# Chapter 7

## Survey Sampling and Inference

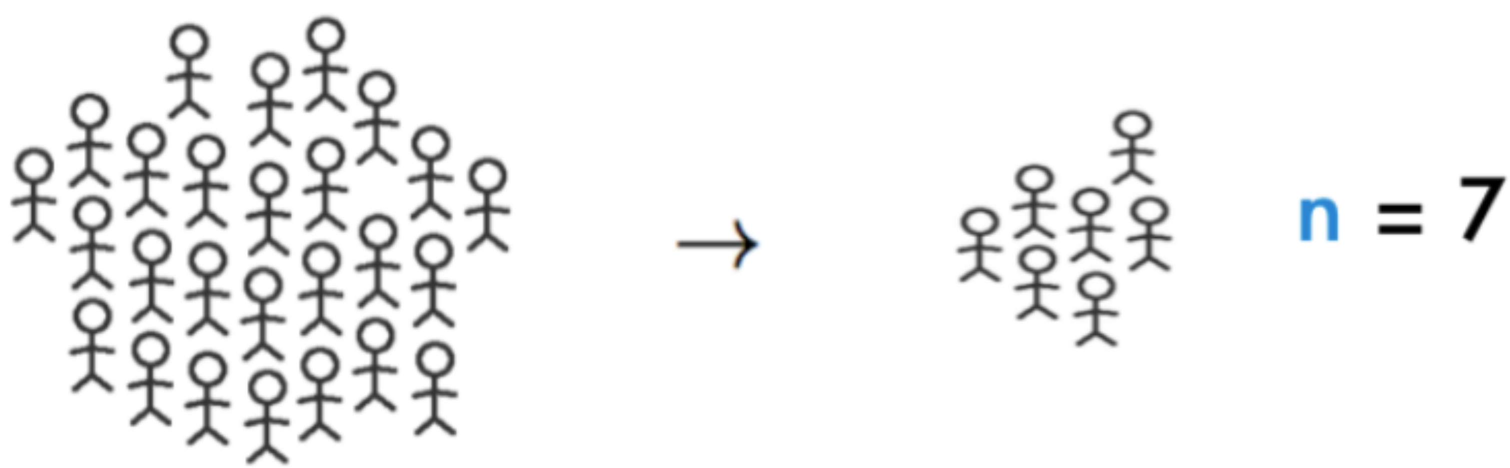
# Survey Terminology

- 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.
- The sample is a collection of people or objects taken from the population.
- A statistics is a number that estimates a population parameter derived from the sample. Sometimes called estimators, and the numbers that result are called estimates.

# 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 parameters.

- Sample is a subset of the population.
- Sample size (n) is the number of observations in a sample.
- Measures that come from the sample are statistics.



# Notation

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

| | Population | Sample |
|---|---|---|
| **Mean** | $\mu$ | $\bar{x}$ |
| **Standard Devation** | $\sigma$ | $s$ |

# Example

Clicker!

In February 2014, the Pew Research Center surveyed 1428 adults in the United States who were married or in a committed partnership. The survey found that 25% of cell phone owners felt that their spouse or partner was distracted by her or his cell phone when they were together.

**What is the population?**

Answer:

A. 1428 adults in the US who were surveyed

B. All American adults who were married or in a committed partnership and owned a cell phone

C. Percentage of adults in the US who were married or in a committed partnership and felt that their spouse or partner was distracted by her or his cell phone when they were together

D. the 25% of cell phone owners who felt that their spouse or partner was distracted by her or his cell phone when they were together

# Example

Clicker!

In February 2014, the Pew Research Center surveyed 1428 adults in the United States who were married or in a committed partnership. The survey found that 25% of cell phone owners felt that their spouse or partner was distracted by her or his cell phone when they were together.

**What is the sample?**

Answer:

A. 1428 adults in the US who were surveyed

B. All American adults who were married or in a committed partnership and owned a cell phone

C. Percentage of adults in the US who were married or in a committed partnership and felt that their spouse or partner was distracted by her or his cell phone when they were together

D. the 25% of cell phone owners who felt that their spouse or partner was distracted by her or his cell phone when they were together

# Example

In February 2014, the Pew Research Center surveyed 1428 adults in the United States who were married or in a committed partnership. The survey found that 25% of cell phone owners felt that their spouse or partner was distracted by her or his cell phone when they were together.

Identify the following.

Population: All American adults who were married or in a committed partnership and owned a cell phone.

Sample: 1428 adults who were surveyed.

# Example

Clicker!

In February 2014, the Pew Research Center surveyed 1428 adults in the United States who were married or in a committed partnership. The survey found that 25% of cell phone owners felt that their spouse or partner was distracted by her or his cell phone when they were together.

**What is the parameter?**

Answer:

A. 1428 adults in the US who were surveyed

B. All American adults who were married or in a committed partnership and owned a cell phone

C. Percentage of adults in the US who were married or in a committed partnership and felt that their spouse or partner was distracted by her or his cell phone when they were together

D. the 25% of cell phone owners who felt that their spouse or partner was distracted by her or his cell phone when they were together

# Example

Clicker!

In February 2014, the Pew Research Center surveyed 1428 adults in the United States who were married or in a committed partnership. The survey found that 25% of cell phone owners felt that their spouse or partner was distracted by her or his cell phone when they were together.

**What is the statistic?**

Answer:

A. 1428 adults in the US who were surveyed
B. All American adults who were married or in a committed partnership and owned a cell phone
C. Percentage of adults in the US who were married or in a committed partnership and felt that their spouse or partner was distracted by her or his cell phone when they were together
D. the 25% of cell phone owners who felt that their spouse or partner was distracted by her or his cell phone when they were together

# Example

In February 2014, the Pew Research Center surveyed 1428 adults in the United States who were married or in a committed partnership. The survey found that 25% of cell phone owners felt that their spouse or partner was distracted by her or his cell phone when they were together.

Identify the following.

Parameter: Percentage of adults in the US who were married or in a committed partnership and felt that their spouse or partner was distracted by her or his cell phone when they were together.

Statistic: The percentage of the sample who felt this way, 25%.

# Census

- Wouldn’t it be better to just include everyone and sample the entire population?
    - This is also known as a census.
- There are problems with taking a census:
    - It costs a lot of time and money.
    - Populations rarely stay constant, so it’s not even possible all the time to take a completely accurate measure.
    - It may be more complex that taking a sample.

# Sampling

- Sampling is natural. It's something we already 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 every tag of every single item in the store. Instead, you would try to check out the price of a variety of items (a representative sample) and based on what you see you would decide if you think the store overall is overpriced or not.
- Other examples: shopping for a car or buying airline tickets.

# Exploratory Analysis to Inference

- While cooking when you taste a spoonful and decide it doesn't taste salty enough, that's exploratory analysis.
- If you generalize and conclude that your soup needs salts, that's 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 presidential nomination, as a Republican, opposing the re-election of FDR.

# The Literary Digest Poll

- The Literary Digest, a popular news magazine, 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.
    - Registered telephone users.
- These groups had incomes well above the national average of the day (it was the 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
    - 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 those 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.

# Convenience 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.

# Nonresponse Bias

- 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 those who do not respond.
- 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, people tend to respond to surveys only if they have strong feelings about the results.
- Measurement bias occurs when asking questions that do not produce a true answer.
    - For example, if we ask people their income, they are likely to inflate the value. In this case we would get a positive (or upward) bias.

# Questions to Ask 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

Clicker!

- A student asked all 250 of her Facebook friends if they preferred Facebook to Instagram.

- Is there bias in this scenario?

  A. Yes

  B. No

# Identify Possible Biases

- A student asked all 250 of her Facebook friends if they preferred Facebook to Instagram.

- Is there bias in this scenario?

  A. Yes

  B. No

# Identify Possible Biases

Clicker!

- 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?"

- Is there bias in this scenario?

  A. Yes

  B. No

# Identify Possible Biases

- 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?"

- Is there bias in this scenario?

  A. Yes

  B. No

# Identify Possible Biases

Clicker!

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

- Is there bias in this scenario?

  A. Yes

  B. No

# Identify Possible Biases

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

- Is there bias in this scenario?

  A. Yes

  B. No

# Identify Possible Biases

Clicker!

- A researcher stood outside a restaurant and asked 100 customers, “Do you eat out at a restaurant at least three times per week?”

- Is there bias in this scenario?

  A. Yes

  B. No

# Identify Possible Biases

- A researcher stood outside a restaurant and asked 100 customers, “Do you eat out at a restaurant at least three times per week?”

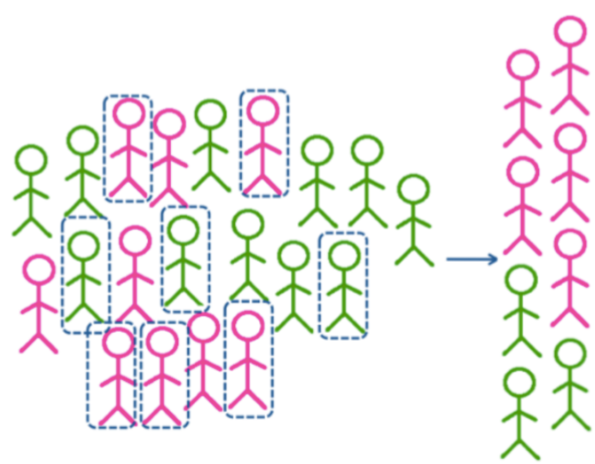
- Is there bias in this scenario?

  A. Yes

  B. No

# Simple Random Sampling

- If a scientific sampling technique is not done, we cannot learn anything about the population by looking at the sample data.
- Simple random sampling, SRS, involves randomly drawing people from the population without replacement.
- Without replacement means that once a subject is selected for a sample, that subject can't be selected again.

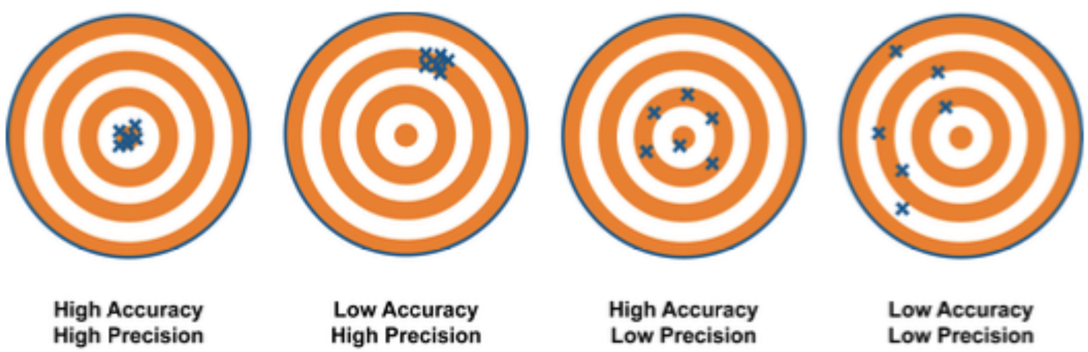


# Simple Random Sampling

- 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 or sampling variability 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

- The accuracy is measured in terms of the bias (taking a good sample).
    - If only basketball players are measured to estimate the proportion of Americans who are taller than 6 feet, then there is a bias for a larger proportion.
- The precision is measured by a number called the standard error (sample size).
    - 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 sample size is small. The estimation method is not very precise.



# Bias and Precision

- For a simple random sample (SRS), the bias is 0, since each individual has an equal chance of being selected.
- For a SRS, the precision is better for larger sample sizes.
    - Also, 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 at 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 will always be some type of variability from sample to sample since the samples will be different from one another.
    - If each one of us took a random sample of people's preference of chocolate M&Ms vs peanut M&Ms on campus we would all get different sample proportions, $\hat{p}$, in favor of chocolate M&Ms.
- This variability can be expressed through a sampling distribution which provides us with both a mean and standard deviation to describe $\hat{p}$.

# Bias and Standard Error

- Bias and standard error are easy to find for a sample proportion under certain conditions.
    - The sample must be randomly selected from the population of interest, either with or without replacement.
    - If the sampling is without replacement, the population needs to be much larger than the sample size; at least 10 times bigger.
- Once these conditions are met, bias of $\hat{p}$ is 0 and the standard error is

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

- Note: In real life, we don't know the true value of the population proportion, p. This means we can't calculate the standard error, but we can come pretty close by using the sample proportion.

# Example

Clicker!

Suppose that in Pet World, the population is 1000 people and 25% of the population are Cat People. Cat People love cats but hate dogs. We are planning a survey in which we take a random sample of 100 people, without replacement. We calculate the proportion of people in our sample who are Cat People.

- What value should we expect for our sample proportion $\hat{p}$?

  A. 25%

  B. 10%

  C. Cannot be determined

# Example

- The sample proportion is unbiased (bias is 0) so we expect it to be the same as the population proportion. Hence, it is 25%.

# Example

Clicker!

Suppose that in Pet World, the population is 1000 people and 25% of the population are Cat People. Cat People love cats but hate dogs. We are planning a survey in which we take a random sample of 100 people, without replacement. We calculate the proportion of people in our sample who are Cat People.

- What's the standard error?

  A. 0.25

  B. 0.000625

  C. 0.0433

  D. 0

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

# Example

- The sample proportion is unbiased (bias is 0) so we expect it to be the same as the population proportion. Hence, it is 25%.
- The standard error is about 4.3%.

$$SE = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.25(1-0.25)}{100}} = 0.0433$$

We can use this formula because the population size (1000) is 10 times larger than the sample size (100). 100 x 10 = 1000

- If we were to take a survey of 100 people from Pet World, we would expect about 25% of them to be Cat People, give or take about 4.3%

# Sample Proportions

- The sampling distribution of $\hat{p}$ gives the probabilities of where our sample proportions will fall, i.e. it tells us how often we would see particular values of $\hat{p}$ if we could repeat our survey infinitely many times.
- We don't need to do infinitely many repetitions to see what the sampling distribution would look like.
- Instead, we can use the Central Limit Theorem (CLT).

# The Central Limit Theorem

- The Central Limit Theorem gives us a very good approximation of the sampling distribution without having to do infinitely many repetitions (or simulations).
- Sampling distributions are important.
    - Enable us to measure bias, standard error and the quality of our estimation methods.
    - Give us the probability that an estimate falls a specified distance from the population value.
    - For example, we don't want to know simply that 18% of <u>our</u> customers are likely to buy new cell phones in the next year. We also want to know the probability that the true percentage might be higher than some particular value, say, 25%.

# The Central Limit Theorem for Sample Proportions

- The Central Limit Theorem for Sample Proportions tells us that if some basic conditions are met, then the sampling distribution of the sample proportion is close to the Normal distribution.
    - Random and Independent: The sample is collected randomly and the trials are independent of each other.
    - Large Sample: The sample size is large enough that the sample expects at least 10 successes $np \geq 10$ and 10 failures. $n(1-p) \geq 10$
    - Big Population: If the sample is collected without replacement, then the population size is at least 10 times larger than the sample size. $N \geq 10n$

# The Central Limit Theorem for Sample Proportions

- The Central Limit Theorem for Sample Proportions tells us that if we take a random sample from a population, and if the sample size is large and the population size is much larger than the sample size, then the sampling distribution of $\hat{p}$ is approximately

$$N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

- If you don't know the value of p, then you can substitute the value of $\hat{p}$ to calculate the standard error.

# Example

Clicker!

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

- Are the conditions for the CLT met?

    1. Random and Independent?

Answer:

A. Yes
B. No

# Example

Clicker!

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

- Are the conditions for the CLT met?

    2. Large sample?

Answer:

A. Yes
B. No

# Example

Clicker!

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

- Are the conditions for the CLT met?

    3. Big population?

Answer:

A. Yes
B. No

# Example

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

Check the conditions:

- The drivers were randomly selected.
- Successes: $200 \times .24 = 48 \geq 10$
- Failures: $200 \times (1 - .24) = 152 \geq 10$
- Population size (# of American drivers) is very large. Much more than 10 times the sample size.

Conclusion: According to the CLT, the sampling distribution is approximately Normal.

# 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:

Clicker!

$$N\left(p, \sqrt{\frac{p(1-p)}{n}}\right) = ?$$

Answer:

A. N(0.03, 0.24)
B. N(0.24, 0.03)
C. N(0.24, 0.0009)

# 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:

$$N\left(p, \sqrt{\frac{p(1-p)}{n}}\right) = N\left(0.24, \sqrt{\frac{0.24(1-0.24)}{200}}\right) = N(0.24, 0.03)$$

# Example

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?

# Example

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?

The probability of 52 or more is the same as the same as saying probability of 52/200 = .26 or more.

Next we find the z-score for 0.26.

$$z = \frac{0.26 - 0.24}{0.03} = 0.67$$

The probability of getting a number bigger than 0.26 is 1 - 0.7486 = 0.25 or 25%. With a sample size of 200, there is about a 25% chance that the $\hat{p}$ will be more than 2 percentage points above 24%.

# Sampling Distribution of $\hat{p}$

- By the Empirical Rule, we know
  - About 68% of all samples will have $\hat{p}$ within 1 standard error of p
  - About 95% of all samples will have $\hat{p}$ within 2 standard errors of p
  - About 99.7% of all samples will have $\hat{p}$ within 3 standard errors of p
- Here, the standard error is the standard deviation for the sampling distribution.

# Realistically…

- It's usually the $\hat{p}$ (sample proportion) not the p (true population proportion) that is known.
- We collect data from a sample of observations, based on this sample we calculate a sample statistic (for example the $\hat{p}$) and use this statistic to make inference about a population parameter.

# Standard Deviation vs Standard Error

- Since we often don't know p, we can't find the true standard deviation of the sampling distribution model, so we use an estimate called the standard error.

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- Standard deviation of p: $SD(p) = SE(p) = \sqrt{\frac{p(1-p)}{n}}$

  ➡ We can only calculate when we know p, the true population proportion. Can use SD and SE interchangeably here (same formula).

- Standard error of $\hat{p}$: $SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

  ➡ We can calculate even if we only know $\hat{p}$, the sample proportion.

# Example

Let's return to Pet World. The population is 1000 people, and the proportion of Cat People is 25%. We'll take a random sample of 100 people.

- What is the probability that the proportion in our sample will be bigger than 29%? Begin by checking conditions for CLT.

# Example

Clicker!

Let's return to Pet World. The population is 1000 people, and the proportion of Cat People is 25%. We'll take a random sample of 100 people.

- Are the conditions for the CLT met?

    1. Random and Independent?

Answer:

A. Yes
B. No

# Example

Clicker!

Let's return to Pet World. The population is 1000 people, and the proportion of Cat People is 25%. We'll take a random sample of 100 people.

- Are the conditions for the CLT met?

    2. Large sample?

Answer:

A. Yes

B. No

# Example

Clicker!

Let's return to Pet World. The population is 1000 people, and the proportion of Cat People is 25%. We'll take a random sample of 100 people.

- Are the conditions for the CLT met?

  3. Big population?

Answer:

A. Yes
B. No

# Example

Let's return to Pet World. The population is 1000 people, and the proportion of Cat People is 25%. We'll take a random sample of 100 people.

- What is the probability that the proportion in our sample will be bigger than 29%? Begin by checking conditions for CLT.

Check the conditions:

- The people were randomly selected.
- Successes: np=100(0.25)=25 is greater than 10
- Failures: n(1-p)=100(0.75)=75 is greater than 10
- Population size is 10 times larger than the sample size, 10x100=1000.

Conclusion: According to the CLT, the sampling distribution will be approximately Normal.

# Example

Let's return to Pet World. The population is 1000 people, and the proportion of Cat People is 25%. We'll take a random sample of 100 people.

- What is the probability that the proportion in our sample will be bigger than 29%? Begin by checking conditions for CLT.

The mean is the same as the population proportion, p=0.25. The standard deviation is the same as the standard error:

$$SE = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.25(1-0.25)}{100}} = 0.0433$$

The probability of getting a value larger than 29 in a N(0.25, 0.0433).

$$z = \frac{0.29-0.25}{0.0433} = 0.92$$

Look up 0.92 in z-table and conclude that the probability is 1-0.8212=0.1788 or about 0.18 or 18%.

# Confidence Interval

- Looking at this from $\hat{p}$ point of view:
  - There is a 95% chance that p is no more than 2 SEs away from $\hat{p}$.
  - So, if we reach out 2 SEs, we are 95% sure that p will be in that interval. In other words, if we reach out 2 SEs in either direction of $\hat{p}$, we can be 95% confident that this interval contains the true population proportion.
  - This is called a 95% confidence interval.

# Why do we calculate CIs?

- Constructing a confidence interval is a way to estimate the true population proportion p when all we have is a sample proportion $\hat{p}$.
- Let's say we are interested in the proportion of UCLA students who have traveled outside the US.
    - We do not have the resources to ask every single UCLA student whether or not they have traveled outside the US.
    - Instead we take a random sample of 100 students and find that 35 of these students have at some point in their lives traveled outside the US, then $\hat{p} = 0.35$.
    - 35% is a reasonable guess for the true population proportion of UCLA students who have traveled outside the US.
    - However, it would be unreasonable to claim that the true population proportion must be equal to 35%.

# Why do we calculate CIs?

- In Statistics, when we are estimating a population parameter we choose to provide an interval where we believe this parameter may be in, and we also accompany our estimate with a measure of how certain we are.
- The way we calculate this interval is by adding and subtracting a margin of error (ME) to the sample statistic.
- Therefore confidence intervals in general have the form

$$\text{estimate} \pm \text{ME}$$

# Confidence Intervals in general have the form estimate ± ME

*Having a Gun in the House -- Safer or More Dangerous?*

Do you think having a gun in the house makes it a safer place to be or a more dangerous place to be?

| | % Safer | % More dangerous |
|---|---|---|
| Men | 67 | 26 |
| Women | 58 | 34 |
| | | |
| Whites | 65 | 28 |
| Nonwhites | 56 | 37 |
| | | |
| East | 59 | 34 |
| Midwest | 62 | 30 |
| South | 68 | 26 |
| West | 59 | 33 |
| | | |
| Republicans | 81 | 15 |
| Independents | 64 | 26 |
| Democrats | 41 | 53 |

Oct. 12-15, 2014

GALLUP

Survey Methods

Results for this Gallup poll are based on telephone interviews conducted Oct. 12-15, 2014, on Gallup U.S. Daily survey, with a random sample of 1,017 adults aged 18 and older, living in all 50 U.S. states and the District of Columbia.

For results based on the total sample of national adults, the margin of sampling error is ± 4 percentage points at the 95% confidence level.

# Calculating ME and CI

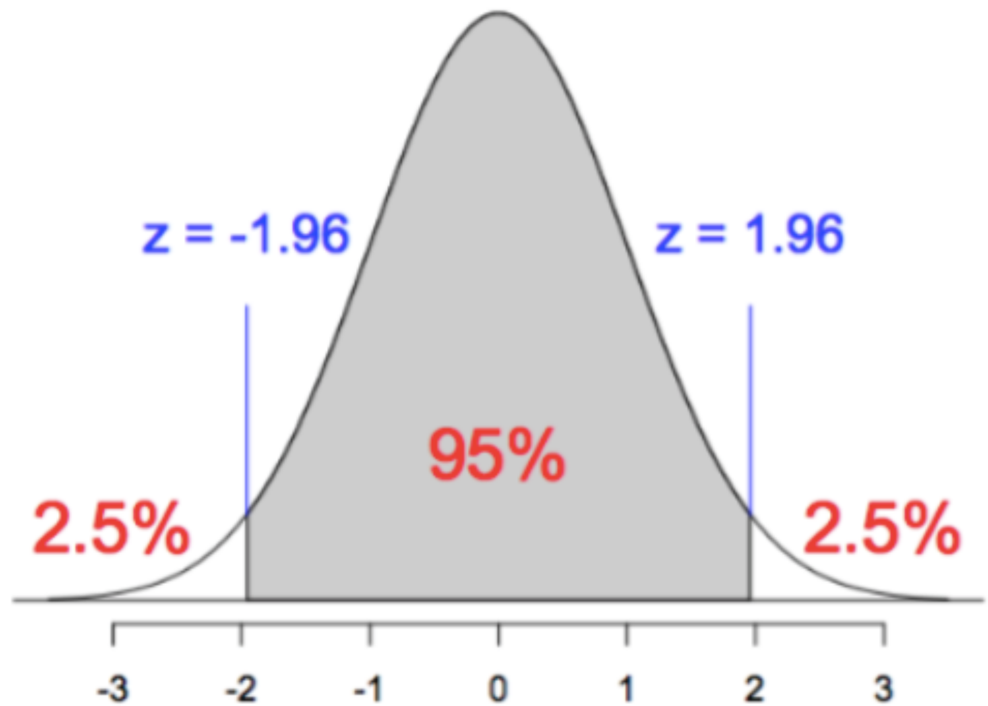
- The margin of error depends on two criteria:
    - how variable is $\hat{p}$
    - how confident do you want to be of your estimate
- If we want to be 95% confident, we should add/subtract about 2 standard errors to/from $\hat{p}$. Therefore a 95% confidence interval is roughly:

$$\hat{p} \pm \underbrace{2 * SE(\hat{p})}_{ME}$$

$$\hat{p} \pm ME$$

# About the Empirical Rule

…it's a bit of a rough estimate.




More accurately for data that is nearly normally distributed, about 95% of the data is within 1.96 standard deviations of the mean.

# Confidence Interval Formula

- Depending on the confidence level, we choose how many SEs to add/subtract from $\hat{p}$. So, a more generic formula for the CI can be written as

$$\hat{p} \pm \underbrace{z^* SE(\hat{p})}_{ME}$$

where $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

and $z^*$ is a critical z-score that depends on the confidence level.

# Critical Z-score

- So far we have seen that z* for a 95% confidence interval is 1.96. The method used to find these z* is a bit tedious.
- The best place to find these critical z-values is at the bottom of the t-table on the ∞ line (in the Appendix of the textbook).

| | | | | | |
|---|---|---|---|---|---|
| ∞ | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 |
| **Confidence levels** | 80% | 90% | 95% | 98% | 99% |

What is the z* for a 98% confidence interval?

Clicker!

A. 1.282

B. 1.645

C. 1.960

D. 2.326

# Critical Z-score

- So far we have seen that $z^*$ for a 95% confidence interval is 1.96. The method used to find these $z^*$ is a bit tedious.
- The best place to find these critical z-values is at the bottom of the t-table on the $\infty$ line (in the Appendix of the textbook).

| | | | | | |
|---|---|---|---|---|---|
| $\infty$ | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 |
| **Confidence levels** | 80% | 90% | 95% | 98% | 99% |

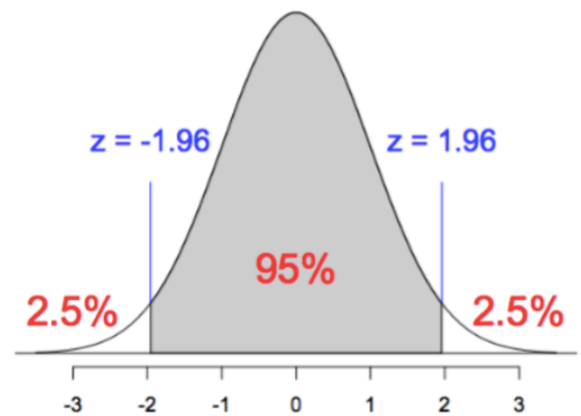
We can see that $z^*$ for a 95% confidence interval is 1.96, which is almost 2. This goes back to the 65-95-99.7 Rule which we interpret as "if we reach out approximately 2 SEs in either direction of $\hat{p}$, we can be 95% confident that this interval contains the true population proportion."

# Critical Z-score

- The t-table only gives critical z-scores ($z^*$) for a select number of confidence levels. These are the confidence levels that are usually used in practice.
- In order to find the critical z-score for any other confidence level we use the z-table.
- But you'll see that a great majority of the time we'll use 95% confidence intervals.

# Example

- For data that is nearly normally distributed, within how many standard deviations of the mean is the middle 95% of the data?




$$\alpha = 0.95$$

$$1-\alpha = 1-0.95 = 0.05$$

$$\frac{0.05}{2} = 0.025$$

If you are using the positive table, 0.025 corresponds to the area to the right of the curve so 1-0.025=0.975.

We use the positive number to identify the critical z-score or z*.

| z | .00 | .01 | .02 | .03 | .04 | .05 | .06 |
|---|---|---|---|---|---|---|---|
| −1.9 | .0287 | .0281 | .0274 | .0268 | .0262 | .0256 | .0250 |
| −1.8 | .0359 | .0351 | .0344 | .0336 | .0329 | .0322 | .0314 |
| −1.7 | .0446 | .0436 | .0427 | .0418 | .0409 | .0401 | .0392 |
| −1.6 | .0548 | .0537 | .0526 | .0516 | .0505 | .0495 | .0485 |

| z | .00 | .01 | .02 | .03 | .04 | .05 | .06 |
|---|---|---|---|---|---|---|---|
| 1.6 | .9452 | .9463 | .9474 | .9484 | .9495 | .9505 | .9515 |
| 1.7 | .9554 | .9564 | .9573 | .9582 | .9591 | .9599 | .9608 |
| 1.8 | .9641 | .9649 | .9656 | .9664 | .9671 | .9678 | .9686 |
| 1.9 | .9713 | .9719 | .9726 | .9732 | .9738 | .9744 | .9750 |

# Example

Clicker!

- For data that is nearly normally distributed, within how many standard deviations of the mean is the middle 90% of the data? Choose the closest value.

A. 1.282

B. 1.645

C. 1.96

D. 1.745

| z | .00 | .01 | .02 | .03 | .04 | .05 | .06 |
|---|---|---|---|---|---|---|---|
| −1.6 | .0548 | .0537 | .0526 | .0516 | .0505 | .0495 | .0485 |
| −1.5 | .0668 | .0655 | .0643 | .0630 | .0618 | .0606 | .0594 |
| −1.4 | .0808 | .0793 | .0778 | .0764 | .0749 | .0735 | .0721 |

| z | .00 | .01 | .02 | .03 | .04 | .05 | .06 |
|---|---|---|---|---|---|---|---|
| 1.4 | .9192 | .9207 | .9222 | .9236 | .9251 | .9265 | .9279 |
| 1.5 | .9332 | .9345 | .9357 | .9370 | .9382 | .9394 | .9406 |
| 1.6 | .9452 | .9463 | .9474 | .9484 | .9495 | .9505 | .9515 |
| 1.7 | .9554 | .9564 | .9573 | .9582 | .9591 | .9599 | .9608 |

# Example

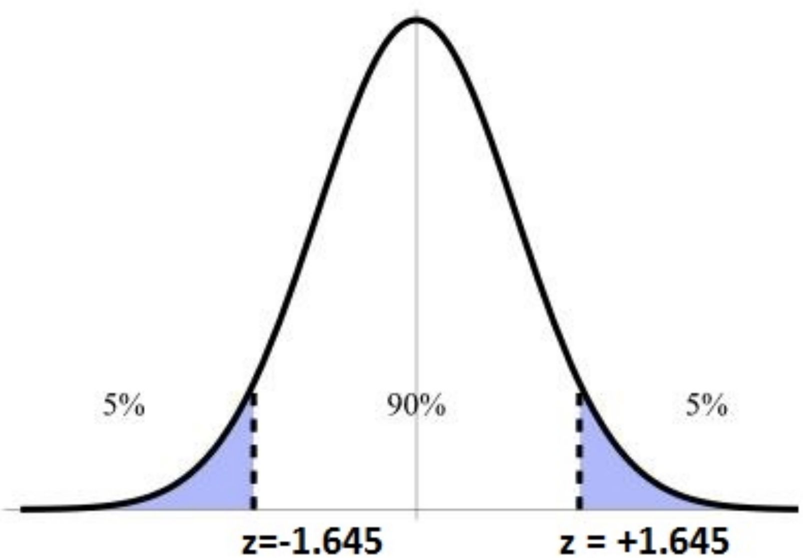
- For data that is nearly normally distributed, within how many standard deviations of the mean is the middle 90% of the data? Choose the closest value.

$$\alpha = 0.90$$

$$1 - \alpha = 1 - 0.90 = 0.10$$

$$\frac{0.10}{2} = 0.05$$

If you are using the positive table, 0.05 corresponds to the area to the right of the curve so 1-0.05=0.95




| z | .00 | .01 | .02 | .03 | .04 | .05 | .06 |
|---|---|---|---|---|---|---|---|
| −1.6 | .0548 | .0537 | .0526 | .0516 | .0505 | .0495 | .0485 |
| −1.5 | .0668 | .0655 | .0643 | .0630 | .0618 | .0606 | .0594 |
| −1.4 | .0808 | .0793 | .0778 | .0764 | .0749 | .0735 | .0721 |

| z | .00 | .01 | .02 | .03 | .04 | .05 | .06 |
|---|---|---|---|---|---|---|---|
| 1.4 | .9192 | .9207 | .9222 | .9236 | .9251 | .9265 | .9279 |
| 1.5 | .9332 | .9345 | .9357 | .9370 | .9382 | .9394 | .9406 |
| 1.6 | .9452 | .9463 | .9474 | .9484 | .9495 | .9505 | .9515 |
| 1.7 | .9554 | .9564 | .9573 | .9582 | .9591 | .9599 | .9608 |

# Interpretation of a 95% CI

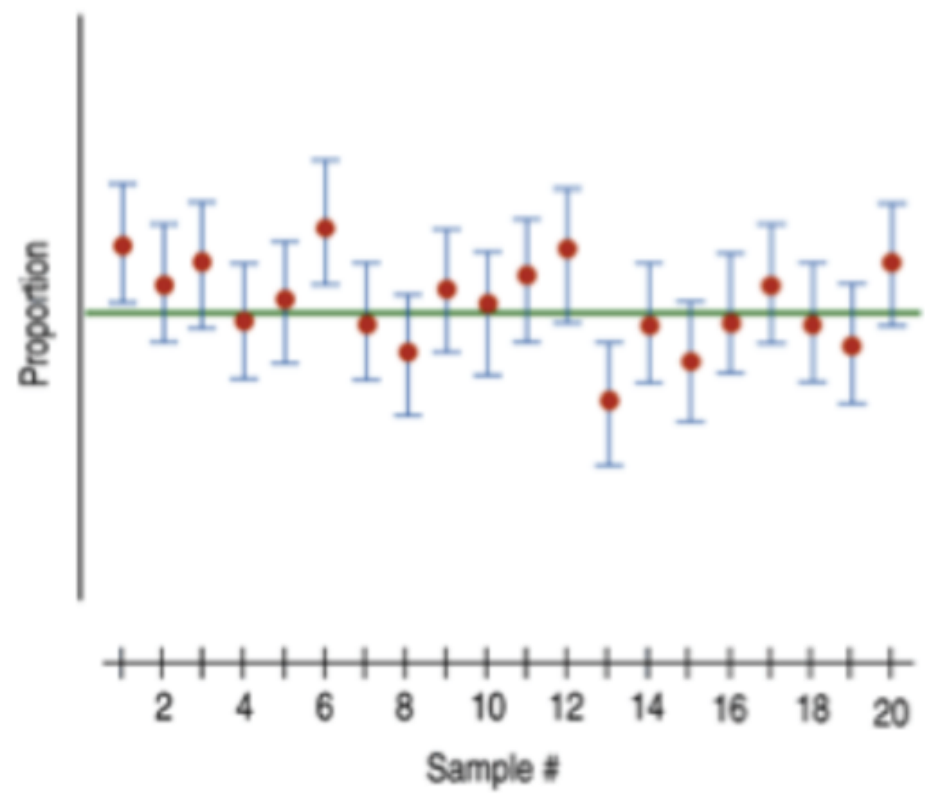
- We survey a group of Americans and learn that 63% believe having a gun in the house makes it a safer place to be with a margin of error of 4%.
- Then we are 95% confident that between 59% to 67% of all Americans think that having a gun makes it a safer place to be.
- More generically: We are 95% confident that the true population proportion lies in the interval.

# What does 95% mean?

- Since samples vary, the confidence intervals constructed based on these samples will vary as well.
- 95% confident means that 95% of random samples of of size n will produce confidence intervals that include the true population proportion.

# What does 95% mean?

This figure shows that some of our confidence intervals capture the true proportion (the green horizontal line), while others do not.



# Example

In a random sample of 100 UCLA students, 35 have traveled outside the US. Assume none of the students are friends or siblings. Estimate the true population proportion of UCLA students who traveled outside the US using a 95% confidence interval.

First, check the conditions.

Clicker!

Are all the conditions met?

A. Yes

B. No

# Example

In a random sample of 100 UCLA students, 35 have traveled outside the US. Estimate the true population proportion of UCLA students who traveled outside the US using a 95% confidence interval.

Check the conditions:

- The UCLA students were randomly selected.
- Successes: $100 \times .35 = 35 \geq 10$
- Failures: $100 \times (1 - .35) = 65 \geq 10$
- Population size is larger than 10 times the sample size.

Conclusion: According to the CLT, the sampling distribution is approximately Normal.

# Example

In a random sample of 100 UCLA students, 35 have traveled outside the US. Estimate the true population proportion of UCLA students who traveled outside the US using a 95% confidence interval.

$$\hat{p} \pm z^{*} SE\left(\hat{p}\right)$$

$$0.35 \pm 1.96 \times \sqrt{\frac{0.35 \times 0.65}{100}}$$

$$0.35 \pm 1.96 \times 0.0477$$

$$0.35 \pm 0.09$$

$$\left(0.35 - 0.09, 0.35 + 0.09\right)$$

$$\left(0.26, 0.44\right)$$

$$0.26 < p < 0.44$$

We are 95% confident that the true population proportion of UCLA students who traveled outside the US is between 26% and 44%.

This means that, 95% of random samples of size n=100 will produce confidence intervals that contain the true population proportion.

# Example

Clicker!

In a random sample of 100 UCLA students, 35 have traveled outside the US.

Estimate the true population proportion of UCLA students who traveled outside the US using a 99% confidence interval.

Answer:

A. (0.26, 0.44)

B. (0.23, 0.47)

C. (0.33, 0.52)

D. (0.18, 0.28)

# Example

In a random sample of 100 UCLA students, 35 have traveled outside the US. Estimate the true population proportion of UCLA students who traveled outside the US using a 99% confidence interval.

$$\hat{p} \pm z^{*} SE(\hat{p})$$

$$0.35 \pm 2.576 \times \sqrt{\frac{0.35 \times 0.65}{100}}$$

$$0.35 \pm 2.576 \times 0.0477$$

$$0.35 \pm 0.123$$

$$(0.35 - 0.123, 0.35 + 0.123)$$

$$(0.23, 0.47)$$

$$0.23 < p < 0.47$$

We are 99% confident that the true population proportion of UCLA students who traveled outside the US is between 23% and 47%.

# Interval Width Based on Confidence

Would you expect a 90% confidence interval to be wider or narrower than the 99% confidence interval you calculated?

Clicker!

HINT: You do not need to do any calculations to answer this question, think about how many standard errors you'll need to add/subtract to/from the sample proportion to capture the middle 90% of the distribution.

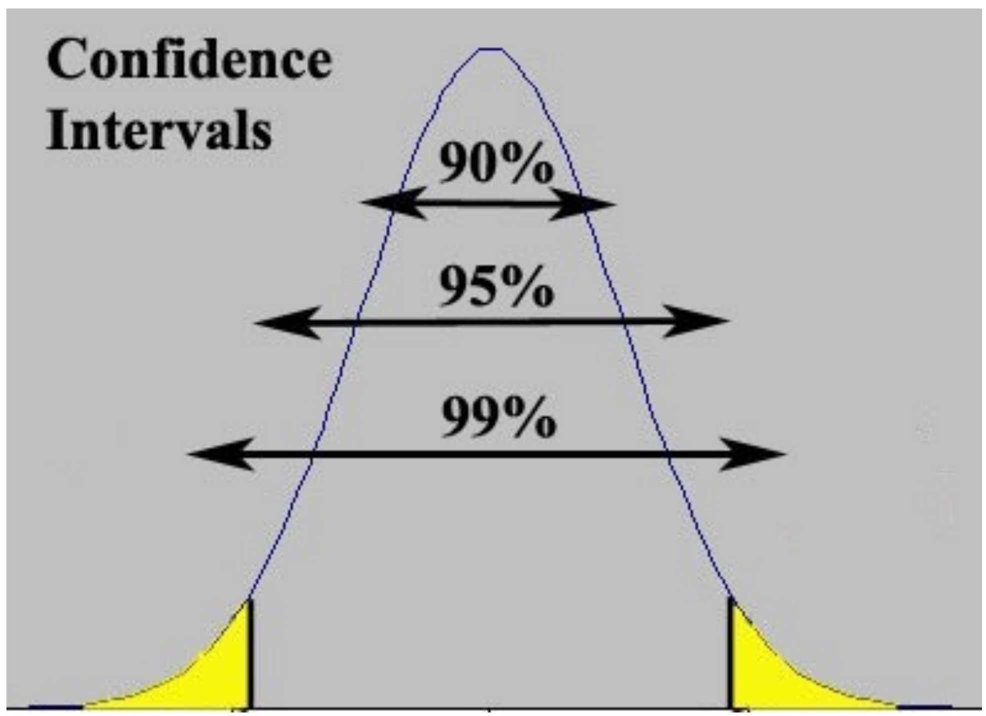
Answer:

A. Wider

B. Narrower

# Interval Width Based on Confidence

Would you expect a 90% confidence interval to be wider or narrower than the 99% confidence interval you calculated?



# How confident do we want to be?

- Being more confident is favorable however intervals that are too wide are not informative.
- Everyone knows that the true proportion of UCLA Students who traveled abroad is between 0% and 100%. Such a wide interval is not informative even if we can be 100% confident.

# Comparing Proportions from 2 Populations

- In 2002, a Pew Poll based on a random sample of 1500 people suggested that 43% of the American public approved of stem cell research.
- In 2009, a new poll of a different sample of 1500 people found that 58% approved.
- Did American opinion change? Perhaps!
- But it is also possible that these two sample proportions are different because the samples used different people.
- The people were randomly selected but random samples can vary.
- Possibly, the sample proportions differed just by chance.
- Even though the sample proportions are different, the population proportions might be the same!

# Comparing Proportions from 2 Populations

- Compare 2 sample proportions each from a different group
  - Are the two groups the same? Are they different?
- Same as the ideas for the one proportion, but the formulas are different.
  - Confidence interval for difference in proportions

# Review: CI for one proportion (Ch 7.4)

- Depending on the confidence level, we choose how many SEs to add/subtract from $\hat{p}$. So, a more generic formula for the CI can be written as

$$\hat{p} \pm \underbrace{z^* SE(\hat{p})}_{ME}$$

where $SE = \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$

and $z^*$ is a critical z-score that depends on the confidence level.

# Confidence Intervals for the difference in proportions

- Confidence intervals are one method to determine whether different sample proportions reflect “real” differences in the population.
    - We find a confidence interval for the difference in proportions $p_1 - p_2$.
    - We check to see whether the interval includes 0.

# CI for the difference in proportions (Ch 7.5)

- When we have two populations to compare, we can also find a confidence interval for the difference between the two population proportions:

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \times SE_{est}$$

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

where $z^*$ is still the critical z-score that depends on the confidence level. The numbers 1 and 2 indicate the two different groups.

# Interpreting CI for difference in proportions

- If the confidence interval captures 0, it means the population proportions could be equal. Because if $p_1$-$p_2$ = 0, then $p_1$=$p_2$ and the proportions are the same.
- If it does not capture 0, then we are confident that the population proportions are not equal, and we should note whether the values in the interval are all positive (the first population proportion is greater than the second) or all negative (the first is less than the second).

# Example

In 2013, 2000 men and 2000 women were randomly selected for a Pew survey. Interviewers reported that the proportion of men who felt that embryonic stem cell research was morally wrong was 0.23, and the proportion of women who felt it was morally wrong was 0.21. We wish to find a 95% confidence interval for the difference in proportions between men and women who feel this way in the population.

First distinguish your samples.

Sample 1: women

Sample 2: men

# Example

Sample 1: women, n=2000, $\hat{p}_1$=0.21

Sample 2: men, n=2000, $\hat{p}_2$=0.23

Next check your conditions!

Random and independent: We are told the samples are random, and we must assume that the observations are independent of each other in both samples.

Large sample: Check both samples!

$$n_1\hat{p}_1 \geq 10 = 2000 \times 0.21 = 420 \geq 10$$

$$n_1\left(1-\hat{p}_1\right) \geq 10 = 2000 \times \left(1-0.21\right) = 1580 \geq 10$$

$$n_2\hat{p}_2 \geq 10 = 2000 \times 0.23 = 460 \geq 10$$

$$n_2\left(1-\hat{p}_2\right) \geq 10 = 2000 \times \left(1-0.23\right) = 1540 \geq 10$$

Big population: Clearly there are more than 10 x 2000 = 20,000 men and more than 20,000 women in the United States.

# Example

Next the 95% confidence interval:

z* is 1.96

$$\left(\hat{p}_1 - \hat{p}_2\right) \pm z^* \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$(0.21 - 0.23) \pm 1.96 \times \sqrt{\frac{0.21(1-0.21)}{2000} + \frac{0.23(1-0.23)}{2000}} = \left(-0.046, 0.006\right)$$

The 95% confidence interval for the difference in proportions between men who believe the research is morally wrong and the proportion of women who believe it is wrong was -0.046 to 0.006. We are 95% confident that the difference in population proportions is between -0.046 and 0.006.

# Example

Last step: Interpret the confidence interval.

The confidence interval is -0.046 to 0.006. The interval contains 0. Thus we can't rule out the possibility that the proportion of men who believe this and the proportion of women who believe it are the same.

# Clicker Test

Clicker!

- Freebie! Click in choice C!

# Example - Physicians' Health Study I

Physicians' Health Study I (to study aspirin's affect on reducing heart attacks).

- Started in 1982 with 22,071 male physicians.
- The physicians were **randomly assigned** into one of two groups.
  - Half took a 325mg aspirin every other day.
  - Half took a placebo.

Source: http://phs.bwh.harvard.edu/phs1.htm

# The Big Idea

- Randomly assigning people to groups tends to balance out all other variables between the groups.
- So variables that could have an effect on the response should be equalized between the two groups
- Thus, cause and effect conclusions are possible in experiments through random assignment.  (It must be a well run experiment.)
- An experiment where subjects are randomly assigned is called a randomized experiment.

# Random vs Random

- With observational studies, **random sampling** is often done. This possibly allows us to make inferences from the sample to the population where the sample was drawn.
- With experiments, **random assignment** is done. This allows us to possibly conclude causation.
- Note: random sampling and random assignment are not the same.

# Random vs Random

- The Physician’s Health Study used random assignment, did it also use random sampling?
- No, hardly any experiments use random sampling. Subjects are chosen in other ways.
- The Physician’s Health Study sent out invitation letters and questionnaires to all 261,248 male physicians between 40 and 84 years of age who lived in the United States.
- Of the 59,285 who were willing to participate in the trial, 26,062 were told they could not because of some medical condition or current medical treatment.

# Random vs Random

- So to what group can we generalize the results that taking aspirin can reduce heart attacks?
  - Just physicians in the study?
  - All male physicians between 40-84 years old?
  - All males physicians?
  - All males between 40-84 years olds?
  - All males?
  - Everyone between 40-84 years old?
  - Everyone?

# Example

A study was done to see which of the two treatments for Crohn’s disease, Inflix injections or Azath pills, was better. Patients were randomly assigned to receive either Inflix or Azath. 169 patients received Inflix, and at the end of the study, 75 of them were in remission (a good outcome). 170 patients received Azath, and at the end of the study, 51 were in remission.

First distinguish your samples.

Sample 1: Inflix group

Sample 2: Azath group

# Example

A study was done to see which of the two treatments for Crohn's disease, Inflix injections or Azath pills, was better. Patients were randomly assigned to receive either Inflix or Azath. 169 patients received Inflix, and at the end of the study, 75 of them were in remission (a good outcome). 170 patients received Azath, and at the end of the study, 51 were in remission.

Next, check the conditions.

Clicker!

Is condition 1 met? Random and independent.

A. Yes

B. No

# Example

A study was done to see which of the two treatments for Crohn's disease, Inflix injections or Azath pills, was better. Patients were randomly assigned to receive either Inflix or Azath. 169 patients received Inflix, and at the end of the study, 75 of them were in remission (a good outcome). 170 patients received Azath, and at the end of the study, 51 were in remission.

Next, check the conditions.

Is condition 1 met? Random and independent.

A. Yes

B. No

Random and independent: The samples are not randomly selected from the population. However, we must assume that the observations are independent of each other in both samples.

# Example

A study was done to see which of the two treatments for Crohn's disease, Inflix injections or Azath pills, was better. Patients were randomly assigned to receive either Inflix or Azath. 169 patients received Inflix, and at the end of the study, 75 of them were in remission (a good outcome). 170 patients received Azath, and at the end of the study, 51 were in remission.

Clicker!

Is condition 2 met? Large Sample.

A. Yes

B. No

# Example

A study was done to see which of the two treatments for Crohn's disease, Inflix injections or Azath pills, was better. Patients were randomly assigned to receive either Inflix or Azath. 169 patients received Inflix, and at the end of the study, 75 of them were in remission (a good outcome). 170 patients received Azath, and at the end of the study, 51 were in remission.

Is condition 2 met? Large Sample.

A. Yes

B. No

Sample 1: Inflix, n=169, $\hat{p}_1$=75/169=0.44

Sample 2: Azath, n=170, $\hat{p}_2$=51/170=0.30

$$n_1\hat{p}_1 \geq 10 = 169 \times 0.44 = 74.36 \geq 10$$

$$n_1\left(1-\hat{p}_1\right) \geq 10 = 169 \times \left(1-0.44\right) = 94.64 \geq 10$$

$$n_2\hat{p}_2 \geq 10 = 170 \times 0.30 = 51 \geq 10$$

$$n_2\left(1-\hat{p}_2\right) \geq 10 = 170 \times \left(1-0.30\right) = 119 \geq 10$$

# Example

A study was done to see which of the two treatments for Crohn's disease, Inflix injections or Azath pills, was better. Patients were randomly assigned to receive either Inflix or Azath. 169 patients received Inflix, and at the end of the study, 75 of them were in remission (a good outcome). 170 patients received Azath, and at the end of the study, 51 were in remission.

Clicker!

Is condition 3 met? Big Population.

A. Yes

B. No

# Example

A study was done to see which of the two treatments for Crohn's disease, Inflix injections or Azath pills, was better. Patients were randomly assigned to receive either Inflix or Azath. 169 patients received Inflix, and at the end of the study, 75 of them were in remission (a good outcome). 170 patients received Azath, and at the end of the study, 51 were in remission.

Is condition 3 met? Big Population.

A. Yes
B. No

# Example

Clicker!

A study was done to see which of the two treatments for Crohn's disease, Inflix injections or Azath pills, was better. Patients were randomly assigned to receive either Inflix or Azath. 169 patients received Inflix, and at the end of the study, 75 of them were in remission (a good outcome). 170 patients received Azath, and at the end of the study, 51 were in remission.

Sample 1: Inflix, n=169, $\hat{p}_1$=75/169=0.44

Sample 2: Azath, n=170, $\hat{p}_2$=51/170=0.30

What is the 95% confidence interval for the difference in proportions?

A. (0.25, 0.60)

B. (0.04, 0.24)

C. (-0.24, -0.04)

D. (0.34, 0.98)

# Example

Next the 95% confidence interval:

z* is 1.96

$$\left(\hat{p}_1 - \hat{p}_2\right) \pm z^{*} \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$\left(0.44 - 0.30\right) \pm 1.96 \times \sqrt{\frac{0.44\left(1-0.44\right)}{169} + \frac{0.30\left(1-0.30\right)}{170}} = \left(0.04, 0.24\right)$$

We are 95% confident that the difference in population proportions is between 0.04 and 0.24.

# Example

Last step: Interpret the confidence interval.

- The confidence interval is 0.04 to 0.24.
- Even though the two samples are not randomly selected from the population, the fact that patients were randomly assigned to one of the two treatments, together with the fact that the other conditions hold, means we can interpret the confidence interval.
- The interval does not contain 0. This tells us that we are confident that one treatment is different from the other.
- Also, the values of the confidence interval are all positive. This means the proportion of people in remission is greater for population 1, which consists of those who took Inflix.

# Example

Conclusion:

- Inflix is the better treatment.
- The percentage of people who will go into remission is at least 4 percentage points greater with Inflix than with Azath, and it could be as much as 24 percentage points greater.