43%</td>
<td>4%</td>
<td>2%</td>
<td>2</td>
<td>1,500</td>
<td>± 2.5%</td>
</tr>
</tbody>
</table>
Realities of Sampling

While a random sample is ideal, often it isn’t feasible. A list of the entire population may not be available, or it may be impossible or too difficult to contact all members of the population.

Sometimes, your population of interest has to be altered to something more feasible to sample from. Generalization of results are limited to the population that was actually sampled from.

In practice, think hard about potential sources of sampling bias, and try your best to avoid them.
Convenience Sample

Individuals who are easily accessible are more likely to be included in the sample.
Simple Random Sample

Each case in the population has an equal chance of being included
Systematic Sampling

Select every kth element of an ordered population


Stratified Sampling

Population is divided into groups called strata. The strata are chosen so that similar cases are grouped together, then a second sampling method, usually simple random sampling, is employed within each stratum.
Cluster Sample

Population is broken into many groups, called clusters. Then we sample a fixed number of clusters and sample the entire cluster.
Multistage Cluster Sample

Population is broken into many groups, called clusters. Then we sample a fixed number of clusters and take a simple random sample the selected clusters.
1. (4 pts) A 2010 poll from the Pew Research Center found that 40% of American adults say that they have tried marijuana. Give two reasons why that estimate might be off.

2. (6 pts) Which of these methods of choosing 400 students for a sample is most likely to reflect the population of all enrolled UNC students? For each method that you do NOT think is best, explain why this method might produce bias.

(a) Stand by the Pit and select 400 students as they come by, using a quota system – make sure the percentage of women, men, students of color, and first generation college students match the percentages of the university.

(b) Sort a list of all enrolled students by name, pick a number at random between 1 and the number of students, and select the next 400 students on the list starting with that position number.

(c) Sort a list of all enrolled students by name, first name, use a random number generator to pick 400 numbers between 1 and the number of students, and select the students whose number on the list equals one of the numbers you picked.