Module: Basics of Statistical Inference

Module Outline

- Introduction
- Sampling
- The main theorems underlying statistical inference:

  Law of Large Numbers and *Central Limit Theorem*
- Confidence Intervals
- Hypothesis testing

Statistics in the Social Sciences

- Answering "What is" questions:

  What is the average age of the Canadian population?

  What fraction of the Canadian population:

  smoke? *support the Liberals/Conservatives*?

  support a tax increase to fund free child-care?

Statistics in the Social Sciences

- Answering "What is" questions:

  What is the average age of the Canadian population?

  What fraction of the Canadian population:

  smoke? *support the Liberals/Conservatives*?

  support a tax increase to fund free child-care?

- *Answering policy impact questions*:

  What is the impact of subsidized child-care on children and parents?

  What is the impact of providing free laptops on children's education?

How to answer questions using data?

- Answering: "Average age of the Cdn population?"

  Or "Is the average age in Canada higher than in Japan?"

How to answer questions using data?

- Answering: "Average age of the Cdn population?"

  Or "Is the average age in Canada higher than in Japan?"

- *Data for the entire population:* rare (depends on the variable)

How to answer questions using data?

- Answering: "Average age of the Cdn population?"

  Or "Is the average age in Canada higher than in Japan?"

- *Data for the entire population:* rare (depends on the variable)

- *Examples*: Census, Crime Records, Customs (Trade data), Sports Leagues, Tax records,....

How to answer questions using data?

- Answering: "Average age of the Cdn population?"

  Or "Is the average age in Canada higher than in Japan?"

- *Data for the entire population:* rare (depends on the variable)

- *Examples*: Census, Crime Records, Customs (Trade data), Sports Leagues, Tax records,....

- *Much more common:* draw a sample (a fraction of the population)

How to answer questions using data?

- Answering: "Average age of the Cdn population?"

  Or "Is the average age in Canada higher than in Japan?"
- *Data for the entire population:* rare (depends on the variable)
- *Examples*: Census, Crime Records, Customs (Trade data), Sports Leagues, Tax records,....
- <u>*Much more common*</u>*:* draw a sample (a fraction of the population)
- *Examples*: Sample surveys, opinion polls

Sample surveys / Polls

- *Examples*:

  Govt. organizations: *Statcan* (LFS, NTS, CSS,....), *US Census Bureau* (ACS, HPS,....), *NSS* (India), *NBS* of China,.....

- <u>*Examples*</u>:

  Govt. organizations: *Statcan* (LFS, NTS, CSS,....), *US Census Bureau* (ACS, HPS,....), *NSS* (India), *NBS* of China,.....

- Commercial organizations: *Pew Research Center*, *Gallup*, *Nielsen*,......

Pew Research Center

Sample surveys / Polls

- *Examples*:

  Govt. organizations: *Statcan* (LFS, NTS, CSS,....), *US Census Bureau* (ACS, HPS,....), *NSS* (India), *NBS* of China,.....

- Commercial organizations: *Pew Research Center*, *Gallup*, *Nielsen*,......

- Individuals: *Teaching surveys, customer experience surveys*,......

Statistical inference: Main questions

- **Main questions**:

  (a) How do we use data from the sample to make inferences about the entire population(s)?

  (b) How much confidence do we have in our findings?

Statistical inference: Main questions

- **Main questions**:

  (a) How do we use data from the sample to make inferences about the entire population(s)?

  (b) How much confidence do we have in our findings?

- More than a guess?

Sampling: Main issues

- Maybe impossible / impractical to examine everyone we are interested in knowing about (*Population*)

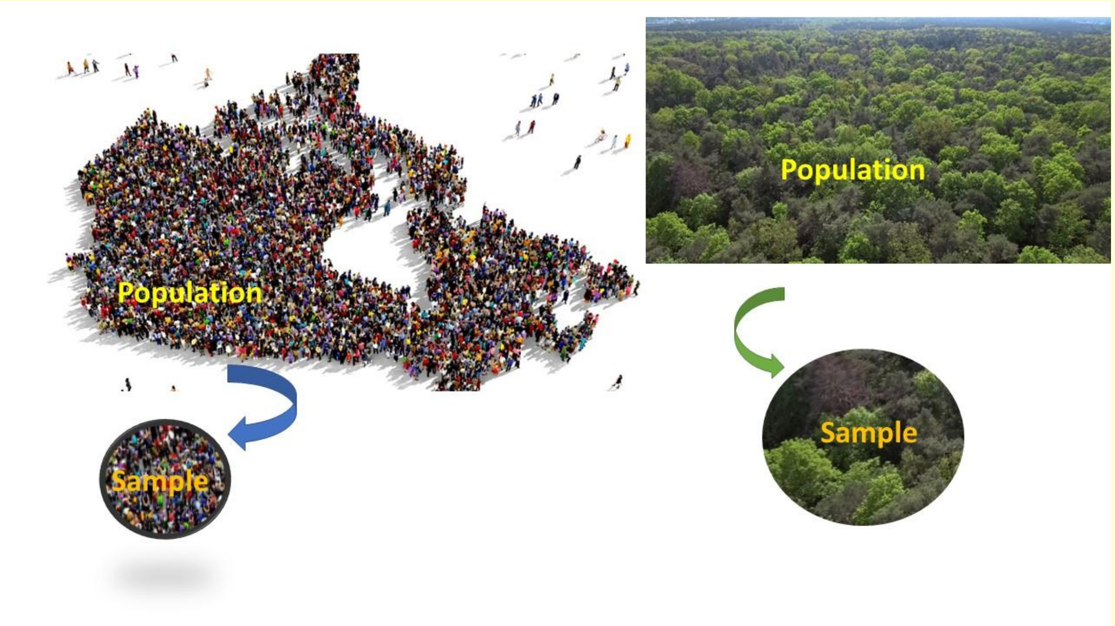
  What is the population?

Sampling: Main issues

- Maybe impossible / impractical to examine everyone we are interested in knowing about (*Population*)

  What is the population?

- Draw a *sample*: a small part of the population



Sampling: Main issues

- Maybe impossible / impractical to examine everyone we are interested in knowing about (*Population*)

  What is the population?

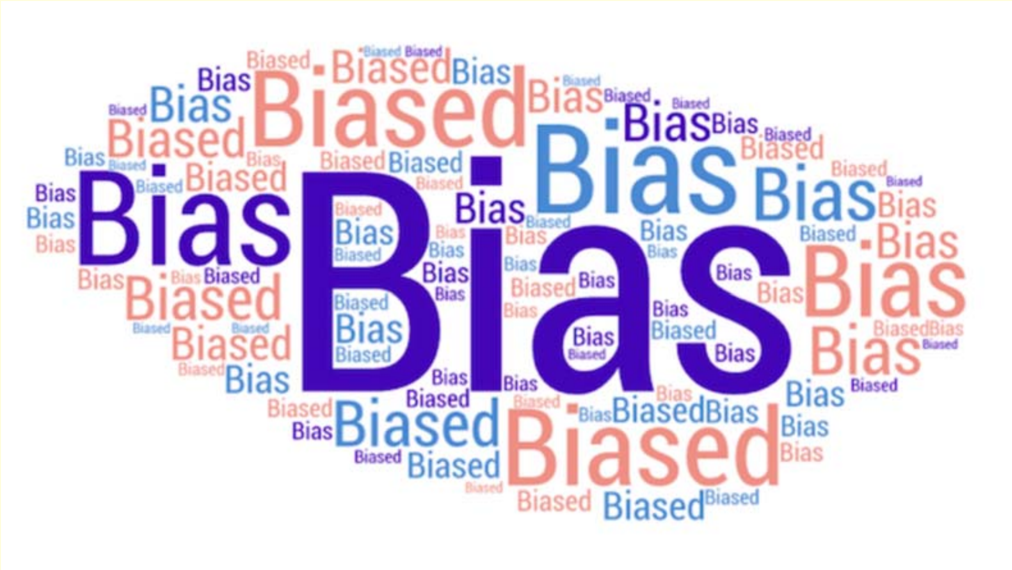
- Draw a *sample*: a small part of the population

- *Issues with drawing a sample:*

  (i) Bias?

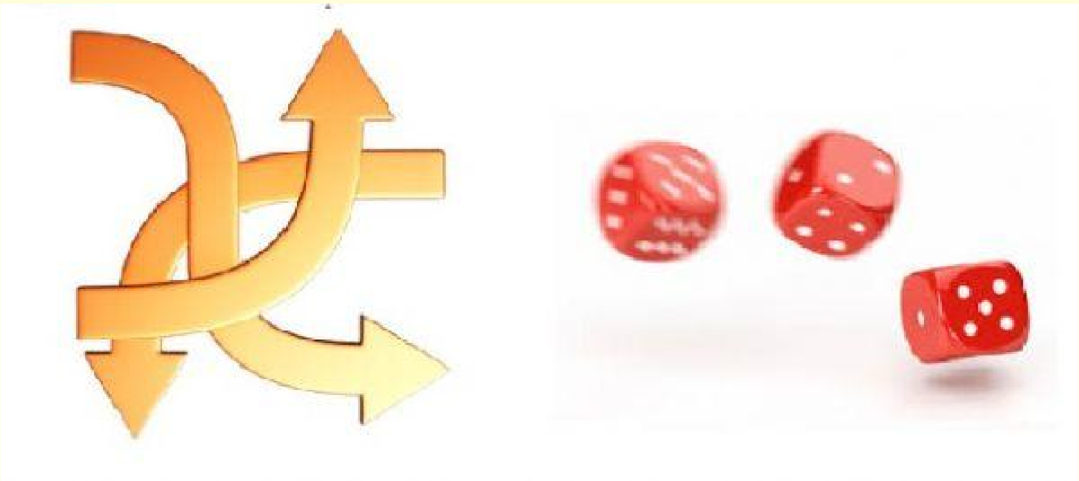
Sampling: Main issues

- Bias may not always be apparent



Sampling: Main issues

- Bias may not always be apparent
- Solution: **Randomize**



Sampling: Main issues

- Bias may not always be apparent
- Solution: **Randomize**
- Randomization ensures that on average, the sample should look like the population

Sampling: Main issues

- Bias may not always be apparent
- Solution: **Randomize**
- Randomization ensures that on average, the sample should look like the population
- (ii) *Make sure the sample is representative of the population*

Sampling: Main issues

- *Choosing a representative sample:*

  Try to match many characteristics of the population

Sampling: Main issues

- *Choosing a representative sample:*

  Try to match many characteristics of the population

- *Typical characteristics*: age, sex, ethnicity, education,.......

Sampling: Main issues

- *Choosing a representative sample:*

  Try to match many characteristics of the population

- *Typical characteristics*: age, sex, ethnicity, education,.......

- (iii) *How large a random sample do we need to choose?*

Sampling: Main issues

- *Choosing a representative sample:*

  Try to match many characteristics of the population

- *Typical characteristics*: age, sex, ethnicity, education,.......

- (iii) *How large a random sample do we need to choose?*

- What matters is the absolute size of the sample

  NOT its size relative to the population

Sampling: Main issues

- *Choosing a representative sample:*

  Try to match many characteristics of the population

- *Typical characteristics*: age, sex, ethnicity, education,.......

- (iii) *How large a random sample do we need to choose?*

- What matters is the <u>absolute size</u> of the sample

  NOT its size relative to the population

- WARNING: If a larger population is also more diverse, then its even more important to have a representative sample

Parameter and Statistic

- *Suppose you have a representative sample*

  Q: What are you trying to find about the population?

Parameter and Statistic

- *Suppose you have a representative sample*

  Q: What are you trying to find about the population?

- **Parameter** μ: Some characteristic about the population (often used in modelling)

Parameter and Statistic

- *Suppose you have a representative sample*

  Q: What are you trying to find about the population?

- **Parameter** μ: Some characteristic about the population (often used in modelling)

- *Examples*: number of daily interactions, loans, fraction who support a tax increase, fraction who drive over the speed limit,.....

Parameter and Statistic

- *Suppose you have a representative sample*

  Q: What are you trying to find about the population?

- **Parameter** μ: Some characteristic about the population (often used in modelling)

- *Examples*: number of daily interactions, loans, fraction who support a tax increase, fraction who drive over the speed limit,.....

- Measure this (or a related) variable X in the sample

Parameter and Statistic

- *Suppose you have a representative sample*

  Q: What are you trying to find about the population?

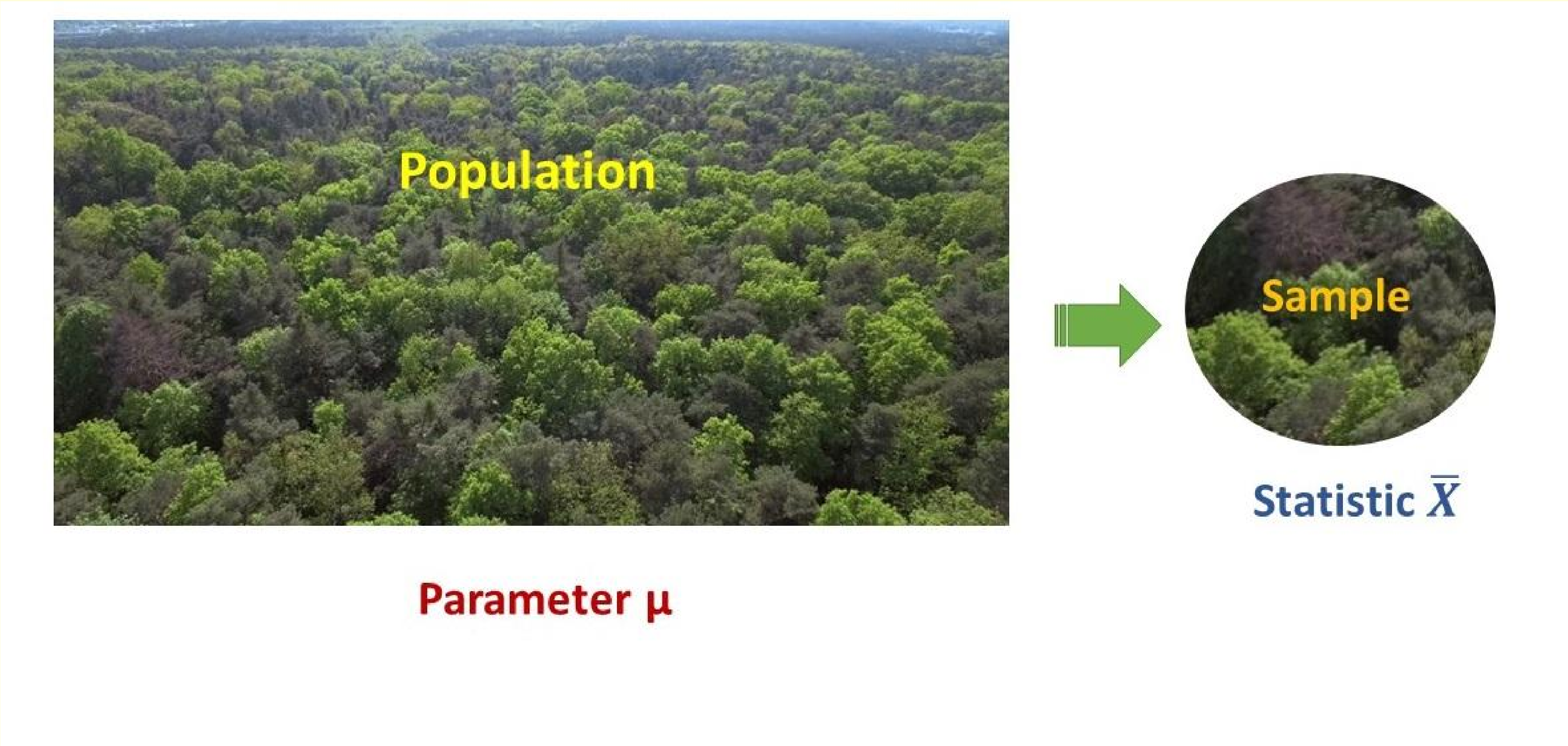
- **Parameter** μ: Some characteristic about the population (often used in modelling)

- *Examples*: number of daily interactions, loans, fraction who support a tax increase, fraction who drive over the speed limit,.....

- Measure this (or a related) variable X in the sample

- **Sample statistic**: Use *data from the sample* to estimate the population parameter e.g. $\overline{X}$

Parameter and Statistic



Estimates

- Have a sample of size n, measure a *variable* X, now what?

Estimates

- Have a sample of size n, measure a *variable* X, now what?
- X **is a quantitative variable**:

  $\overline{X}$ is an estimate of the mean value of X in the population

Estimates

- Have a sample of size n, measure a *variable* X, now what?
- X **is a quantitative variable**:

  $\overline{X}$ is an estimate of the mean value of X in the population

- Examples: Amount of household debt, time spent commuting daily, number of friends,........

Estimates

- X **is a categorical {0,1} variable**:

  $p =$ proportion of people with $X = 1$ in the sample is an estimate of the fraction of people with $X = 1$ in the population

Estimates

- X **is a categorical {0,1} variable**:

  $p =$ proportion of people with $X = 1$ in the sample is an estimate of the fraction of people with $X = 1$ in the population

- Examples: Fraction who smoke, who support policy P, who are single,.........

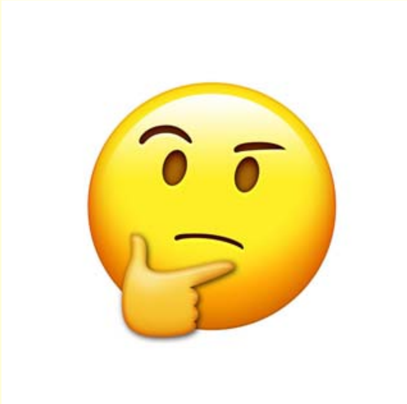
Estimates

- X **is a categorical {0,1} variable**:

  p = proportion of people with $X = 1$ in the sample is an estimate of the fraction of people with $X = 1$ in the population

- Examples: Fraction who smoke, who support policy P, who are single,........

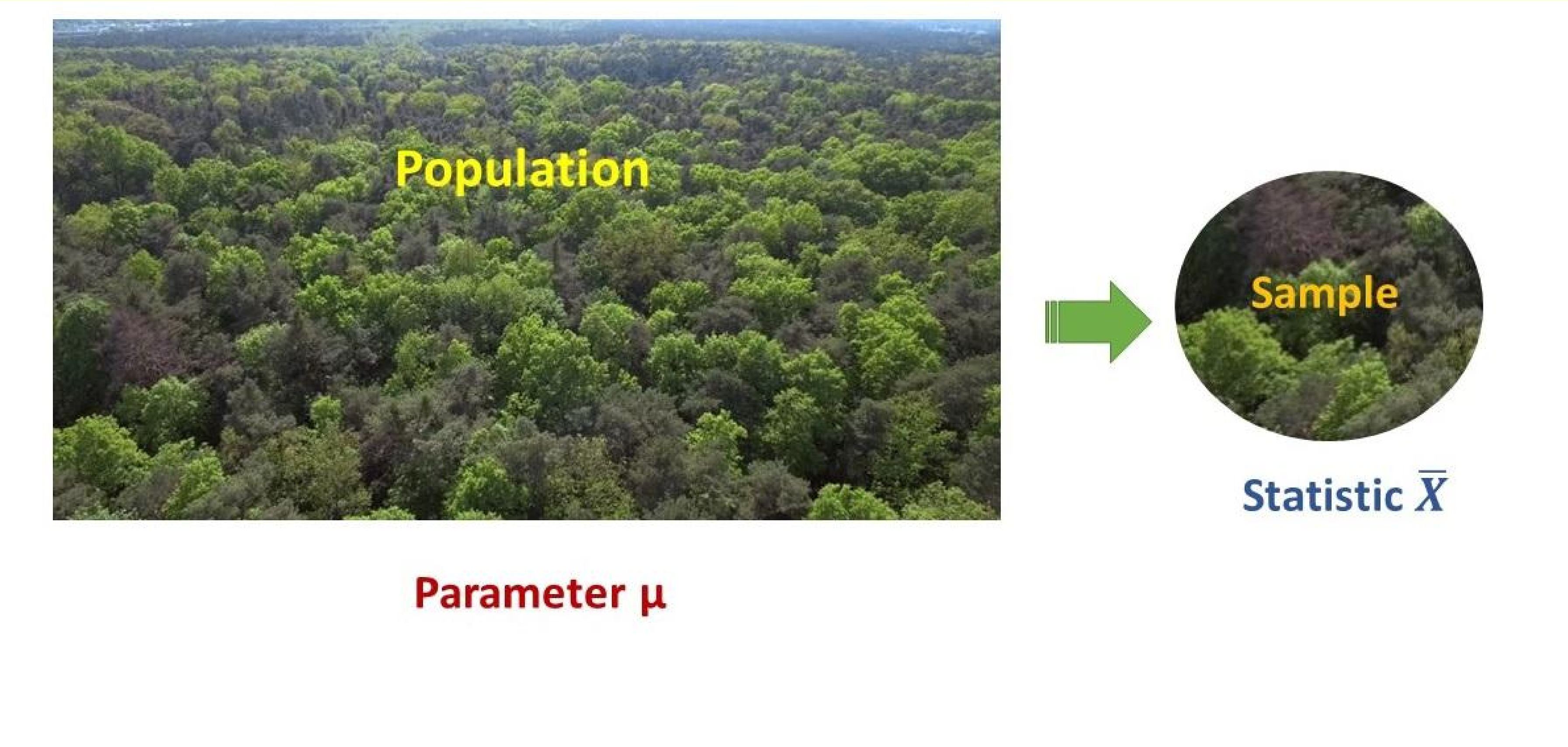
- *Why? Need statistics for this?*



The Probability basis of Statistics

- **Q**: Trying to find *parameter* $\mu = E[X]$ for the population

  Ans: Use statistic $\overline{X}$ (or p) from the sample data



The Probability basis of Statistics

- **Q**: Trying to find *parameter* $\mu = E[X]$ for the population

  Ans: Use statistic $\overline{X}$ (or p) from the sample data

- *Why is this a good estimator?*

The Probability basis of Statistics

- **Q**: Trying to find *parameter* $\mu = E[X]$ for the population

  Ans: Use statistic $\overline{X}$ (or p) from the sample data

- *Why is this a good estimator?*

- **Law of Large Numbers**: As $n \longrightarrow \infty$, $\overline{X}$ approaches μ with probability 1

The Probability basis of Statistics

- **Q**: Trying to find *parameter* $\mu = E[X]$ for the population

  Ans: Use statistic $\overline{X}$ (or p) from the sample data

- *Why is this a good estimator?*

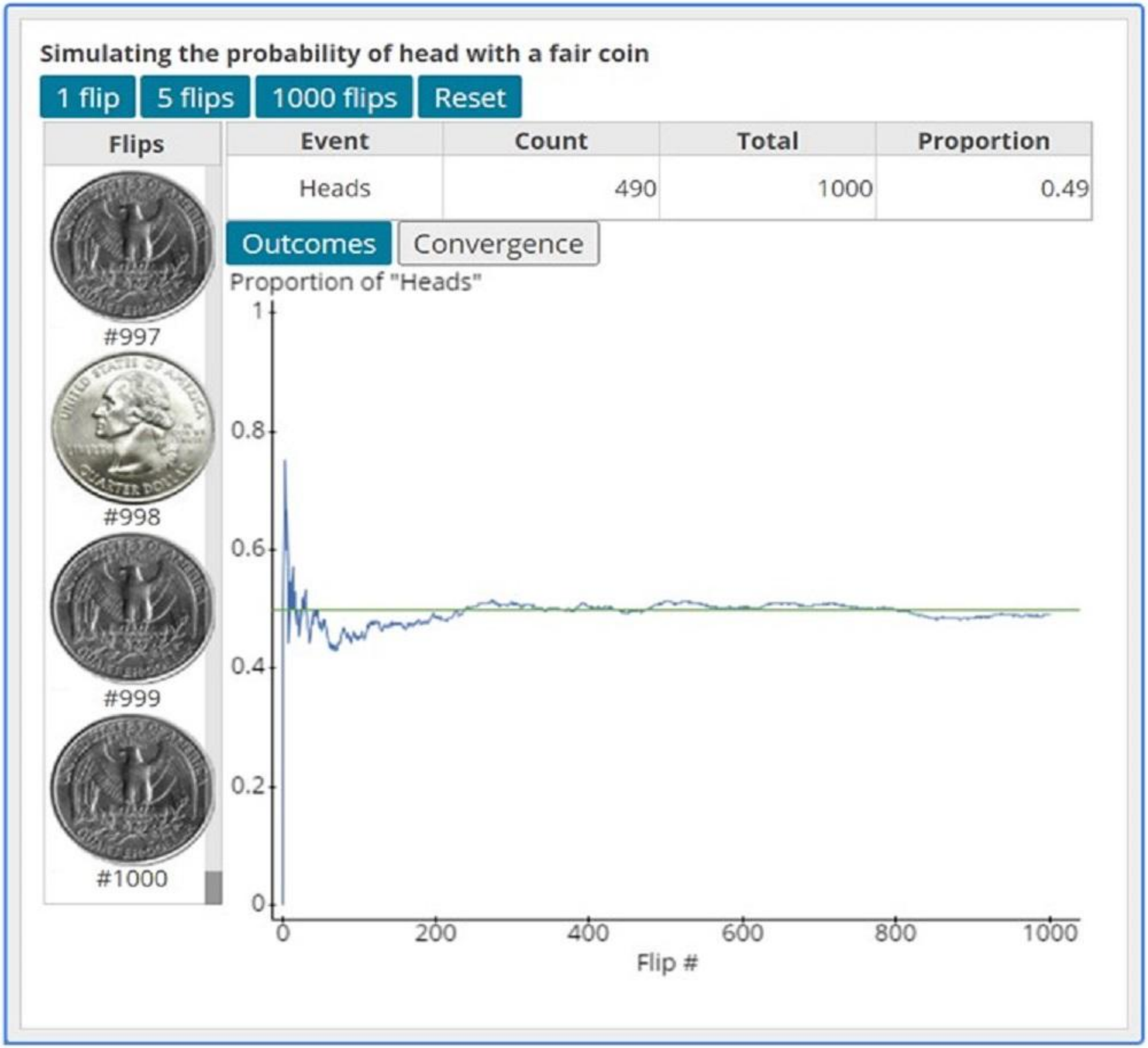
- **Law of Large Numbers**: As $n \longrightarrow \infty$, $\overline{X}$ approaches μ with probability 1

- As the size of the sample gets bigger and bigger, there is more and more chance that your estimate $\overline{X}$ is close to the population parameter.

Interpretation of probability

- *Frequentist*:

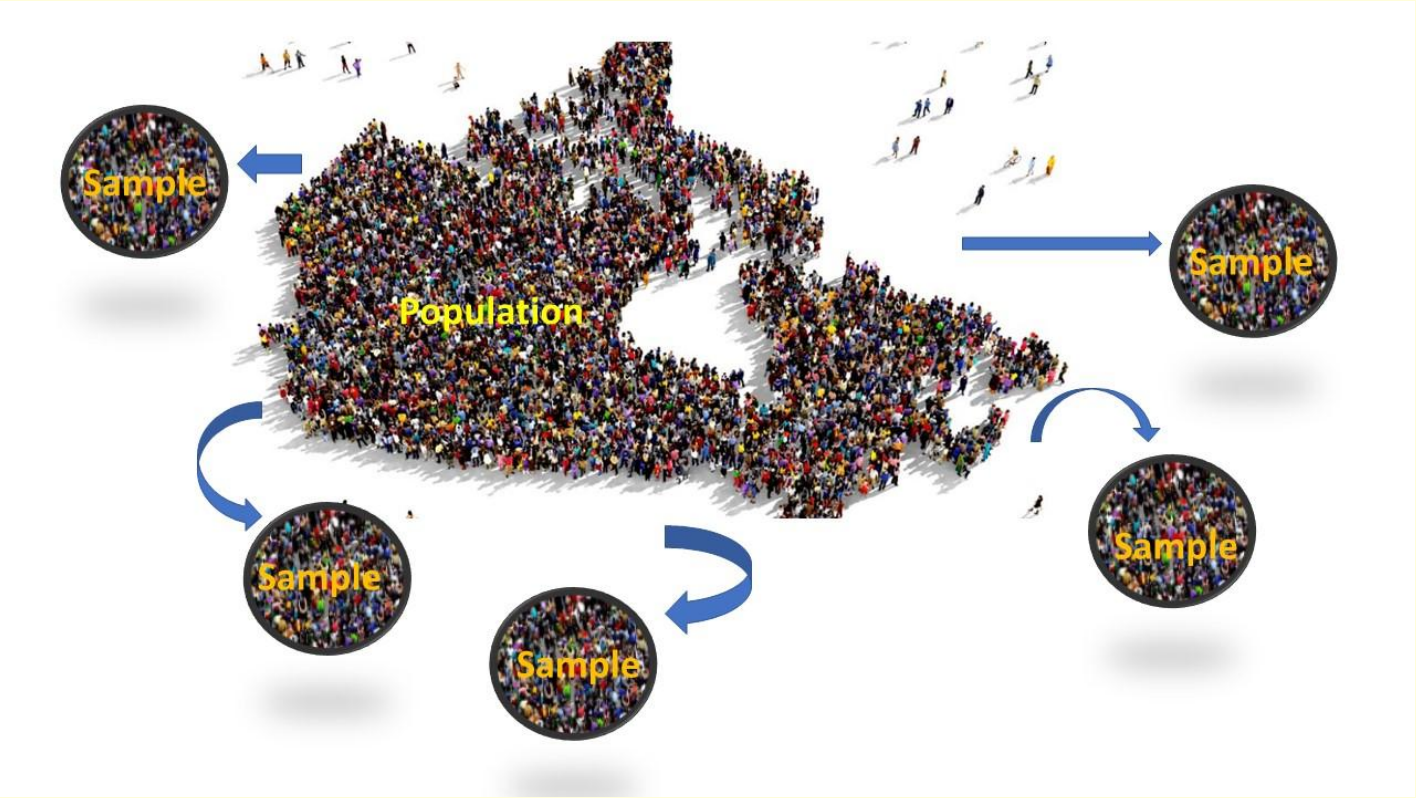
  If you repeat the experiment many, many times, the fraction will go to probability


Simulating the probability of head with a fair coin
1 flip
5 flips
1000 flips
Reset
Flips
Event
Count
Total
Proportion
Heads
490
1000
0.49
Outcomes
Convergence
Proportion of "Heads"
1
0.8
0.6
0.4
0.2
0
#997
#998
#999
#1000
0
200
400
600
800
1000
Flip #

The Probability basis of Statistics

- $\overline{X}$: from the sample data

  For different samples, $\overline{X}$ will be different



The Probability basis of Statistics

- $\overline{X}$: from the sample data

  For different samples, $\overline{X}$ will be different

- **Q**: *How much confidence do you have in the $\overline{X}$ measured from your sample?*

The Probability basis of Statistics

- $\overline{X}$: from the sample data

  For different samples, $\overline{X}$ will be different

- **Q**: *How much confidence do you have in the* $\overline{X}$ *measured from your sample?*

- Ans: **Central Limit Theorem**

The Probability basis of Statistics

- Population: μ (unknown)   Sample (size n): $\overline{X}$

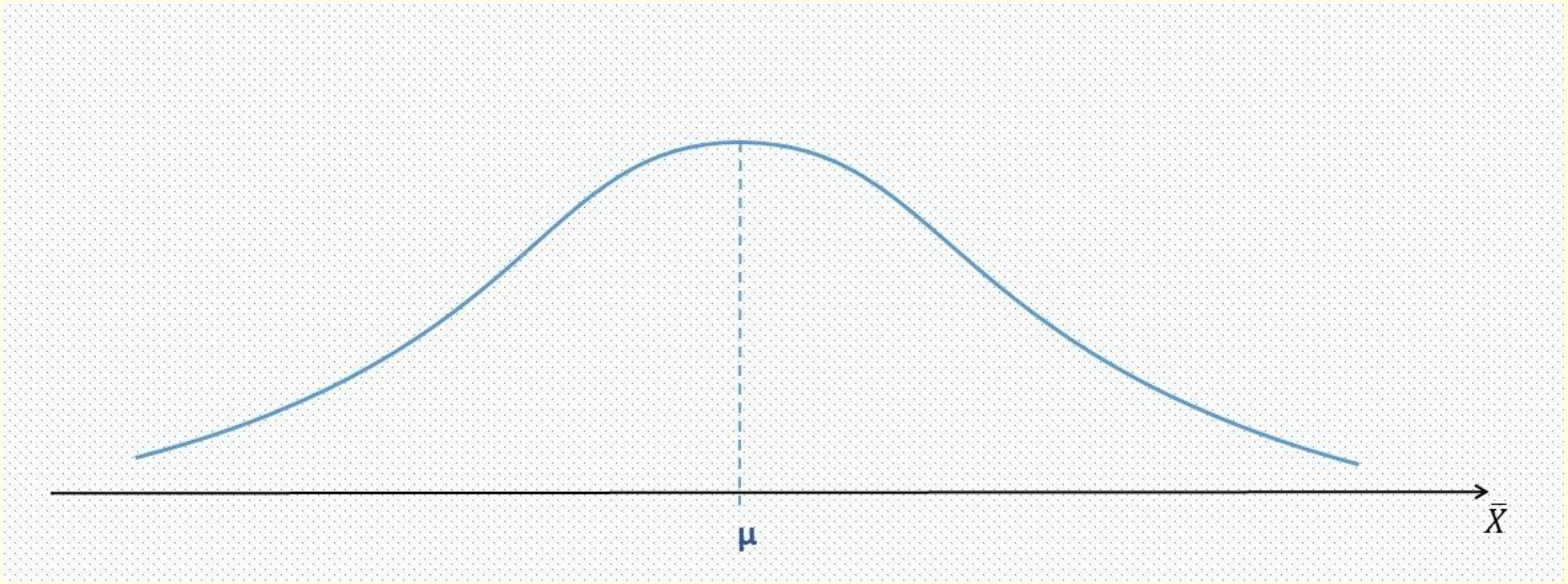
The Probability basis of Statistics

- Population: μ (unknown) Sample (size n): $\overline{X}$
- **Central Limit Theorem**: As $n \longrightarrow \infty$, $\overline{X}$ will be normally distributed around μ with variance $\frac{\sigma^2}{n}$

  i.e. $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$ where $\sigma^2 =$ variance of the population

Central Limit Theorem: Implications

- **Central Limit Theorem**: As $n \longrightarrow \infty$, $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$



Central Limit Theorem: Implications

- **Central Limit Theorem**: As $n \longrightarrow \infty$, $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$
- Typically, $n \geq 30$ enough to apply *CLT*

Central Limit Theorem: Implications

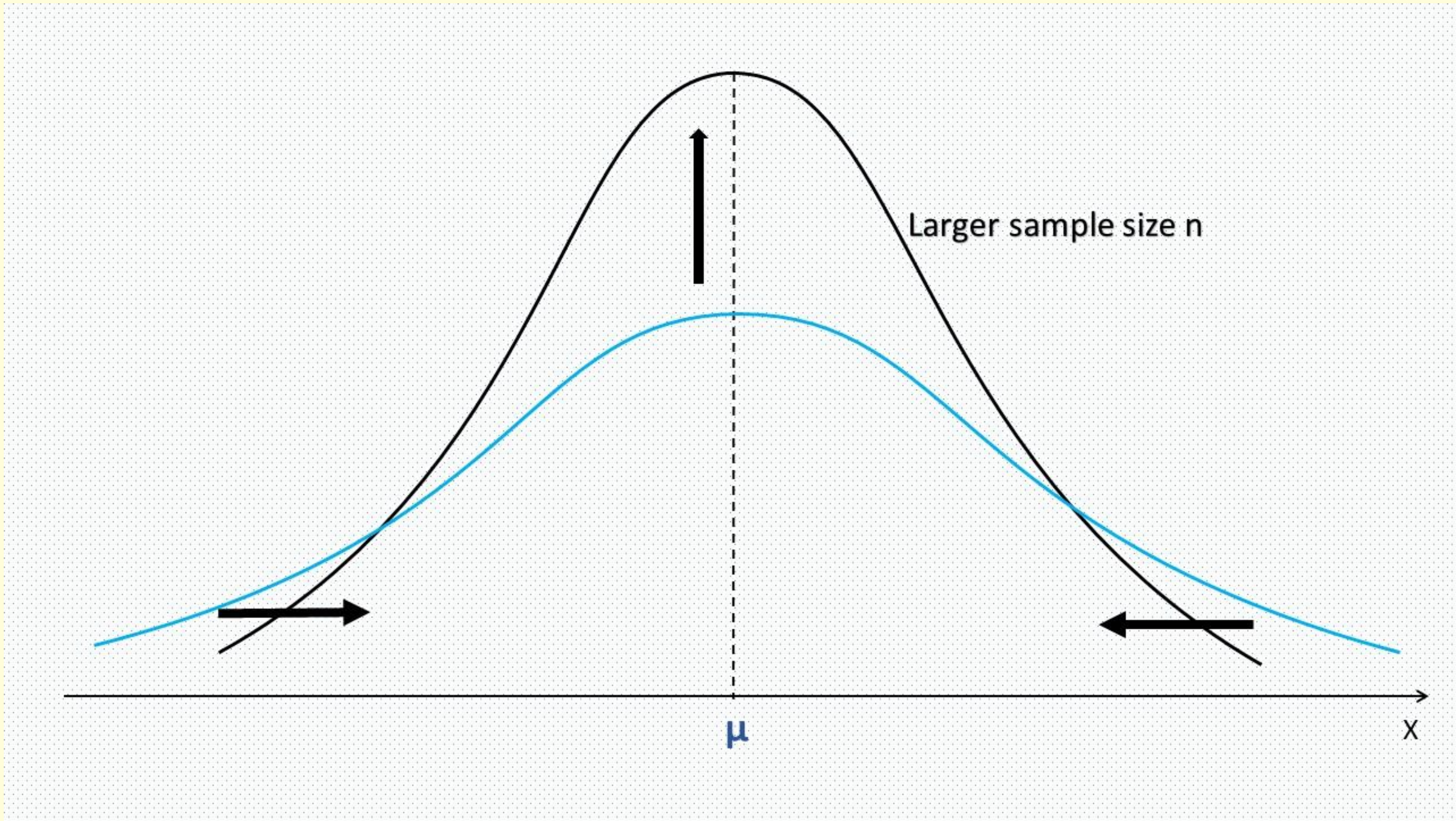
- **Central Limit Theorem**: As $n \longrightarrow \infty$, $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$
- Typically, $n \geq 30$ enough to apply CLT
- Variance $= \frac{\sigma^2}{n}$; standard deviation $= \frac{\sigma}{\sqrt{n}}$

Central Limit Theorem: Implications

- **Central Limit Theorem**: As $n \longrightarrow \infty$, $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$
- Typically, $n \geq 30$ enough to apply CLT
- Variance $= \frac{\sigma^2}{n}$; standard deviation $= \frac{\sigma}{\sqrt{n}}$
- Effect of larger n : reduce $\frac{\sigma}{\sqrt{n}}$

Estimate and Standard Error

- **Central Limit Theorem**: As $n \longrightarrow \infty$, $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$



Estimate and Standard Error

- **Central Limit Theorem**: As $n \longrightarrow \infty$, $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$
- For a quantitative variable X, **estimate of mean** $\mu = \overline{X}$

Estimate and Standard Error

- **Central Limit Theorem**: As $n \longrightarrow \infty$, $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$
- For a quantitative variable X, **estimate of mean** $\mu = \overline{X}$
- Spread given by standard deviation $= \frac{\sigma}{\sqrt{n}}$

Estimate and Standard Error

- **Central Limit Theorem**: As $n \longrightarrow \infty$, $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$
- For a quantitative variable X, **estimate of mean** $\mu = \overline{X}$
- Spread given by standard deviation $= \frac{\sigma}{\sqrt{n}}$
- **Standard error** of estimate $= \frac{s}{\sqrt{n}}$
  where $s =$ sample standard deviation

Estimate and Standard Error

- **Central Limit Theorem**: As $n \longrightarrow \infty$, $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$
- For a quantitative variable X, **estimate of mean** $\mu = \overline{X}$
- Spread given by standard deviation $= \frac{\sigma}{\sqrt{n}}$
- **Standard error** of estimate $= \frac{s}{\sqrt{n}}$
  where $s =$ sample standard deviation
- Recall: $s = \sqrt{\frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \overline{x})^2}$

Example: Quantitative Variable

- How much sleep S an average teenager gets on weeknights?

Example: Quantitative Variable

- How much sleep S an average teenager gets on weeknights?
- *Sample of size 6*:

  Data: 8, 8, 10, 8, 10, 10

Example: Quantitative Variable

- How much sleep S an average teenager gets on weeknights?
- *Sample of size 6*:

  Data: 8, 8, 10, 8, 10, 10
- *Estimate*: $\overline{S} =$

Example: Quantitative Variable

- How much sleep S an average teenager gets on weeknights?
- *Sample of size 6*:

  Data: 8, 8, 10, 8, 10, 10
- *Estimate*: $\overline{S} =$
- *Standard error:* $\frac{s}{\sqrt{6}} =$

  $$s = \sqrt{\frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

Estimate and Standard Error

- For a categorical variable $X \in \{0, 1\}$, interested in estimating the fraction of the population with $X = 1$.

Estimate and Standard Error

- For a categorical variable $X \in \{0, 1\}$, interested in estimating the fraction of the population with $X = 1$.

- **Estimate of proportion** $= p$: fraction of sample with $X = 1$

Estimate and Standard Error

- For a categorical variable $X \in \{0, 1\}$, interested in estimating the fraction of the population with $X = 1$.
- **Estimate of proportion** $= p$: fraction of sample with $X = 1$
- **Standard error** of estimate $= \frac{s}{\sqrt{n}}$
  where $s = \sqrt{p(1-p)}$

Example: Underlying Variable is Categorical

- What fraction support policy A?

Example: Underlying Variable is Categorical

- What fraction support policy A?
- *Sample of size 10*:

  Data: Y, Y, N, Y, N, N, Y, N, N, N

Example: Underlying Variable is Categorical

- What fraction support policy A?
- *Sample of size 10*:

  Data: Y, Y, N, Y, N, N, Y, N, N, N
- *Estimate*: $p =$

Example: Underlying Variable is Categorical

- What fraction support policy A?
- *Sample of size 10*:

  Data: Y, Y, N, Y, N, N, Y, N, N, N
- *Estimate*: $p =$
- *Standard error:* $\frac{\sqrt{p(1-p)}}{\sqrt{10}} =$

Using EXCEL for Estimate and Standard Error

- Quantitative Variable: $\overline{X}$, $\frac{s}{\sqrt{n}}$

  AVERAGE, *STDEV*

Using EXCEL for Estimate and Standard Error

- Quantitative Variable: $\overline{X}$, $\frac{s}{\sqrt{n}}$

  AVERAGE, *STDEV*

- Proportion: p, $\frac{\sqrt{p(1-p)}}{\sqrt{n}}$

  COUNTIF

Confidence Interval

- *So far*:

  Estimate: $\overline{X}$ (or p), **Standard error** (std. devn.): $\frac{s}{\sqrt{n}}$

  CLT: As $n \longrightarrow \infty$, $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$

Confidence Interval

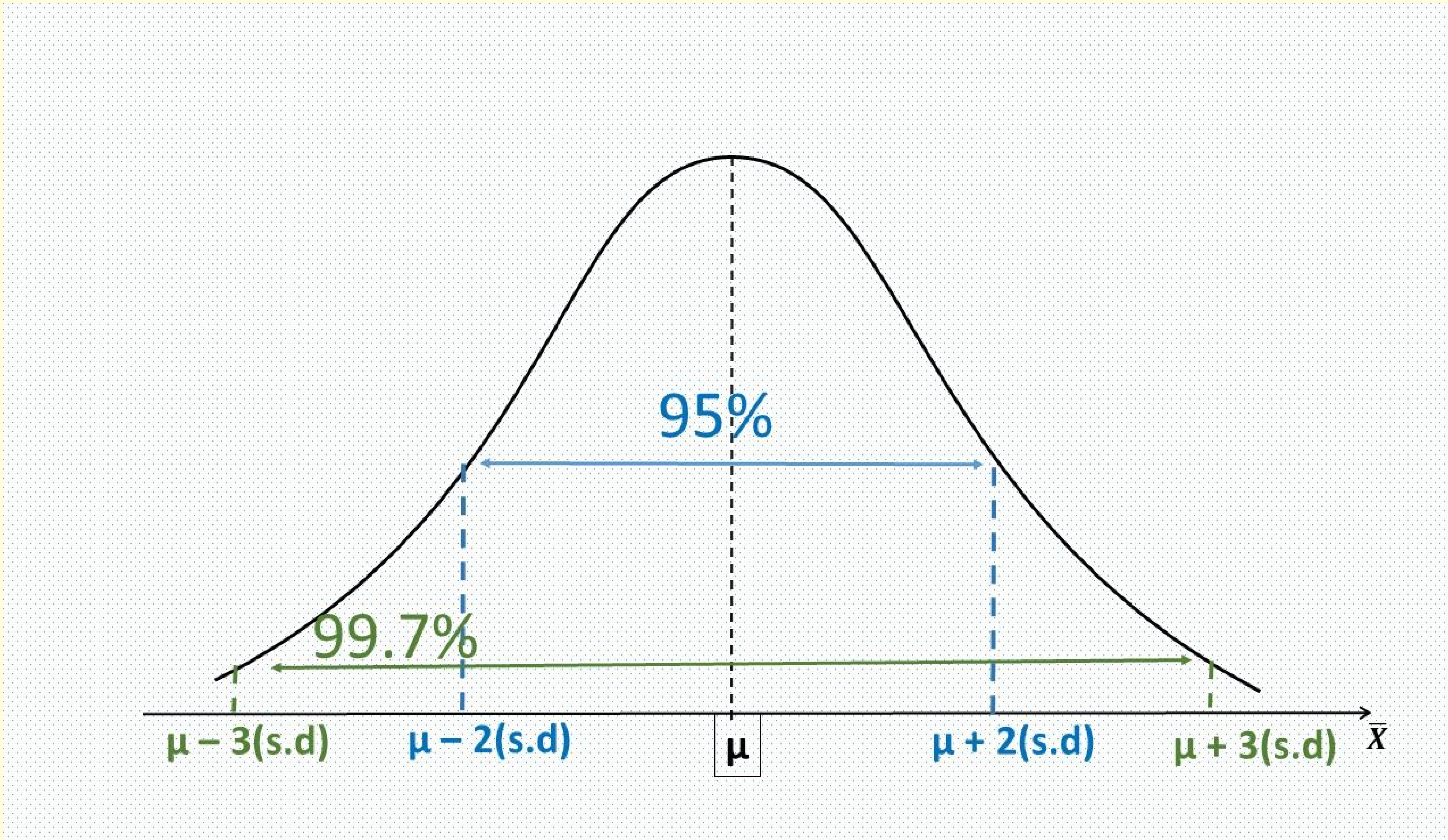
- *So far*:

  Estimate: $\overline{X}$ (or p), **Standard error** (std. devn.): $\frac{s}{\sqrt{n}}$

  CLT: As $n \longrightarrow \infty$, $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$

- *Question:* How much confidence do we have in our estimate?

Confidence Interval



Confidence Interval

-
- Prob$[\ \mu - 2(s.d) \leq \overline{X} \leq \mu + 2(s.d)] = 0.95$

Confidence Interval

-

- Prob[$\mu - 2(s.d) \leq \overline{X} \leq \mu + 2(s.d)$] $= 0.95$

- *Implies*:

  95% confidence interval for μ: $[\overline{X} - \frac{2s}{\sqrt{n}}, \overline{X} + \frac{2s}{\sqrt{n}}]$

Confidence Interval

-
- Prob[$\mu - 2(s.d) \le \overline{X} \le \mu + 2(s.d)] = 0.95$
- *Implies*:

  95% confidence interval for μ: $[\overline{X} - \frac{2s}{\sqrt{n}}, \overline{X} + \frac{2s}{\sqrt{n}}]$
- **Point estimate** *of* $\mu : \overline{X}$

  Confidence interval: estimate $\pm \frac{2s}{\sqrt{n}}$

Confidence Interval

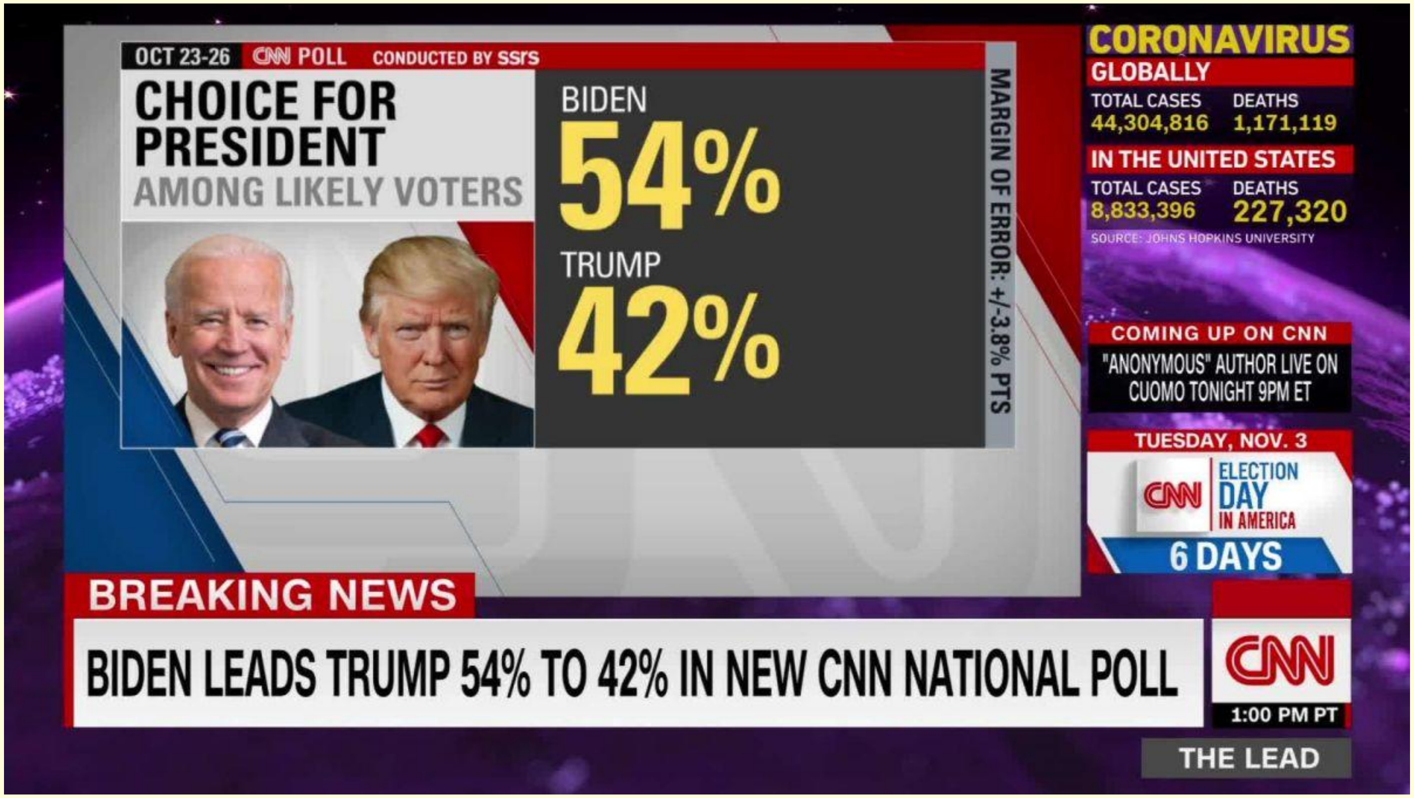
-
- Prob[$\mu - 2(s.d) \leq \overline{X} \leq \mu + 2(s.d)] = 0.95$
- *Implies*:

  95% confidence interval for μ: $[\overline{X} - \frac{2s}{\sqrt{n}}, \overline{X} + \frac{2s}{\sqrt{n}}]$
- **Point estimate** *of* $\mu : \overline{X}$

  Confidence interval: estimate $\pm \frac{2s}{\sqrt{n}}$
- **Margin of error**: $\frac{2s}{\sqrt{n}}$

Confidence Interval Real Examples

- **2020** US Presidential election CNN poll of 886 voters (Oct): **Biden**: 54%, **Trump**: 42%, *Margin of error*: $\pm 3.6\%$



Confidence Interval Real Examples

- **2020** US Presidential election CNN poll of 886 voters (Oct):
  Biden: 54%, **Trump**: 42%, *Margin of error*: $\pm 3.6\%$
- **2016** US election Bloomberg poll of 799 voters (Nov):
  Clinton: 46%, **Trump**: 43%, *Margin of error*: $\pm 3.5\%$

Confidence Interval Real Examples

- **2020** US Presidential election CNN poll of 886 voters (Oct):
  Biden: 54%, **Trump**: 42%, *Margin of error*: $\pm 3.6\%$
- **2016** US election Bloomberg poll of 799 voters (Nov):
  Clinton: 46%, **Trump**: 43%, *Margin of error*: $\pm 3.5\%$
- **2021** Cdn. Federal election polls (338canada.com):
  Projected number of seats:
  Liberals: $[115, 186]$, *Cons.*: $[90, 148]$, *NDP*: $[19, 54]$

Confidence Interval Real Examples

- **2020** US Presidential election CNN poll of 886 voters (Oct):

  Biden: 54%, **Trump**: 42%, *Margin of error*: $\pm 3.6\%$

- **2016** US election Bloomberg poll of 799 voters (Nov):

  Clinton: 46%, **Trump**: 43%, *Margin of error*: $\pm 3.5\%$

- **2021** Cdn. Federal election polls (338canada.com):

  Projected number of seats:

  Liberals: $[115, 186]$, *Cons.*: $[90, 148]$, *NDP*: $[19, 54]$

- **Final outcome:** *Liberals*: 160, *Cons.*: 119, *NDP*: 25

Confidence interval: Example with sleep data

- **Data**: 8, 8, 10, 8, 10, 10

Confidence interval: Example with sleep data

- **Data**: 8, 8, 10, 8, 10, 10
- *Point estimate*: $\overline{S} = 9$; *standard error:* $\frac{s}{\sqrt{6}} = 0.45$

Confidence interval: Example with sleep data

- **Data**: 8, 8, 10, 8, 10, 10
- *Point estimate*: $\overline{S} = 9$; *standard error:* $\frac{s}{\sqrt{6}} = 0.45$
- *95% confidence interval*: $[\overline{X} - \frac{2s}{\sqrt{n}}, \overline{X} + \frac{2s}{\sqrt{n}}] \equiv$

Confidence interval: Example with sleep data

- **Data**: 8, 8, 10, 8, 10, 10
- *Point estimate*: $\overline{S} = 9$; *standard error:* $\frac{s}{\sqrt{6}} = 0.45$
- *95% confidence interval*: $[\overline{X} - \frac{2s}{\sqrt{n}}, \overline{X} + \frac{2s}{\sqrt{n}}] \equiv$
- Warning!

Confidence Interval: Properties

- **Point estimate** *of* $\mu : \overline{X}$

  95% confidence interval: $[\overline{X} - \frac{2s}{\sqrt{n}}, \overline{X} + \frac{2s}{\sqrt{n}}]$

Confidence Interval: Properties

- **Point estimate** *of* $\mu : \overline{X}$

  95% confidence interval: $[\overline{X} - \frac{2s}{\sqrt{n}}, \overline{X} + \frac{2s}{\sqrt{n}}]$

- More variability: $s \uparrow \Rightarrow$ wider CI

Confidence Interval: Properties

- **Point estimate** *of* $\mu : \overline{X}$

  95% confidence interval: $[\overline{X} - \frac{2s}{\sqrt{n}}, \overline{X} + \frac{2s}{\sqrt{n}}]$

- More variability: $s \uparrow \Rightarrow$ wider CI

- As $n \uparrow \Rightarrow$ narrower CI

Confidence Interval when underlying variable is categorical

- **Point estimate** *of* fraction: p

  95% confidence interval: $[p - \frac{2s}{\sqrt{n}}, p + \frac{2s}{\sqrt{n}}]$

  where $s = \sqrt{p(1-p)}$

CI: Example with support for policy A

- **Data**: Y, Y, N, Y, N, N, Y, N, N, N

CI: Example with support for policy A

- **Data**: Y, Y, N, Y, N, N, Y, N, N, N
- *Point estimate*: $p = 0.4$, *Standard error:* $\frac{\sqrt{p(1-p)}}{\sqrt{10}} = 0.155$

CI: Example with support for policy A

- **Data**: Y, Y, N, Y, N, N, Y, N, N, N
- *Point estimate*: $p = 0.4$, *Standard error:* $\frac{\sqrt{p(1-p)}}{\sqrt{10}} = 0.155$
- *95% confidence interval*: $[p - \frac{2s}{\sqrt{n}}, p + \frac{2s}{\sqrt{n}}] \equiv$

Interpretation of Confidence Interval

- *Example*: Support for policy A among Canadians

  Point estimate: $p = 0.4$, *95% confidence interval*: $[0.09, 0.71]$

Interpretation of Confidence Interval

- *Example*: Support for policy A among Canadians

  Point estimate: $p = 0.4$, *95% confidence interval*: $[0.09, 0.71]$

- What does it mean?

  "40% of Canadians support policy A"

Interpretation of Confidence Interval

- *Example*: Support for policy A among Canadians

  Point estimate: $p = 0.4$, *95% confidence interval*: $[0.09, 0.71]$

- What does it mean?

  "40% of Canadians support policy A"

- **"It is probably true that 40% of Canadians support policy A"**

Interpretation of Confidence Interval

- *Example*: Support for policy A among Canadians

  Point estimate: $p = 0.4$, *95% confidence interval*: $[0.09, 0.71]$

- What does it mean?

  "40% of Canadians support policy A"

- **"It is probably true that 40% of Canadians support policy A"**

- **"We don't exactly know what fraction of Canadians support policy A, but we know it is between 9% and 71%."**

Interpretation of Confidence Interval

- *Example*: Support for policy A among Canadians

  Point estimate: $p = 0.4$, *95% confidence interval*: $[0.09, 0.71]$
- What does it mean?

  "40% of Canadians support policy A"
- **"It is probably true that 40% of Canadians support policy A"**
- **"We don't exactly know what fraction of Canadians support policy A, but we know it is between 9% and 71%."**
- Close, but No!

Interpretation of Confidence Interval

- *Example*: Support for policy A among Canadians

  Point estimate: $p = 0.4$, *95% confidence interval*: $[0.09, 0.71]$

- **"We don't exactly know what fraction of Canadians support policy A, but we know that fraction is probably between 9% and 71%."**

Interpretation of Confidence Interval

- *Example*: Support for policy A among Canadians

  Point estimate: $p = 0.4$, *95% confidence interval*: $[0.09, 0.71]$

- **"We don't exactly know what fraction of Canadians support policy A, but we know that fraction is probably between 9% and 71%."**

- Very close

Interpretation of Confidence Interval

- *Example*: Support for policy A among Canadians

  Point estimate: $p = 0.4$, *95% confidence interval*: $[0.09, 0.71]$

- **"We don't exactly know what fraction of Canadians support policy A, but we know that fraction is probably between 9% and 71%."**

- Very close

- **"We don't exactly know what fraction of Canadians support policy A, but are 95% confident that it is between 9% and 71%."**

Interpretation of Confidence Interval

- *Example*: Mean Weight of Americans

  Source: 2015-16 NCHS Survey

Interpretation of Confidence Interval

- *Example*: Mean Weight of Americans

  Source: 2015-16 NCHS Survey

- **Women**: *Point estimate*: 77.3 *kg*, *95% CI*: $[75.7, 78.9]$

  Sample size: 2760

Interpretation of Confidence Interval

- *Example*: Mean Weight of Americans

  Source: 2015-16 NCHS Survey

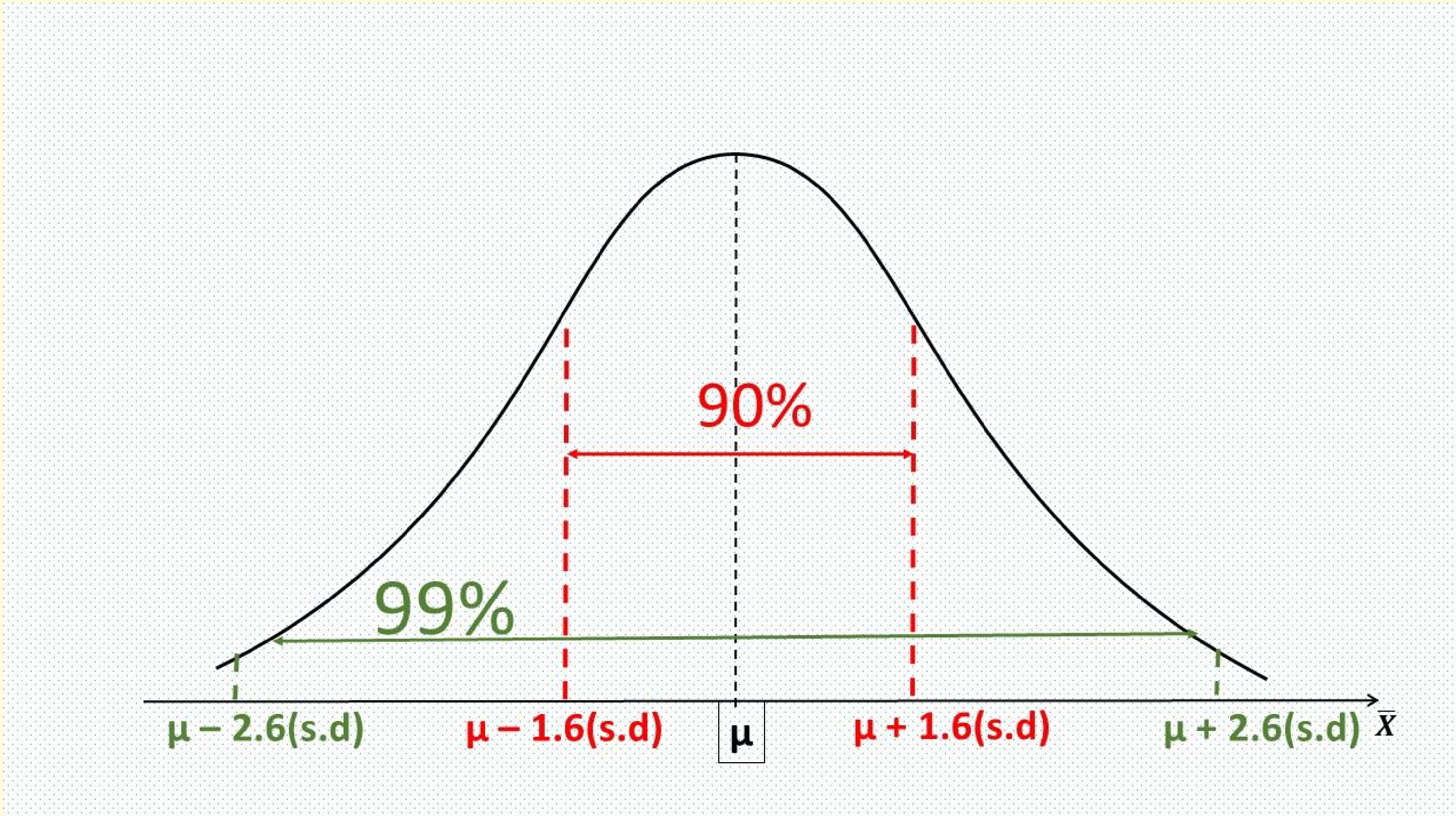
- **Women**: *Point estimate*: 77.3 *kg*, *95% CI*: $[75.7, 78.9]$

  Sample size: 2760

- **Men**: *Point estimate*: 89.7 *kg*, *95% CI*: $[87.9, 91.5]$

  Sample size: 2591

Other levels of confidence



Other levels of confidence

- Normal pdf
- $\text{Prob}[\ \mu - 2.6(s.d) \leq \overline{X} \leq \mu + 2.6(s.d)] = 0.99$

Other levels of confidence

- Normal pdf
- Prob[$\mu - 2.6(s.d) \leq \overline{X} \leq \mu + 2.6(s.d)] = 0.99$
- *Implies*:

  99% confidence interval: $[\overline{X} - \frac{2.6s}{\sqrt{n}}, \overline{X} + \frac{2.6s}{\sqrt{n}}]$

Other levels of confidence

- Normal pdf
- Prob[$\mu - 2.6(s.d) \leq \overline{X} \leq \mu + 2.6(s.d)] = 0.99$
- *Implies*:

  99% confidence interval: $[\overline{X} - \frac{2.6s}{\sqrt{n}}, \overline{X} + \frac{2.6s}{\sqrt{n}}]$
- Wider interval than 95%

Other levels of confidence

- **99% confidence interval**: $[\overline{X} - \frac{2.6s}{\sqrt{n}}, \overline{X} + \frac{2.6s}{\sqrt{n}}]$

Other levels of confidence

- **99% confidence interval**: $[\overline{X} - \frac{2.6s}{\sqrt{n}}, \overline{X} + \frac{2.6s}{\sqrt{n}}]$
- **90% confidence interval**: $[\overline{X} - \frac{1.6s}{\sqrt{n}}, \overline{X} + \frac{1.6s}{\sqrt{n}}]$

  Narrower than 95%

Other levels of confidence

- **99% confidence interval**: $[\overline{X} - \frac{2.6s}{\sqrt{n}}, \overline{X} + \frac{2.6s}{\sqrt{n}}]$
- **90% confidence interval**: $[\overline{X} - \frac{1.6s}{\sqrt{n}}, \overline{X} + \frac{1.6s}{\sqrt{n}}]$

  Narrower than 95%

- If sample size $n < 30$, use the *t-distribution* [not covered here]

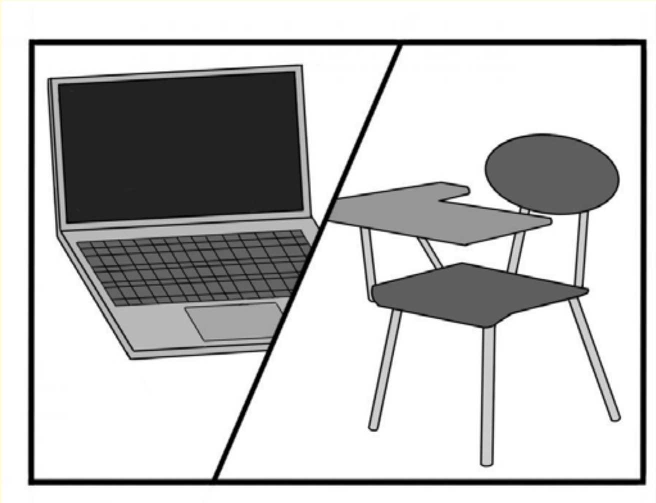
Hypothesis Testing

- *Hypothesis / Conjecture:* Can we use data to test them?

Hypothesis Testing

- *Hypothesis / Conjecture:* Can we use data to test them?
- *Examples*:
  Online learning has had an impact on student learning

  Policy A (free laptops, subsidized child-care, raise in minimum wage) will have a positive/negative impact on outcome B (education, children and parents, number of jobs)



Hypothesis Testing

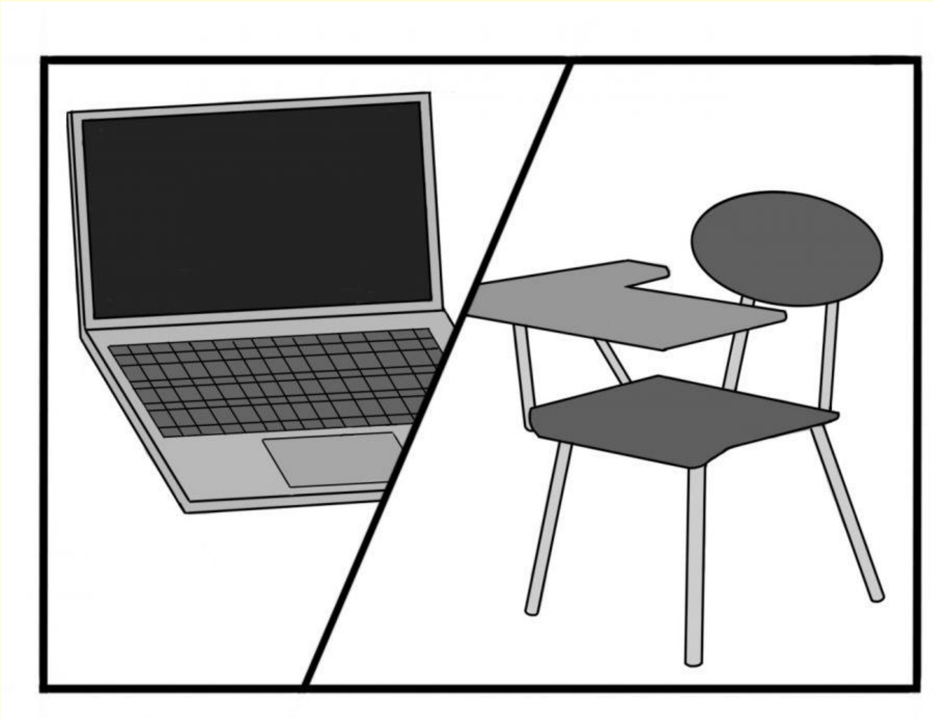
- *Hypothesis / Conjecture:* Can we use data to test them?
- *Examples*:
  Online learning has had an impact on student learning

  Policy A (free laptops, subsidized child-care, raise in minimum wage) will have a positive/negative impact on outcome B (education, children and parents, number of jobs)
- A country's CO_2 emissions are strongly related to its GDP

  There is a strong connection between a city's crime-rate and its unemployment rate

Null and Alternate Hypothesis

- *Hypothesis:* Online learning has had an impact on student learning



Null and Alternate Hypothesis

- *Hypothesis:* Online learning has had an impact on student learning
- Data? Suppose you have test scores of n students before and after online learning

Null and Alternate Hypothesis

- *Hypothesis:* Online learning has had an impact on student learning
- Data? Suppose you have test scores of n students before and after online learning
- $X = S_{\text{online}} - S_{\text{in-person}} \qquad E[X] = \mu?$

Null and Alternate Hypothesis

- *Hypothesis:* Online learning has had an impact on student learning
- Data? Suppose you have test scores of n students before and after online learning
- $X = S_{\text{online}} - S_{\text{in-person}} \qquad E[X] = \mu?$
- **Null hypothesis** $H_0 : \mu = 0$ (skeptical view)

Null and Alternate Hypothesis

- *Hypothesis:* Online learning has had an impact on student learning
- Data? Suppose you have test scores of n students before and after online learning
- $X = S_{\text{online}} - S_{\text{in-person}} \qquad E[X] = \mu?$
- **Null hypothesis** $H_0 : \mu = 0$ (skeptical view)
- **Alternate hypothesis** $H_A : \mu \neq 0$

Hypothesis Testing: Basic idea

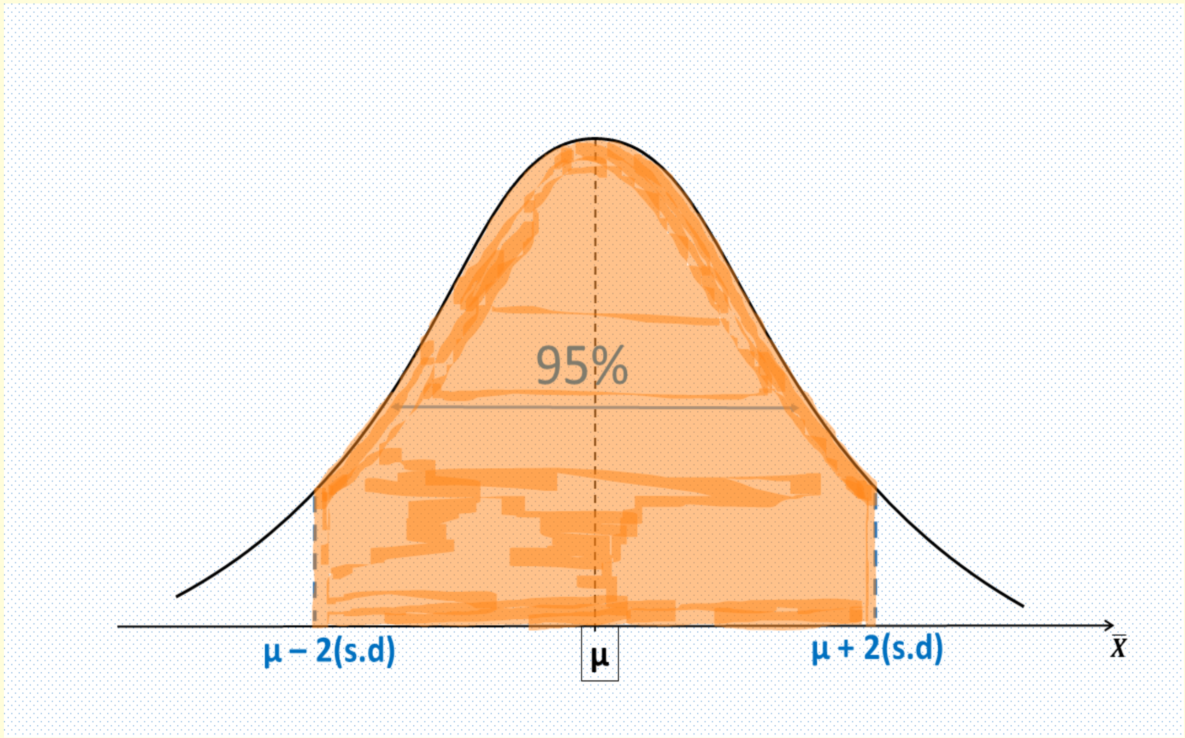
- How to test $H_0 : \mu = 0$?

Hypothesis Testing: Basic idea

- How to test $H_0 : \mu = 0$?
- Again **CLT:** $\overline{X} \sim N(\mu, \frac{\sigma^2}{n}) \equiv N(0, \frac{\sigma^2}{n})$ if H_0 is true

Hypothesis Testing: Basic idea

- How to test $H_0 : \mu = 0$?
- Again **CLT:** $\overline{X} \sim N(\mu, \frac{\sigma^2}{n}) \equiv N(0, \frac{\sigma^2}{n})$ if H_0 is true
-



Hypothesis Testing: Basic idea

- How to test $H_0 : \mu = 0$?   **CLT:** $\overline{X} \sim N(0, \frac{\sigma^2}{n})$ if H_0 is true

- <u>If H_0 is true</u>, Prob$[\ \mu - 2(s.d) \leq \overline{X} \leq \mu + 2(s.d)] = 0.95$

Hypothesis Testing: Basic idea

- How to test $H_0 : \mu = 0$? **CLT:** $\overline{X} \sim N(0, \frac{\sigma^2}{n})$ if H_0 is true

- If H_0 is true, Prob[$\mu - 2(s.d) \leq \overline{X} \leq \mu + 2(s.d)] = 0.95$

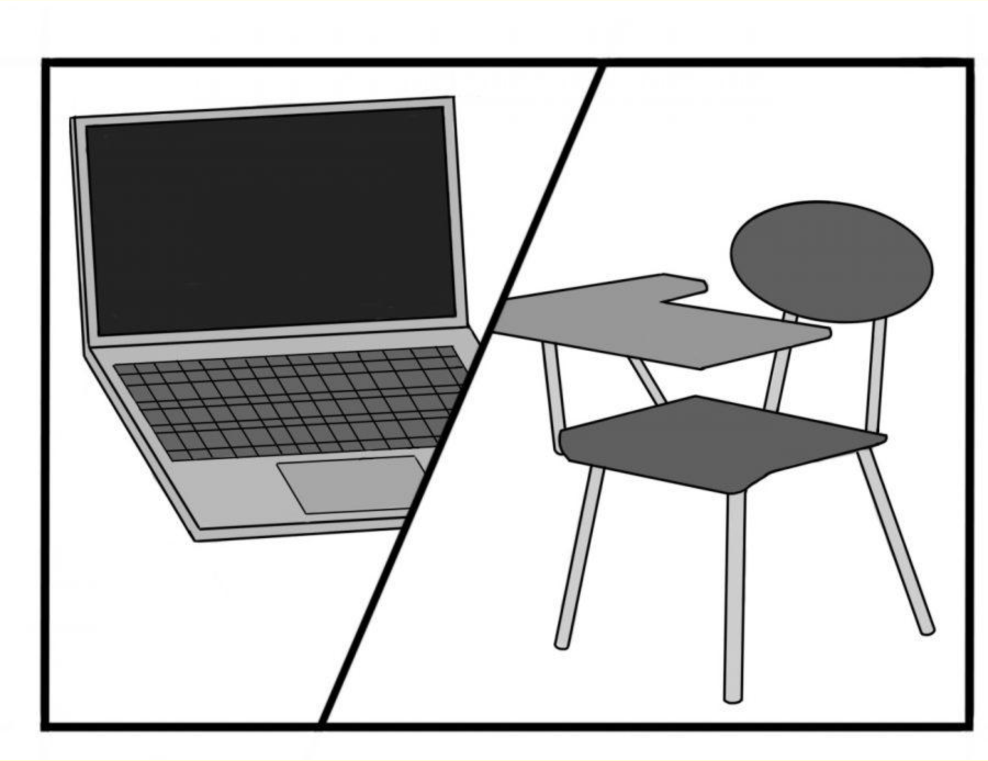
- **Test**:

  If $\overline{X}$ *outside* $[\mu - \frac{2s}{\sqrt{n}}, \mu + \frac{2s}{\sqrt{n}}]$: reject H_0 in favor of H_A

  if not: then cannot reject H_0

Example: Hypothesis testing

- *Online learning*: $n = 100$ students, $X = S_{\text{online}} - S_{\text{in-person}}$



Example: Hypothesis testing

- *Online learning*: $n = 100$ students, $X = S_{\text{online}} - S_{\text{in-person}}$
- *Data:* $\overline{X} = -3$, $s = 2.5$

Example: Hypothesis testing

- *Online learning*: $n = 100$ students, $X = S_{\text{online}} - S_{\text{in-person}}$
- *Data:* $\overline{X} = -3$, $s = 2.5$
- $H_0 : \mu = 0$ vs $H_A : \mu \neq 0$

  <u>If H_0 is true</u>, $\text{Prob}[\, \mu - \frac{2s}{\sqrt{n}} \leq \overline{X} \leq \mu + \frac{2s}{\sqrt{n}}] = 0.95$

Example: Hypothesis testing

- *Online learning*: $n = 100$ students, $X = S_{\text{online}} - S_{\text{in-person}}$
- *Data:* $\overline{X} = -3$, $s = 2.5$
- $H_0 : \mu = 0$ vs $H_A : \mu \neq 0$

  If H_0 is true, Prob$[\ \mu - \frac{2s}{\sqrt{n}} \leq \overline{X} \leq \mu + \frac{2s}{\sqrt{n}}] = 0.95$
- Here $\overline{X}$ outside $[-0.5, 0.5]$: reject H_0

Mechanics of Hypothesis testing

- Test $H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$

  If H_0 is true, $\overline{X} \sim N(\mu_0, \frac{\sigma^2}{n})$

Mechanics of Hypothesis testing

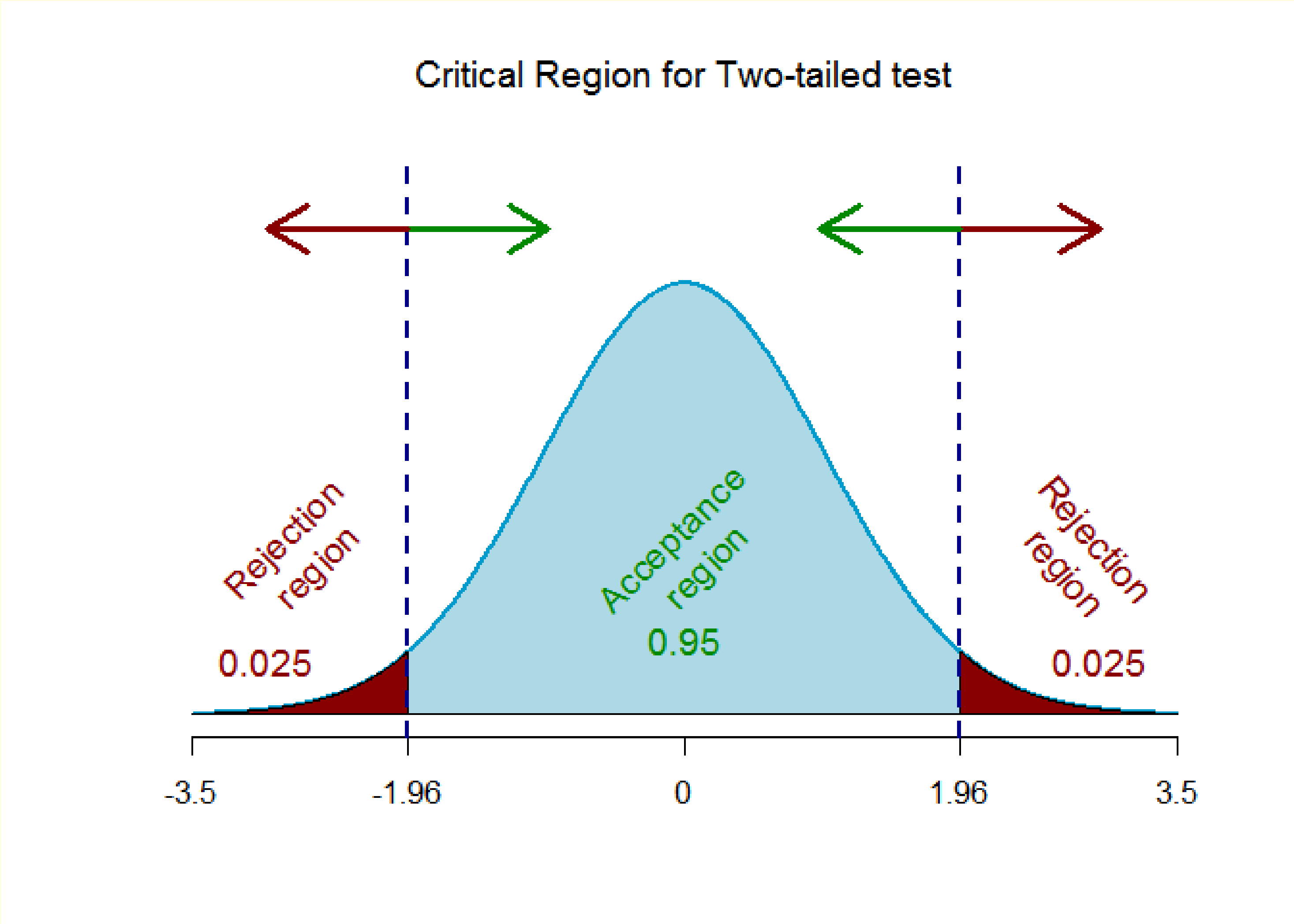
- Test $H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$

  If H_0 is true, $\overline{X} \sim N(\mu_0, \frac{\sigma^2}{n})$

- Property of a normal distribution: $\frac{\overline{X} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0, 1)$

Mechanics of Hypothesis testing

- $N(0,1)$

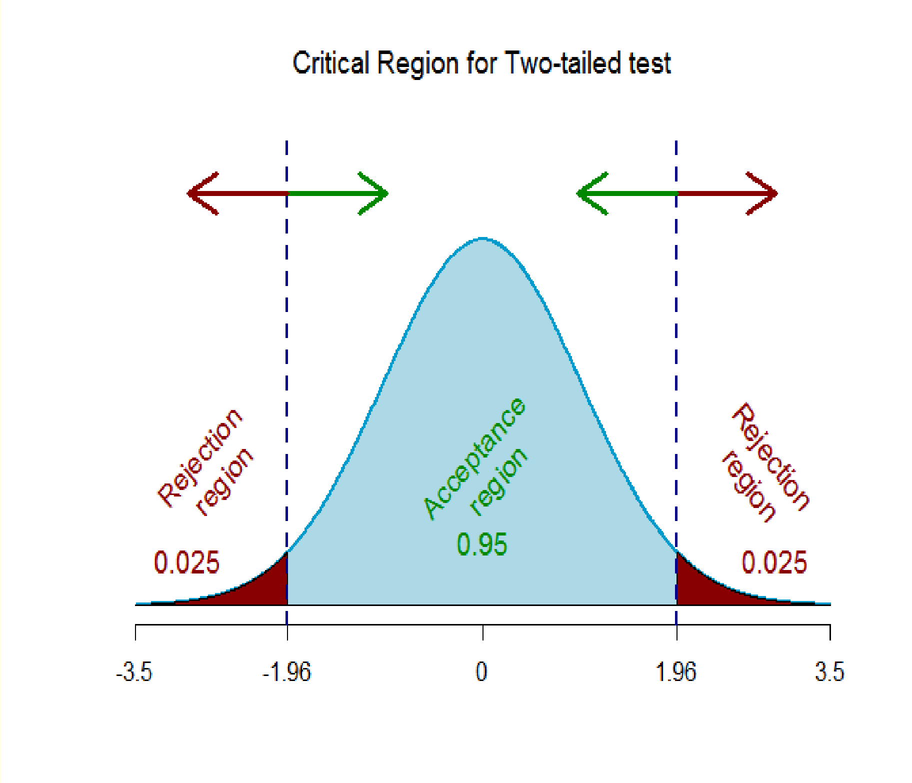


Mechanics of Hypothesis testing

- $N(0,1)$
- $\frac{\overline{X}-\mu_0}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$

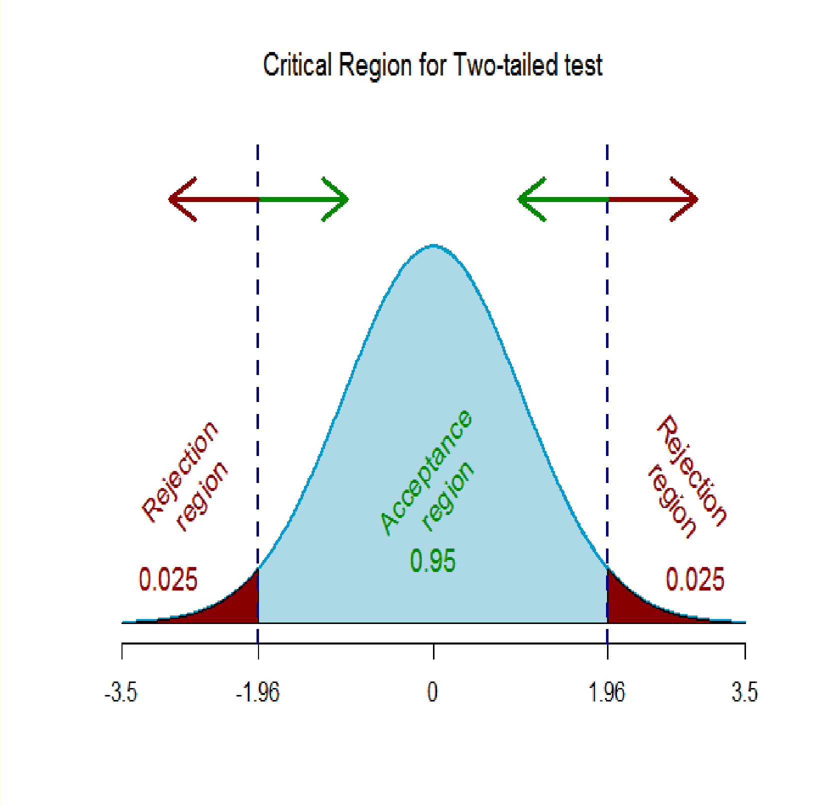
Mechanics of Hypothesis testing

- $N(0,1)$
- $\frac{\overline{X}-\mu_0}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$
- Construct the z-statistic: $\frac{\overline{X}-\mu_0}{\frac{s}{\sqrt{n}}}$



Mechanics of Hypothesis testing

- $N(0,1)$
- $\frac{\overline{X}-\mu_0}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$
- Construct the z-statistic: $\frac{\overline{X}-\mu_0}{\frac{s}{\sqrt{n}}}$
- **Check**: is z outside $[-1.96, 1.96]$?



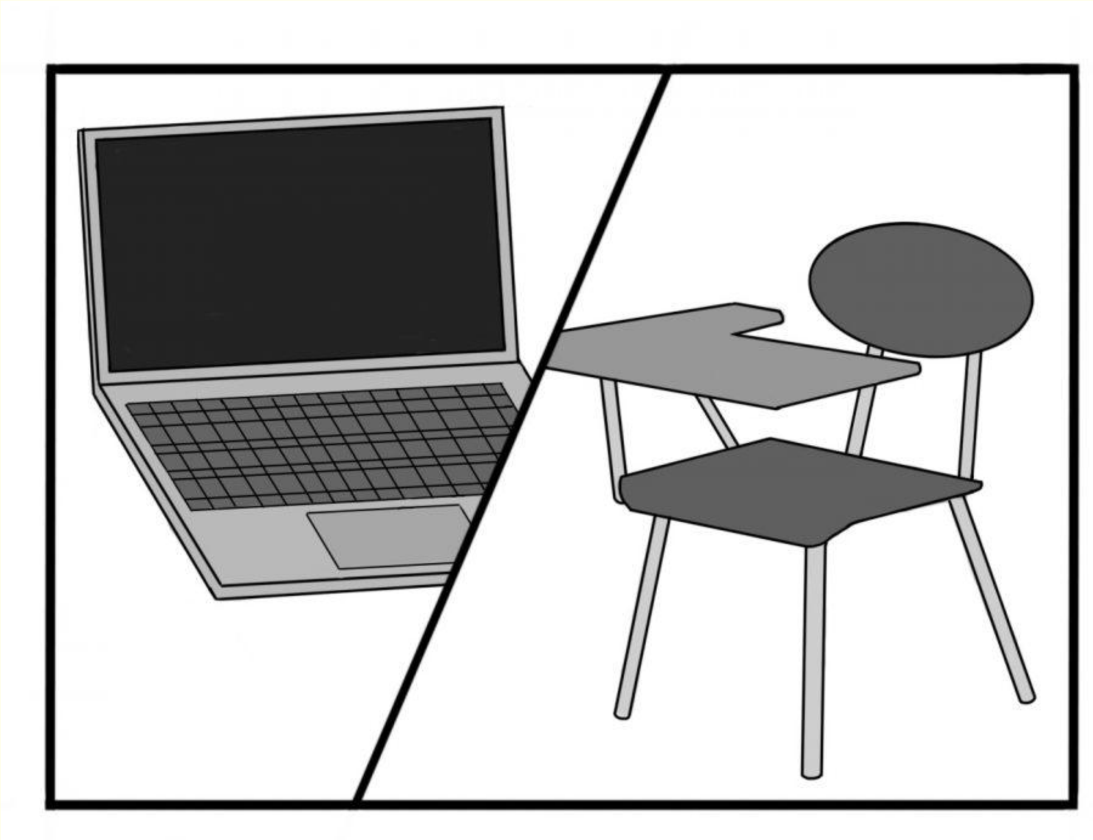
Mechanics of Hypothesis testing

- $N(0,1)$
- $\frac{\overline{X}-\mu_0}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$
- Construct the z-statistic: $\frac{\overline{X}-\mu_0}{\frac{s}{\sqrt{n}}}$
- **Check**: is z outside $[-1.96, 1.96]$?
- *Yes* : reject H_0 in favor of H_A

  if not: then cannot reject H_0

Example: Online learning

- *Online learning*: $n = 100$ students, $X = S_{\text{online}} - S_{\text{in-person}}$



Example: Online learning

- *Online learning*: $n = 100$ students, $X = S_{\text{online}} - S_{\text{in-person}}$
- $n = 100$ students, $X = S_{\text{online}} - S_{\text{in-person}}$, $\overline{X} = -3$, $s = 2.5$

  $H_0 : \mu = 0$ vs $H_A : \mu \neq 0$

Example: Online learning

- *Online learning*: $n = 100$ students, $X = S_{\text{online}} - S_{\text{in-person}}$
- $n = 100$ students, $X = S_{\text{online}} - S_{\text{in-person}}$, $\overline{X} = -3$, $s = 2.5$

  $H_0 : \mu = 0$ vs $H_A : \mu \neq 0$

- $z = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}} =$

Example: Online learning

- *Online learning*: $n = 100$ students, $X = S_{\text{online}} - S_{\text{in-person}}$
- $n = 100$ students, $X = S_{\text{online}} - S_{\text{in-person}}$, $\overline{X} = -3$, $s = 2.5$

  $H_0 : \mu = 0$ vs $H_A : \mu \neq 0$

- $z = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}} =$
- outside $[-1.96, 1.96]$: reject H_0

Example: Hypothesis testing with proportions

- Test hypothesis that **ads have an impact on perceptions about politicians**

Example: Hypothesis testing with proportions

- Test hypothesis that **ads have an impact on perceptions about politicians**
- Current favorable perception: 0.4

Example: Hypothesis testing with proportions

- Test hypothesis that **ads have an impact on perceptions about politicians**
- Current favorable perception: 0.4
- Run ad campaign. Then survey 200 people. Find 90 of them now have favorable perception.

Example: Hypothesis testing with proportions

- Test hypothesis that **ads have an impact on perceptions about politicians**
- Current favorable perception: 0.4
- Run ad campaign. Then survey 200 people. Find 90 of them now have favorable perception.
- $H_0 : \mu_0 = 0.4$ vs $H_A : \mu_0 \neq 0.4$

Example: Hypothesis testing with proportions

- Test hypothesis that **ads have an impact on perceptions about politicians**
- Current favorable perception: 0.4
- Run ad campaign. Then survey 200 people. Find 90 of them now have favorable perception.
- $H_0 : \mu_0 = 0.4$ vs $H_A : \mu_0 \neq 0.4$
- $p =$ $, s =$

Example: Hypothesis testing with proportions

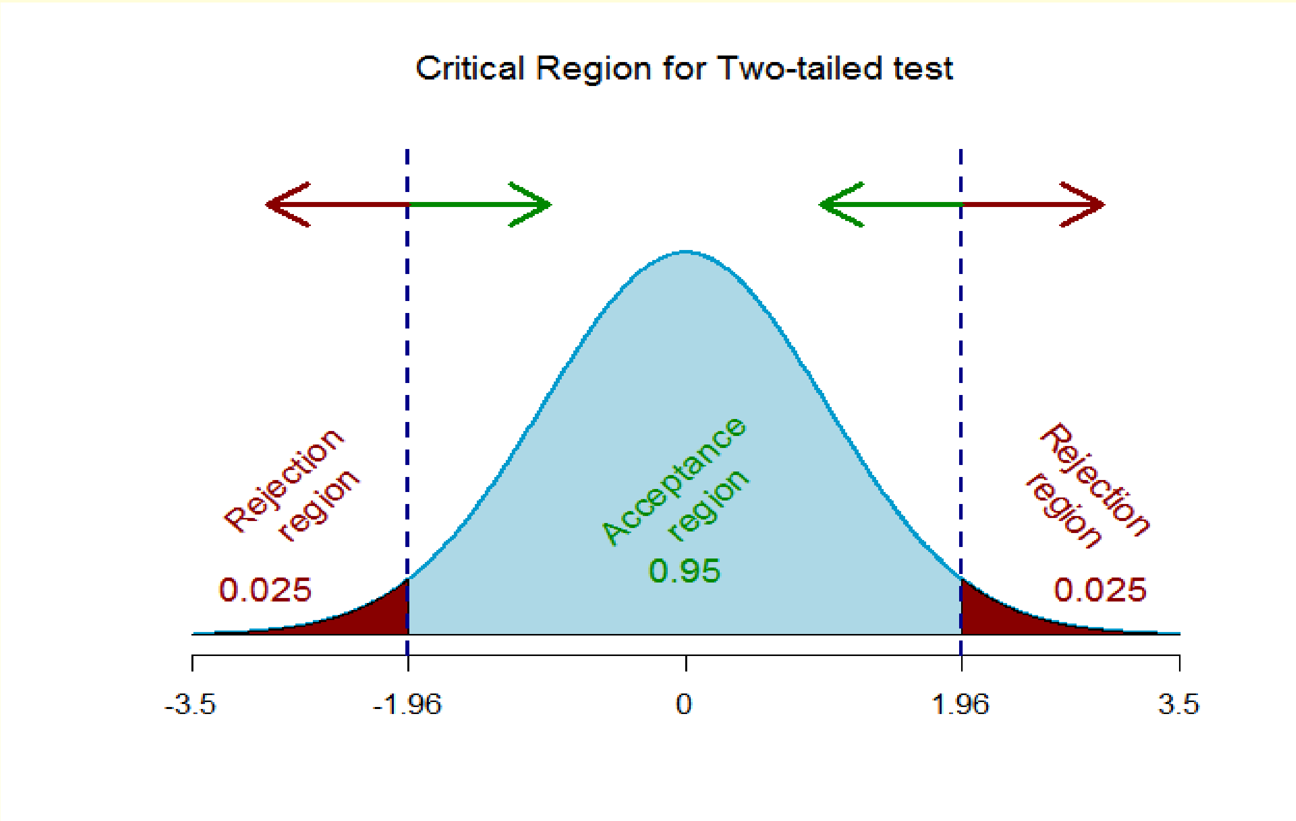
- Test hypothesis that **ads have an impact on perceptions about politicians**
- Current favorable perception: 0.4
- Run ad campaign. Then survey 200 people. Find 90 of them now have favorable perception.
- $H_0 : \mu_0 = 0.4$ vs $H_A : \mu_0 \neq 0.4$
- $p =$ $, s =$
- $z = \frac{p-\mu_0}{\frac{s}{\sqrt{n}}} =$

Example: Hypothesis testing with proportions

- Test hypothesis that **ads have an impact on perceptions about politicians**
- Current favorable perception: 0.4
- Run ad campaign. Then survey 200 people. Find 90 of them now have favorable perception.
- $H_0 : \mu_0 = 0.4$ vs $H_A : \mu_0 \neq 0.4$
- $p =$ $\qquad\qquad$ $, s =$
- $z = \frac{p-\mu_0}{\frac{s}{\sqrt{n}}} =$
- Check against: $[-1.96, 1.96]$

P-values

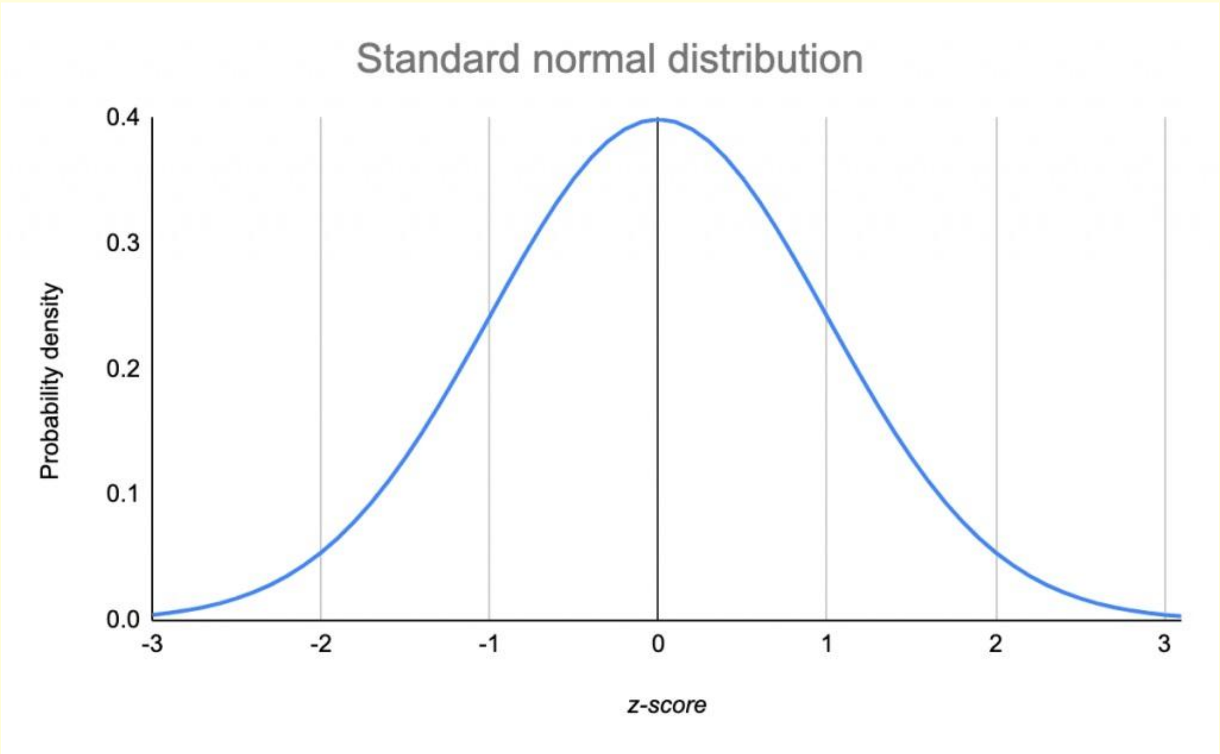
- **P-value**: What is the probability of seeing extreme data like this (in either direction) given that H_0 is true?



P-values

- **P-value**: What is the probability of seeing extreme data like this (in either direction) given that H_0 is true?

- **Reject H_0 if the P-value is small** (less than 0.05)



P-values

- **P-value**: What is the probability of seeing extreme data like this (in either direction) given that H_0 is true?
- **Reject H_0 if the P-value is small** (less than 0.05)
- In ad campaign example: $H_0 : \mu_0 = 0.4$

  Data: $n = 200$; $p = 0.45$, $z = 1.42$

P-values

- **P-value**: What is the probability of seeing extreme data like this (in either direction) given that H_0 is true?
- **Reject H_0 if the P-value is small** (less than 0.05)
- In ad campaign example: $H_0 : \mu_0 = 0.4$

  Data: $n = 200$; $p = 0.45$, $z = 1.42$
- *P-value*: $\Pr(Z \leq -1.42 \text{ or } Z \geq 1.42) = 2\Pr(Z \leq -1.42)$

P-values

- **P-value**: What is the probability of seeing extreme data like this (in either direction) given that H_0 is true?
- **Reject H_0 if the P-value is small** (less than 0.05)
- In ad campaign example: $H_0 : \mu_0 = 0.4$

  Data: $n = 200$; $p = 0.45$, $z = 1.42$
- *P-value*: $\Pr(Z \leq -1.42 \text{ or } Z \geq 1.42) = 2\Pr(Z \leq -1.42)$
- Using Excel $(NORM.DIST) : Prob(Z \leq -1.42) = 0.0778$

  P-value:

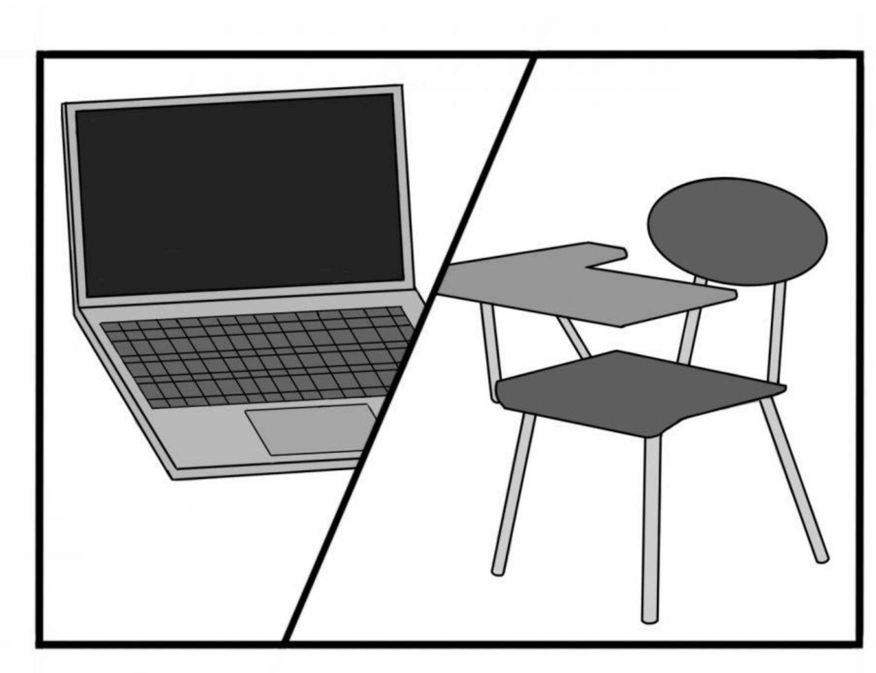
Two-sided vs one-sided tests

- So far $H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$ **(two-sided test)**

Two-sided vs one-sided tests

- So far $H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$ **(two-sided test)**
- What if $H_0 : \mu = \mu_0$ vs $H_A : \mu \leq \mu_0$ **(one-sided test)**

  Example : **Hypothesis**: Online learning has a negative impact on student learning



Two-sided vs one-sided tests

- So far $H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$ **(two-sided test)**
- What if $H_0 : \mu = \mu_0$ vs $H_A : \mu \leq \mu_0$ **(one-sided test)**

  Example : **Hypothesis**: Online learning has <u>a negative impact</u> on student learning

- Now $H_0 : \mu = 0$ vs $H_A : \mu \leq 0$

Two-sided vs one-sided tests

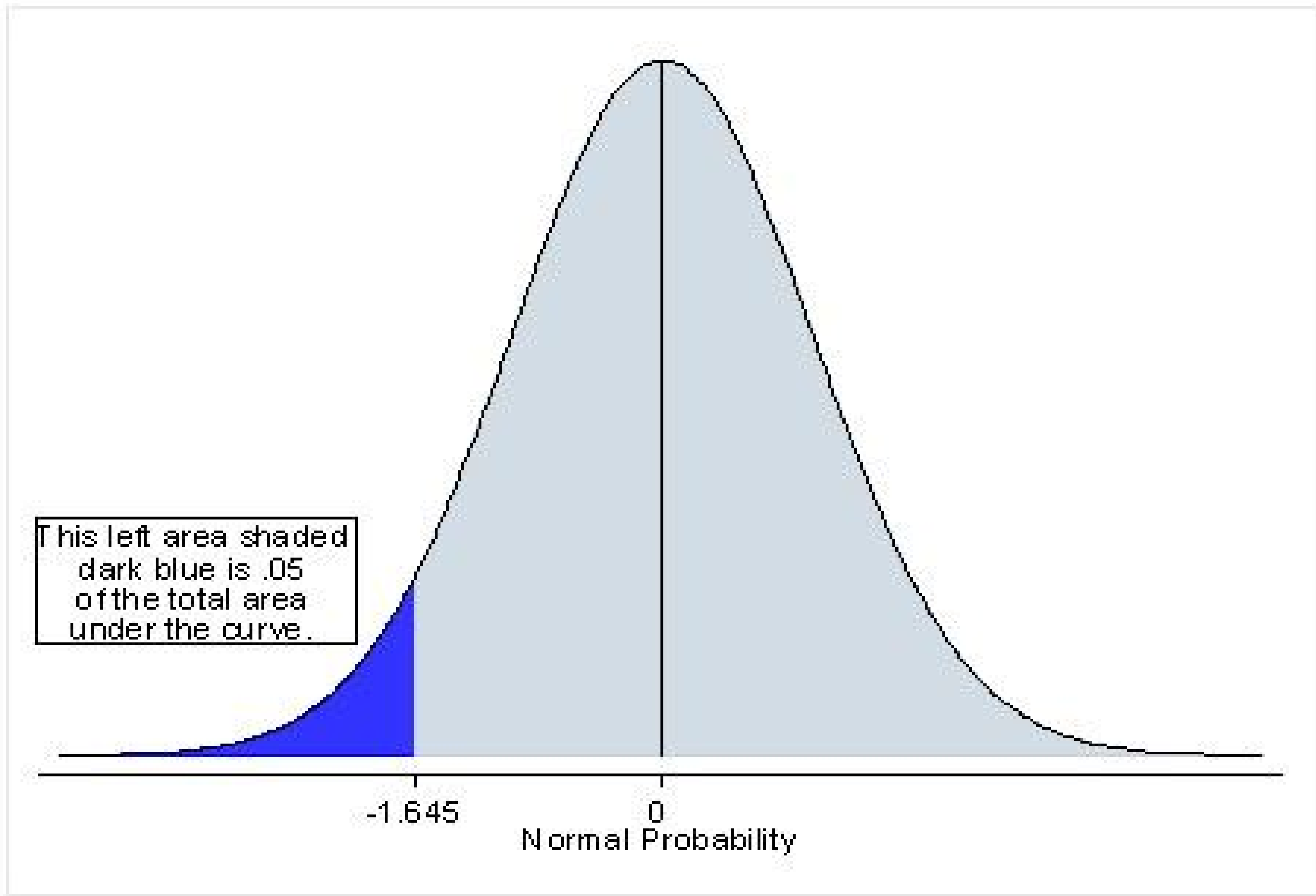
- So far $H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$ **(two-sided test)**
- What if $H_0 : \mu = \mu_0$ vs $H_A : \mu \leq \mu_0$ **(one-sided test)**

  Example : **Hypothesis**: Online learning has <u>a negative impact</u> on student learning

- Now $H_0 : \mu = 0$ vs $H_A : \mu \leq 0$
- **Same mechanics for testing:** $z = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}}$,

  only critical value different: Need $z \leq -1.645$ to reject H_0

One-sided tests

- 

One-sided tests

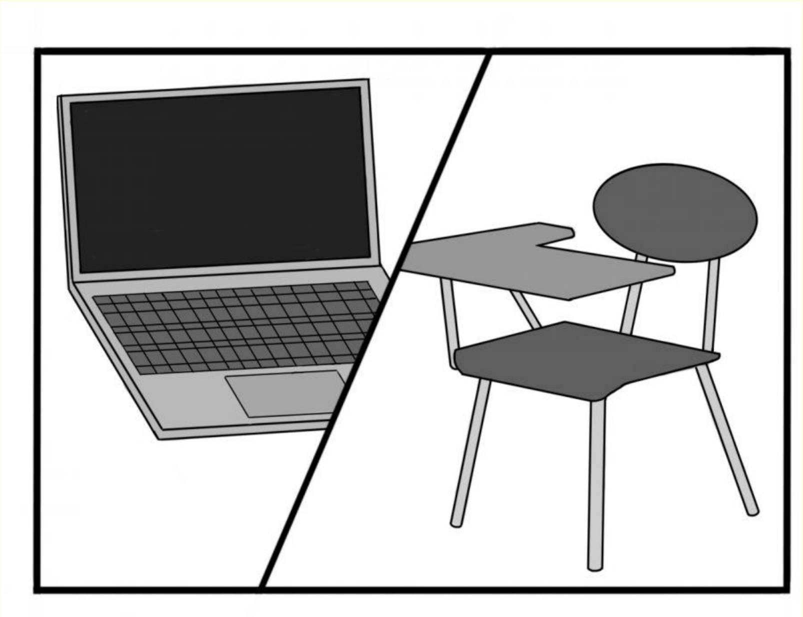
- Normal distribution with one-sided test
- Critical 95% values: $\mathbf{z} \leq -\mathbf{1.645}$ (for $H_A : \mu \leq \mu_0$)

  $\mathbf{z} \geq \mathbf{1.645}$ (for $H_A : \mu \geq \mu_0$)

One-sided tests

- Normal distribution with one-sided test
- Critical 95% values: $\mathbf{z} \leq -\mathbf{1.645}$ (for $H_A : \mu \leq \mu_0$)

  $\mathbf{z} \geq \mathbf{1.645}$ (for $H_A : \mu \geq \mu_0$)
- *Online learning example*: $n = 100$, $X = S_{\text{online}} - S_{\text{in-person}}$



One-sided tests

- Normal distribution with one-sided test
- Critical 95% values: $\mathbf{z} \leq -\mathbf{1.645}$ (for $H_A : \mu \leq \mu_0$)

  $\mathbf{z} \geq \mathbf{1.645}$ (for $H_A : \mu \geq \mu_0$)
- *Online learning example*: $n = 100$, $X = S_{\text{online}} - S_{\text{in-person}}$
- **Hypothesis**: Online learning has a $-$ive impact

  Now $H_0 : \mu = 0$ vs $H_A : \mu \leq 0$

One-sided tests

- Normal distribution with one-sided test
- Critical 95% values: $\mathbf{z} \leq -\mathbf{1.645}$ (for $H_A : \mu \leq \mu_0$)

  $\mathbf{z} \geq \mathbf{1.645}$ (for $H_A : \mu \geq \mu_0$)
- *Online learning example*: $n = 100$, $X = S_{\text{online}} - S_{\text{in-person}}$
- **Hypothesis**: Online learning has a $-$ive impact

  Now $H_0 : \mu = 0$ vs $H_A : \mu \leq 0$
- $\overline{X} = -3$, $s = 2.5$, $z = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}} =$

One-sided tests

- Normal distribution with one-sided test
- Critical 95% values: $\mathbf{z} \leq -\mathbf{1.645}$ (for $H_A : \mu \leq \mu_0$)

  $\mathbf{z} \geq \mathbf{1.645}$ (for $H_A : \mu \geq \mu_0$)
- *Online learning example*: $n = 100$, $X = S_{\text{online}} - S_{\text{in-person}}$
- **Hypothesis**: Online learning has a $-$ive impact

  Now $H_0 : \mu = 0$ vs $H_A : \mu \leq 0$
- $\overline{X} = -3$, $s = 2.5$, $z = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}} =$
- $z \leq -1.645$: reject H_0

One-sided test example

- Q: Do food for education programs enhance school enrollment?

One-sided test example

- Q: Do food for education programs enhance school enrollment?
- $H_0 : \mu = 0$ vs $H_A : \mu \geq 0$ **(one-sided test)**

One-sided test example

- Q: Do food for education programs enhance school enrollment?
- $H_0 : \mu = 0$ vs $H_A : \mu \geq 0$ **(one-sided test)**
- Study from Uganda (Alderman, Gilligan, Lehrer 2012):

One-sided test example

- Q: Do food for education programs enhance school enrollment?
- $H_0 : \mu = 0$ vs $H_A : \mu \geq 0$ **(one-sided test)**
- Study from Uganda (Alderman, Gilligan, Lehrer 2012):
- *Find for enrollment rates for ages* $6 - 9$*:*

  SFP pgm: estimate: 0.094, std. error: 0.040

  THR pgm: estimate: 0.054, std. error: 0.058

One-sided test example

- Q: Do food for education programs enhance school enrollment?
- $H_0 : \mu = 0$ vs $H_A : \mu \geq 0$ **(one-sided test)**
- Study from Uganda (Alderman, Gilligan, Lehrer 2012):
- *Find for enrollment rates for ages* $6 - 9$*:*

  SFP pgm: estimate: 0.094, std. error: 0.040

  THR pgm: estimate: 0.054, std. error: 0.058

- $z = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}}$; $z_{SFP} =$, $z_{THR} =$

One-sided test example

- Q: Do food for education programs enhance school enrollment?
- $H_0 : \mu = 0$ vs $H_A : \mu \geq 0$ **(one-sided test)**
- Study from Uganda (Alderman, Gilligan, Lehrer 2012):
- *Find for enrollment rates for ages* $6 - 9$*:*

  SFP pgm: estimate: 0.094, std. error: 0.040

  THR pgm: estimate: 0.054, std. error: 0.058

- $z = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}}$; $z_{SFP} =$, $z_{THR} =$
- $z_{SFP} \geq 1.645$: reject H_0

Connection with Regressions

- **Q**: Is there a significant relationship between a country's GDP and its CO_2 emissions?

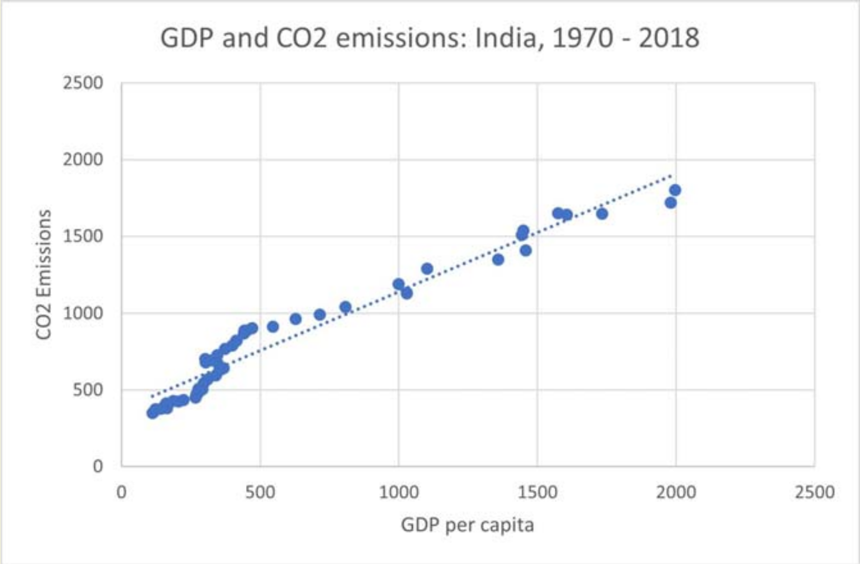
  Q: Is there a strong connection between unemployment and crime?

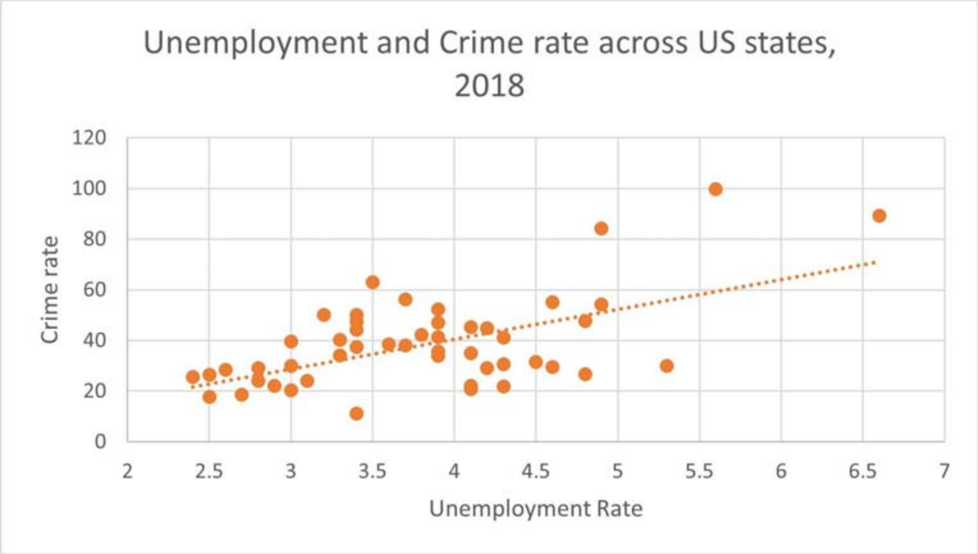
Connection with Regressions

- **Q**: Is there a significant relationship between a country's GDP and its CO_2 emissions?

  Q: Is there a strong connection between unemployment and crime?

- Tools so far: *Scatter plot* and *Regression*





Connection with Regressions

- **Q**: Is there a significant relationship between a country's GDP and its CO_2 emissions?

  Q: Is there a strong connection between unemployment and crime?

- Tools so far: *Scatter plot* and *Regression*

- **Q**: Is this relationship strong/significant?

Connection with Regressions

- *Regression:* $Y = \alpha + \beta X$

Connection with Regressions

- *Regression:* $Y = \alpha + \beta X$
- **Q**: Is this relationship strong/significant?

  Test if $H_0 : \beta = 0$ vs $H_A : \beta \neq 0$

Connection with Regressions

- *Regression:* $Y = \alpha + \beta X$
- **Q**: Is this relationship strong/significant?

  Test if $H_0 : \beta = 0$ vs $H_A : \beta \neq 0$
- **Ans**: Look at *p-values* [less than 0.05?]

  and compare *z-statistic* [outside the interval $(-1.96, 1.96)$?]

Connection with Regressions

- *Regression:* $Y = \alpha + \beta X$
- **Q**: Is this relationship strong/significant?

  Test if $H_0 : \beta = 0$ vs $H_A : \beta \neq 0$
- **Ans**: Look at *p-values* [less than 0.05?]

  and compare *z-statistic* [outside the interval $(-1.96, 1.96)$?]
- **Excel**: Data Analysis –> Regression –> Confidence Interval

Regression Examples from Research Papers

- "**Economic Shocks and Civil Conflict**" by Miguel, Satyanath and Sergenti (*Journal of Political Economy 2004*)

Miguel, Satyanath and Sergenti (2004)

TABLE 3
RAINFALL AND CIVIL CONFLICT (Reduced-Form)

Explanatory Variable	Dependent Variable: Civil Conflict ≥25 Deaths (OLS) (1)	Dependent Variable: Civil Conflict ≥1,000 Deaths (OLS) (2)
Growth in rainfall, t	−.024	−.062**
	(.043)	(.030)
Growth in rainfall, $t-1$	−.122**	−.069**
	(.052)	(.032)
Country fixed effects	yes	yes
Country-specific time trends	yes	yes
R^2	.71	.70
Root mean square error	.25	.22
Observations	743	743

NOTE.—Huber robust standard errors are in parentheses. Regression disturbance terms are clustered at the country level. A country-specific year time trend is included in all specifications (coefficient estimates not reported).

* Significantly different from zero at 90 percent confidence.
** Significantly different from zero at 95 percent confidence.
*** Significantly different from zero at 99 percent confidence.

Regression Examples from Research Papers

- "**Economic Shocks and Civil Conflict**" by Miguel, Satyanath and Sergenti (*Journal of Political Economy 2004*)
- "**Speed Limit Enforcement and Road Safety**" by Bauernschuster and Rekers (*IZA Discussion paper 2019*)

Table 4

The Effect of the Blitzmarathons on Traffic Accidents

	(1)	(2)	(3)	(4)	(5)	(6)
(a) Number of accidents						
[Mean: 2.362; N: 493,518]						
Blitzmarathon	-0.121***	-0.146***	-0.171***	-0.174***	-0.161***	-0.178***
	(0.044)	(0.045)	(0.045)	(0.046)	(0.046)	(0.047)
R^2	0.669	0.671	0.672	0.706	0.709	0.710
(b) Number of slightly injured						
[Mean: 1.916; N: 493,518]						
Blitzmarathon	-0.126**	-0.132**	-0.154***	-0.165***	-0.155***	-0.163***
	(0.052)	(0.052)	(0.052)	(0.052)	(0.051)	(0.052)
R^2	0.582	0.583	0.584	0.620	0.623	0.624
(c) Number of severely injured						
[Mean: 0.367; N: 493,518]						
Blitzmarathon	-0.036*	-0.032	-0.035*	-0.031	-0.029	-0.033
	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)
R^2	0.123	0.124	0.124	0.130	0.130	0.128
(d) Number of fatally injured						
[Mean: 0.021; N: 493,518]						
Blitzmarathon	-0.002	-0.001	-0.002	-0.001	-0.001	-0.001
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
R^2	0.122	0.123	0.124	0.129	0.129	0.128
County FE	×	×	×	×	×	×
Time FE	×	×	×	×	×	×
Weather		×	×	×	×	×
Vacation			×	×	×	×
County × Time FE				×	×	×
County × Weather					×	×
County × Vacation						×

Notes: The table shows the effect of the Blitzmarathons on the number of traffic accidents [Panel (a)], slightly injured [Panel (b)], severely injured [Panel (c)], and fatally injured [Panel (d)]. Each column in each row presents a separate regression. All regressions are run at the county-day level. "Blitzmarathon" is as a dummy variable indicating the Blitzmarathon is in force in a specific county on a specific day. All regressions include county and time fixed effects. Time fixed effects include day-of-week, month-of-year, and year fixed effects. Weather controls include atmospheric temperature, amount of precipitation, and a dummy for snow cover. Additionally, we include dummies indicating missing atmospheric temperature, missing amount of precipitation, and missing snow cover. Vacation controls include dummies for school vacation, the last school day before a school vacation, and the last day of a school vacation. County × Time, County × Weather, and County × Vacation are interaction of county indicators with all time fixed effects, weather controls, and vacation controls, respectively. Standard errors (in parentheses) are clustered at the county level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Module Recap

- Introduction to Statistical Inference
- Sampling
- The main theorems underlying statistical inference:

  Law of Large Numbers and *Central Limit Theorem*
- Point estimate, Standard Error and Confidence Intervals
- Hypothesis testing and connection with Regressions