Chapter 7: Survey Sampling and Inference
Survey Terminology

• The **Population** is the group of people or objects we wish to study.

• A **Parameter** is a numerical value that characterizes some aspect of the population.

• A **Census** is a survey in which every member of the population is measured.

• A **Sample** is a collection of people or objects taken from the population.

• A **Statistic** is a number that estimates a population parameter derived from the sample.
Population vs. Sample

- **Population** is the collection of ALL data values.
- **Population size** is usually very large, often unknown, and usually impossible to obtain all values.
- Measures that come from the population are called **parameters**.

- **Sample** is a subset of the population and we can measure characteristics.
- **Sample size** is the number of observations in a sample, $n$.
- Measures that come from the sample are called **statistics**.

$n = 7$
Notation

- Typically, we use Greek letters to represent population parameters and Latin letters to represent sample statistics.

<table>
<thead>
<tr>
<th></th>
<th>Population</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>$\mu$</td>
<td>$\bar{x}$</td>
</tr>
<tr>
<td>SD</td>
<td>$\sigma$</td>
<td>$s$</td>
</tr>
</tbody>
</table>
Example

- Are the values below parameters or statistics?

(a) parameters          (b) statistics

*Since Barack Obama took office, do you think the Republican Party has become more conservative, or not?*

*(Asked of those who think the Republican Party has gotten more conservative) Do you think the Republican Party has become more conservative mostly because of the influence of the Tea Party movement or mostly for other reasons?*

<table>
<thead>
<tr>
<th></th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes, more conservative</td>
<td>54</td>
</tr>
<tr>
<td>(Mostly because of Tea Party influence)</td>
<td>(19)</td>
</tr>
<tr>
<td>(Mostly for other reasons)</td>
<td>(35)</td>
</tr>
<tr>
<td>No, not more conservative</td>
<td>40</td>
</tr>
<tr>
<td>No opinion</td>
<td>6</td>
</tr>
</tbody>
</table>
Census

- Wouldn't it be better to just include everyone and sample the entire population?
  - Such a special sample is called a **census**.

- There are problems with taking a census:
  - It can be difficult to complete a census: there always seem to be some individuals who are hard to locate or hard to measure. And there may be a certain characteristic about those individuals who are hard to locate.
  - Populations rarely stand still. Even if you could take a census, the population changes while you work, so its never possible to get a perfect measure.
  - Taking a census may be more complex than sampling.
Sampling is natural

- Sampling is a natural thing to do.
  - Think about sampling something you are cooking - you taste (examine) a small part of what you're cooking to get an idea about the dish as a whole.
  - If you walk into a clothing store that you've never heard of before, in order to decide if the store is affordable you wouldn't check the tag of every single item in the store. You would instead try to check out the price of a variety of items (a representative sample) and based on what you see decide if you think the store overall is overpriced or not.
Exploratory Analysis to Inference

• When you taste a spoonful of soup and decide it doesn't taste salty enough, that's exploratory analysis

• If you generalize and conclude that your soup needs salt, that's an inference

• For your inference to be valid the spoonful you tasted (the sample) needs to be representative of the entire pot (the population)
  • If your spoonful comes only from the surface and the salt is collected at the bottom of the pot, what you tasted is probably not representative of the whole pot.
  • If you first stir the soup thoroughly before you taste, your spoonful will more likely be representative of the whole pot.
Statistical Inference

- **Statistical inference** is the art and science of drawing conclusions about a population on the basis of observing only a small subset of that population (i.e. a sample).

- **Statistical inference** always involves uncertainty, so an important component of this science is measuring our uncertainty.
Sample Surveys

- Opinion polls are examples of sample surveys designed to ask questions of a small group of people in the hope of learning something about the entire population.
  - Professional pollsters try to ensure that the sample they take is representative of the population.
  - If not, the sample can give misleading information about the population.
Landon vs. FDR

- In 1936, Landon sought the Republican presidential nomination opposing the re-election of FDR.
The Literary Digest Poll

- The Literary Digest polled about 10 million Americans, and got responses from about 2.4 million.
- The poll showed that Landon would likely be the overwhelming winner and FDR would get only 43% of the votes.
- Election result: FDR won, with 62% of the votes.
  - The magazine was completely discredited because of the poll and was soon discontinued.
The Literary Digest Poll - what went wrong?

- The magazine had surveyed
  - Its own readers
  - Registered automobile owners, and
  - Registered telephone users
- These groups had incomes well above the national average of the day (remember, this is Great Depression era) which resulted in lists of voters far more likely to support Republicans than a truly *typical* voter of the time, i.e. the sample was not representative of the American population at the time.
Bias

• A method is **biased** if it has a tendency to produce an untrue value.

• **Sampling bias** results from taking a sample that is not representative of the population.
  • Convenience sampling and voluntary response sampling

• **Measurement bias** comes from asking questions that do not produce a true answer.
  • Confusing wording, misleading questions
Voluntary Samples are problematic

• In a voluntary response sample a large group of individuals are invited to respond, and all who do respond are counted.

  • Voluntary response samples are almost always biased, and so conclusions drawn from them are almost always wrong.

• Voluntary response samples are often biased toward those with strong opinions or those who are strongly motivated.

• Since the sample is not representative, the resulting voluntary response bias invalidates the survey.
Convenient samples are not so convenient...

- In **convenience sampling** we simply include the individuals who are convenient.
  - Unfortunately, this group may not be representative of the population.
- For example, if you are strongly against smoking chances are none of your close friends are smokers, so your friends wouldn't make a good sample to ask about anti-smoking laws.
- Convenience sampling is not only a problem for students or other beginning samplers.
  - In fact, it is a widespread problem in the business world - the easiest people for a company to sample are its own customers.
Non-response

• A common and serious potential source of bias for most surveys is nonresponse bias.

• No survey succeeds in getting responses from everyone.
  • The problem is that those who don't respond may differ from those who do.
  • And they may differ on just the variables we care about.

• **Non-response error** occurs when those who respond may differ from whose who do not

• For example, surveys sent home with students may show that parents have no trouble sparing time to spend with their children. But which parents return the surveys?
Design of Survey Questions

- Surveys that are too long are more likely to be refused, reducing the response rate and biasing all the results.

- Work hard to avoid influencing responses.
  - Response bias refers to anything in the survey design that influences the responses.
  - For example, the wording of a question can influence the responses.

- Measurement bias occurs when anything in the survey design influences the responses (e.g. wording of survey questions).


Measurement Bias

THE WIZARD OF ID

HOW WOULD YOU RATE THE KING'S JOB PERFORMANCE?

WOULDN'T YOU SAY HE'S DOING A GOOD JOB, AN OUTSTANDING JOB OR A TERRIFIC JOB?

I THINK I KNOW WHO IS PAYING FOR THIS POLL

OPINION POLL

2-23

© 1981 CREATORS SYNDICATE, INC. WWW.CREATORS.COM
Questions to ask to when thinking about bias

- What percentage of people who were asked to participate actually did so?

- Did the researchers choose people to participate in the survey or did the people themselves choose to participate?

- Did the researcher leave out whole segments of the population who are likely to answer the question differently from the rest of the population?
Identify possible biases

• A student asked all 2500 of her Facebook friends if they preferred Facebook to Twitter.

• A researcher asked 500 randomly selected people, “Are you in favor of the unfair tax burden that the hard working successful business people have so that the lazy unemployed can receive a paycheck without working?”

• On July 4, CNN posted on their website a question asking if they supported the current US military operations. 18,943 people responded.

• A researcher stood outside a grocery store and asked 250 shoppers, “Do you eat out at a restaurant at least three times per week?”
Simple Random Sampling

- **Simple Random Sampling**, SRS, involves randomly drawing people from the population without replacement.

- If a scientific sampling technique is not done, we cannot learn anything about the population by looking at the sample data.
SRS continued...

• We want to make sure that every possible sample of the size we plan to draw has an equal chance of being selected.

• Such samples also guarantee that each individual has an equal chance of being selected.

• With this method each combination of people has an equal chance of being selected as well.

• This ensures that sample-to-sample differences, also called precision, is small.

• **Precision**: Sample-to-sample differences caused by the fact that each draw of random numbers selects different people for the sample.
Accuracy and Precision, Bias and Standard Error

• Bias is a measure of the accuracy.
  
  • If only basketball players are measured to estimated the proportion of Americans who are taller than 6 feet, then there is a bias for a larger proportion.

• Standard Error is a measure of precision.
  
  • If the sample size is only three, the estimate of the proportion of tall people using the sample is likely to be far from the proportion of tall people in the US. The standard error will be large.
Bias, Precision, Mean, and Standard Error

• For a SRS, the bias is 0.
  • Equivalent to the statement that the mean of all the sample proportions equals the population proportion.

• For a SRS, the precision is better for larger sample sizes.
  • Equivalent to the statement that the standard error is smaller for larger sample sizes.

• The precision and bias are independent of the population size as long as the population size is as least 10 times larger than the sample size.
Sampling Distributions

- We take random samples of populations to make some inference about a population parameter.

- There is some inherent variability of using a sample as an estimate of the population parameter.
  - If each one of us took a random sample of people's preference of Obama vs. Romney on UCLA campus we would all get different sample proportions, $\hat{p}$, in favor of Obama.

- This variability can be expressed through a sampling distribution, which provides us with both a mean and standard deviation to describe $\hat{p}$. 
Reese's Pieces Simulation

The Central Limit Theorem for Sample Proportions

- **The Central Limit Theorem for Sample Proportions:** If the trials are random and independent and the sample and population sizes are large then the sampling distribution of $\hat{p}$ is approximately normal and follows:

$$CLT: \hat{p} \sim N \left( mean(\hat{p}) = p, SD(\hat{p}) = \sqrt{\frac{pq}{n}} \right)$$
Conditions for the Central Limit Theorem for Sample Proportions

- **Random and Independent:** The sample is collected randomly and the trials are independent of each other.

- **Large Sample:**
  - The sample has at least 10 successes, \( np \geq 10 \), and
  - at least 10 failures \( n(1 - p) \geq 10 \).

- **Large Population:** If the sample is collected without replacement, then the population size is at least 10 times the sample size.
The Central Limit Theorem Example

- **200** randomly selected American drivers were asked if they text while driving. **24%** of all American drivers admit to texting while driving.

- Check your conditions:
  - The drivers were randomly selected.
  - Successes: $200 \times (0.24) = 48 \geq 10$,
    
    Failures: $200 \times (1 - 0.24) = 152 \geq 10$
  - Population Size (# American Drivers) is very large.

- **Conclusion**: The distribution is approximately normal based on the CLT.
The Central Limit Theorem Example

- 200 randomly selected American drivers were asked if they text while driving. 24% of all American drivers admit to texting while driving.

- The distribution of a sample proportion, $\hat{p}$, is distributed normally as such:

$$N(\mu = 0.24, SD(\hat{p}) = \sqrt{\frac{0.24 \times 0.76}{200}})$$
Finding Probabilities with the Central Limit Theorem

- 78% of all laboratory mice can make it through a maze. If 600 randomly selected mice attempt the maze, what is the probability that more than 80% of them will make it through the maze?

- Note that all requirements are met.

  - Random Sample
  - \# Successes \((np) = 600 \times 0.78 = 468 \geq 10\)
  - \# Failures \((n(1-p)) = 600 \times 0.22 = 132 \geq 10\)
  - Large population size: All mice in existence.
Finding Probabilities with the Central Limit Theorem

- 78% of all laboratory mice can make it through a maze. If 600 randomly selected mice attempt the maze, what is the probability that more than 80% of them will make it through the maze?

- By CLT the distribution for all possible sample proportions, the sampling distribution, is approximately Normal.

\[ N(\mu = 0.78, SD(\hat{p}) = \sqrt{\frac{0.80 \times 0.20}{600}}) \]
Finding Probabilities with the Central Limit Theorem

- An iPod has 5,000 songs. Assume that 15% of the songs in this iPod are classical. We randomly pick 100 songs for a playlist. What is the probability that no more than 10% of the songs will be classical?
Finding Probabilities with the Central Limit Theorem

• 200 randomly selected American drivers were asked if they text while driving. 24% of all American drivers admit to texting while driving. What is the probability that 52 or more of the randomly selected drivers text while driving?