Lecture 11:
Intro to Statistical Inference
API-201Z

Maya Sen

Harvard Kennedy School
http://scholar.harvard.edu/msen
Announcements

▶ Midterm #1 nearly done
▶ Will post distribution & solutions ASAP
▶ Will email when exams ready for pick up from Melissa Kappotis
▶ Quick note about final exercise group sizes (2-4)
Announcements

- Midterm #1 nearly done
Announcements

- Midterm #1 nearly done
- Will post distribution & solutions ASAP
Announcements

- Midterm #1 nearly done
- Will post distribution & solutions ASAP
- Will email when exams ready for pick up from Melissa Kappotis
Announcements

- Midterm #1 nearly done
- Will post distribution & solutions ASAP
- Will email when exams ready for pick up from Melissa Kappotis
- Quick note about final exercise group sizes (2-4)
Roadmap

- What do we mean by statistical inference?
- Estimation and estimators
- Sampling distributions
- Central Limit Theorem (and Law of Large Numbers)
- Applications of the CLT
Roadmap

- What do we mean by statistical inference?
Roadmap

- What do we mean by statistical inference?
- Estimation and estimators
Roadmap

- What do we mean by statistical inference?
- Estimation and estimators
- Sampling distributions
Roadmap

- What do we mean by statistical inference?
- Estimation and estimators
- Sampling distributions
- Central Limit Theorem (and Law of Large Numbers)
Roadmap

- What do we mean by statistical inference?
- Estimation and estimators
- Sampling distributions
- Central Limit Theorem (and Law of Large Numbers)
- Applications of the CLT
Big Picture

Probability: If we knew truth about how data were generated, then probability tells us what data we should expect.

Statistical Inference: Learning about the population given a sample/data.
Big Picture

- Probability: If we knew truth about how data were generated, then probability tells us what data we should expect
Big Picture

- **Probability**: If we knew truth about how data were generated, then probability tells us what data we should expect
- **Statistical Inference**: Learning about the population given a sample/data
Big Picture

- Probability: If we knew truth about how data were generated, then probability tells us what data we should expect
- Statistical Inference: Learning about the population given a sample/data
Big Picture

- **Probability**: If we knew truth about how data were generated, then probability tells us what data we should expect.
- **Statistical Inference**: Learning about the population given a sample/data.
Big Picture

- **Probability**: If we knew truth about how data were generated, then probability tells us what data we should expect.
- **Statistical Inference**: Learning about the population given a sample/data.
Big Picture

- **Probability:** If we knew truth about how data were generated, then probability tells us what data we should expect.
- **Statistical Inference:** Learning about the population given a sample/data.

![Diagram](image)
Statistical Inference

In this class, we will learn about the population from a sample in 2 ways:

1. Estimation: Use data from a sample to estimate value of population parameter of interest
   - Examine series of polls to estimate true population U.S. presidential approval
   - Sample 100 women to estimate true average BMI for Ugandan women

2. Hypothesis Testing: Use data from a sample to assess a particular belief or theory
   - Examine income data to assess whether mean household income has changed since 2010
   - Sample 100 U.K. men and women to see if men earn more than women

Will start next time
In this class, we will learn about the population from a sample in 2 ways:

1. Estimation: Use data from a sample to estimate value of population parameter of interest
   - Examine series of polls to estimate true population U.S. presidential approval
   - Sample 100 women to estimate true average BMI for Ugandan women

2. Hypothesis Testing: Use data from a sample to assess a particular belief or theory
   - Examine income data to assess whether mean household income has changed since 2010
   - Sample 100 U.K. men and women to see if men earn more than women

Will start next time
Statistical Inference

In this class, we will learn about the population from a sample in 2 ways:

1. **Estimation**: Use data from a sample to estimate value of population parameter of interest

   ▶ Examine series of polls to estimate true population U.S. presidential approval
   ▶ Sample 100 women to estimate true average BMI for Ugandan women

2. **Hypothesis Testing**: Use data from a sample to assess a particular belief or theory

   ▶ Examine income data to assess whether mean household income has changed since 2010
   ▶ Sample 100 U.K. men and women to see if men earn more than women

Will start next time
In this class, we will learn about the population from a sample in 2 ways:

1. **Estimation**: Use data from a sample to estimate value of population parameter of interest

   - Examine series of polls to estimate true population U.S. presidential approval
   - Sample 100 women to estimate true average BMI for Ugandan women

2. **Hypothesis Testing**: Use data from a sample to assess a particular belief or theory

   - Examine income data to assess whether mean household income has changed since 2010
   - Sample 100 U.K. men and women to see if men earn more than women

   Will start next time
Statistical Inference

In this class, we will learn about the population from a sample in 2 ways:

1. **Estimation**: Use data from a sample to estimate value of population parameter of interest
   - Examine series of polls to estimate true population U.S. presidential approval

2. **Hypothesis Testing**: Use data from a sample to assess a particular belief or theory
   - Examine income data to assess whether mean household income has changed since 2010
   - Sample 100 U.K. men and women to see if men earn more than women

Will start next time
In this class, we will learn about the population from a sample in 2 ways:

1. **Estimation**: Use data from a sample to estimate value of population parameter of interest
   - Examine series of polls to estimate true population U.S. presidential approval
   - Sample 100 women to estimate true average BMI for Ugandan women
In this class, we will learn about the population from a sample in 2 ways:

1. **Estimation**: Use data from a sample to estimate value of population parameter of interest
   - Examine series of polls to estimate true population U.S. presidential approval
   - Sample 100 women to estimate true average BMI for Ugandan women

2. **Hypothesis Testing**: Use data from a sample to assess a particular belief or theory
Statistical Inference

In this class, we will learn about the population from a sample in 2 ways:

1. **Estimation**: Use data from a sample to estimate value of population parameter of interest
   - Examine series of polls to estimate true population U.S. presidential approval
   - Sample 100 women to estimate true average BMI for Ugandan women

2. **Hypothesis Testing**: Use data from a sample to assess a particular belief or theory
   - Examine income data to assess whether mean household income has changed since 2010
In this class, we will learn about the population from a sample in 2 ways:

1. **Estimation:** Use data from a sample to estimate value of population parameter of interest
   - Examine series of polls to estimate true population U.S. presidential approval
   - Sample 100 women to estimate true average BMI for Ugandan women

2. **Hypothesis Testing:** Use data from a sample to assess a particular belief or theory
   - Examine income data to assess whether mean household income has changed since 2010
   - Sample 100 U.K. men and women to see if men earn more than women
Statistical Inference

In this class, we will learn about the population from a sample in 2 ways:

1. **Estimation**: Use data from a sample to estimate value of population parameter of interest
   - Examine series of polls to estimate true population U.S. presidential approval
   - Sample 100 women to estimate true average BMI for Ugandan women

2. **Hypothesis Testing**: Use data from a sample to assess a particular belief or theory
   - Examine income data to assess whether mean household income has changed since 2010
   - Sample 100 U.K. men and women to see if men earn more than women
   - Will start next time
Estimation

Estimator: Rule or formula that we apply to a sample that produces a number called an estimate.

Estimator often denoted with "hat" symbol.

Ex) \( \hat{\mu} \) is an estimator for the population mean.

Often use \( \hat{\theta} \) as generic notation.

Because it's a function of the data, it's a random variable.

Estimate: Our guess (usually our best one) of the true value of the population parameter.

Realization of the random variable that is the estimator.

In other words: Using an estimator on a particular sample gives us an estimate.
Estimation

- **Estimator**: Rule or formula that we apply to a sample that produces a number called an estimate

- Estimator often denoted with "hat" symbol

- Ex) $\hat{\mu}$ is an estimator for the population mean

- Often use $\hat{\theta}$ as generic notation

- Because it's a function of the data, it's a random variable

- Estimate: Our guess (usually our best one) of the true value of the population parameter

- Realization of the random variable that is the estimator

- In other words: Using an estimator on a particular sample gives us an estimate
Estimation

- **Estimator**: Rule or formula that we apply to a sample that produces a number called an estimate
  - Estimator often denoted with “hat” symbol
Estimation

- **Estimator**: Rule or formula that we apply to a sample that produces a number called an estimate
  - Estimator often denoted with “hat” symbol
  - Ex) $\hat{\mu}$ is an estimator for the population mean
Estimation

- **Estimator**: Rule or formula that we apply to a sample that produces a number called an estimate
  - Estimator often denoted with “hat” symbol
  - Ex) \( \hat{\mu} \) is an estimator for the population mean
  - Often use \( \hat{\theta} \) as generic notation
Estimation

- **Estimator**: Rule or formula that we apply to a sample that produces a number called an estimate
  - Estimator often denoted with “hat” symbol
  - Ex) $\hat{\mu}$ is an estimator for the population mean
  - Often use $\hat{\theta}$ as generic notation
  - Because it’s a function of the data, it’s a random variable

- **Estimate**: Our guess (usually our best one) of the true value of the population parameter
  - Realization of the random variable that is the estimator
  - In other words: Using an estimator on a particular sample gives us an estimate
Estimation

- **Estimator**: Rule or formula that we apply to a sample that produces a number called an estimate
  - Estimator often denoted with “hat” symbol
  - Ex) $\hat{\mu}$ is an estimator for the population mean
  - Often use $\hat{\theta}$ as generic notation
  - Because it’s a function of the data, it’s a random variable

- **Estimate**: Our guess (usually our best one) of the true value of the population parameter
Estimation

- **Estimator**: Rule or formula that we apply to a sample that produces a number called an estimate
  - Estimator often denoted with “hat” symbol
  - Ex) \( \hat{\mu} \) is an estimator for the population mean
  - Often use \( \hat{\theta} \) as generic notation
  - Because it’s a function of the data, it’s a random variable

- **Estimate**: Our guess (usually our best one) of the true value of the population parameter
  - Realization of the random variable that is the estimator
Estimation

- **Estimator**: Rule or formula that we apply to a sample that produces a number called an estimate
  - Estimator often denoted with “hat” symbol
  - Ex) \( \hat{\mu} \) is an estimator for the population mean
  - Often use \( \hat{\theta} \) as generic notation
  - Because it’s a function of the data, it’s a random variable
- **Estimate**: Our guess (usually our best one) of the true value of the population parameter
  - Realization of the random variable that is the estimator
- In other words: Using an estimator on a particular sample gives us an estimate
Estimation

We actually have covered several frequently used estimators:

▶ Sample mean: Estimator of true population mean
▶ Sample proportion: Estimator of true population proportion
▶ Sample median: Estimator of true population median
▶ Sample standard deviation: Estimator of true population spread (population standard deviation)
We actually have covered several frequently used estimators:
Estimation

We actually have covered several frequently used estimators:

- Sample mean: Estimator of true population mean
We actually have covered several frequently used estimators:

- Sample mean: Estimator of true population mean
- Sample proportion: Estimator of true population proportion
We actually have covered several frequently used estimators:

- Sample mean: Estimator of true population mean
- Sample proportion: Estimator of true population proportion
- Sample median: Estimator of true population median
We actually have covered several frequently used estimators:

- Sample mean: Estimator of true population mean
- Sample proportion: Estimator of true population proportion
- Sample median: Estimator of true population median
- Sample standard deviation: Estimator of true population spread (population standard deviation)
Estimation

Some estimators are better/worse than others.

▶ What is a good estimator for the population mean, $\mu$?

▶ $\hat{\theta} = 3$ → Estimate is always 3.

▶ $\hat{\theta} = Y_1$ → Use the first observation.

▶ $\hat{\theta} = \max(Y_1, \ldots, Y_n)$ → Use largest observation.

▶ $\hat{\theta} = \bar{Y}$ → Use the sample mean.

Statisticians have developed criteria for evaluating the goodness of an estimator (for now, we’ll skip!)
Estimation

Some estimators are better/worse than others

\[ \hat{\theta} = 3 \rightarrow \text{Estimate is always 3} \]

\[ \hat{\theta} = Y_1 \rightarrow \text{Use the first observation} \]

\[ \hat{\theta} = \max(Y_1, \ldots, Y_n) \rightarrow \text{Use largest observation} \]

\[ \hat{\theta} = \bar{Y} \rightarrow \text{Use the sample mean} \]

Statisticians have developed criteria for evaluating the goodness of an estimator (for now, we'll skip!)
Some estimators are better/worse than others

- What is a good estimator for the population mean, $\mu_y$?
Estimation

Some estimators are better/worse than others

- What is a good estimator for the population mean, $\mu_y$?
- $\hat{\theta} = 3 \rightarrow$ Estimate is always 3
Some estimators are better/worse than others

- What is a good estimator for the population mean, $\mu_y$?
- $\hat{\theta} = 3 \rightarrow$ Estimate is always 3
- $\hat{\theta} = Y_1 \rightarrow$ Use the first observation
Some estimators are better/worse than others

- What is a good estimator for the population mean, $\mu_y$?
- $\hat{\theta} = 3 \rightarrow$ Estimate is always 3
- $\hat{\theta} = Y_1 \rightarrow$ Use the first observation
- $\hat{\theta} = \max(Y_1, \ldots, Y_n) \rightarrow$ Use largest observation
Estimation

Some estimators are better/worse than others

- What is a good estimator for the population mean, $\mu_y$?
- $\hat{\theta} = 3 \rightarrow$ Estimate is always 3
- $\hat{\theta} = Y_1 \rightarrow$ Use the first observation
- $\hat{\theta} = \max(Y_1, \ldots, Y_n) \rightarrow$ Use largest observation
- $\hat{\theta} = \bar{Y} \rightarrow$ Use the sample mean
Estimation

Some estimators are better/worse than others

- What is a good estimator for the population mean, $\mu_y$?
- $\hat{\theta} = 3 \rightarrow$ Estimate is always 3
- $\hat{\theta} = Y_1 \rightarrow$ Use the first observation
- $\hat{\theta} = \max(Y_1, \ldots, Y_n) \rightarrow$ Use largest observation
- $\hat{\theta} = \bar{Y} \rightarrow$ Use the sample mean

Statisticians have developed criteria for evaluating the goodness of an estimator (for now, we’ll skip!)
Sampling Fluctuation

To use an estimator, you need a sample.

Ex) simple random sample, stratified sample, clustered sample

However: Every time you sample, you get slightly different observations in the sample.

Ex) Health researcher studying nutrition samples 1000 Ugandan women, another research replicates study

Ex) 5 friends each sample 20 HKS faculty (out of 120)

Ex) Lobbying firm randomly choose 5 Senators to call on Day 1, randomly choose 5 Senators to call on Day 2

These sampling fluctuations mean that different samples generate different estimates.
Sampling Fluctuation

- To use an estimator, you need a sample
Sampling Fluctuation

- To use an estimator, you need a sample
  - Ex) simple random sample, stratified sample, clustered sample
Sampling Fluctuation

- To use an estimator, you need a sample
  - Ex) simple random sample, stratified sample, clustered sample
- However: Every time you sample, you get slightly different observations in the sample
Sampling Fluctuation

- To use an estimator, you need a sample
  - Ex) simple random sample, stratified sample, clustered sample
- However: Every time you sample, you get slightly different observations in the sample
  - Ex) Health researcher studying nutrition samples 1000 Ugandan women, another research replicates study
Sampling Fluctuation

- To use an estimator, you need a sample
  - Ex) simple random sample, stratified sample, clustered sample
- However: Every time you sample, you get slightly different observations in the sample
  - Ex) Health researcher studying nutrition samples 1000 Ugandan women, another research replicates study
  - Ex) 5 friends each sample 20 HKS faculty (out of 120)
Sampling Fluctuation

- To use an estimator, you need a sample
  - Ex) simple random sample, stratified sample, clustered sample
- However: Every time you sample, you get slightly different observations in the sample
  - Ex) Health researcher studying nutrition samples 1000 Ugandan women, another research replicates study
  - Ex) 5 friends each sample 20 HKS faculty (out of 120)
  - Ex) Lobbying firm randomly choose 5 Senators to call on Day 1, randomly choose 5 Senators to call on Day 2
Sampling Fluctuation

- To use an estimator, you need a sample
  - Ex) simple random sample, stratified sample, clustered sample
- However: Every time you sample, you get slightly different observations in the sample
  - Ex) Health researcher studying nutrition samples 1000 Ugandan women, another research replicates study
  - Ex) 5 friends each sample 20 HKS faculty (out of 120)
  - Ex) Lobbying firm randomly choose 5 Senators to call on Day 1, randomly choose 5 Senators to call on Day 2
- These sampling fluctuations mean that different samples generate different estimates
Political Polls Example

Many private companies poll the US population regularly – YouGov, Rasmussen, NBC, NYT, etc.

Suppose we are interested in true share of population who believe country moving in “right direction”

Population of interest? (The entire US population)

Population parameter of interest? (US population mean)
Political Polls Example

Many private companies poll the US population regularly – YouGov, Rasmussen, NBC, NYT, etc.
Political Polls Example

- Many private companies poll the US population regularly – YouGov, Rasmussen, NBC, NYT, etc.
- Suppose we are interested in true share of population who believe country moving in “right direction”
Political Polls Example

- Many private companies poll the US population regularly – YouGov, Rasmussen, NBC, NYT, etc.
- Suppose we are interested in true share of population who believe country moving in “right direction”
- Population of interest?
Political Polls Example

- Many private companies poll the US population regularly – YouGov, Rasmussen, NBC, NYT, etc.
- Suppose we are interested in true share of population who believe country moving in “right direction”
- Population of interest? (The entire US population)
Political Polls Example

- Many private companies poll the US population regularly – YouGov, Rasmussen, NBC, NYT, etc.
- Suppose we are interested in true share of population who believe country moving in “right direction”
- Population of interest? (The entire US population)
- Population parameter of interest?
Many private companies poll the US population regularly – YouGov, Rasmussen, NBC, NYT, etc.

Suppose we are interested in true share of population who believe country moving in “right direction”

Population of interest? (The entire US population)

Population parameter of interest? (US population mean)
Political Polls Example
Political Polls Example

- Real Clear Politics: Is the country moving in the right direction?
Political Polls Example

- Real Clear Politics: Is the country moving in the right direction?
- Hundreds and hundreds of polls
Political Polls Example

▶ Real Clear Politics: Is the country moving in the right direction?
▶ Hundreds and hundreds of polls
Political Polls Example

Why does the share fluctuate across polls?

If you kept repeating the polling, the estimates would take on their own distribution (remember, they are random variables!)

That distribution is called the sampling distribution

We can use sampling distribution to get a better sense of what underlying population distribution may look like

Turns out, the sampling distribution has attractive properties
Political Polls Example

- Why does the share fluctuate across polls?

If you kept repeating the polling, the estimates would take on their own distribution (remember, they are random variables!). That distribution is called the sampling distribution. We can use sampling distribution to get a better sense of what underlying population distribution may look like. Turns out, the sampling distribution has attractive properties.
Political Polls Example

- Why does the share fluctuate across polls?
- If you kept repeating the polling, the estimates would take on their own distribution
Political Polls Example

- Why does the share fluctuate across polls?
- If you kept repeating the polling, the estimates would take on their own distribution (remember, they are random variables!)
Political Polls Example

- Why does the share fluctuate across polls?
- If you kept repeating the polling, the estimates would take on their own distribution (remember, they are random variables!)
- That distribution is called the sampling distribution
Political Polls Example

- Why does the share fluctuate across polls?
- If you kept repeating the polling, the estimates would take on their own distribution (remember, they are random variables!)
- That distribution is called the sampling distribution
- We can use sampling distribution to get a better sense of what underlying population distribution may look like
Political Polls Example

- Why does the share fluctuate across polls?
- If you kept repeating the polling, the estimates would take on their own distribution (remember, they are random variables!)
- That distribution is called the sampling distribution
- We can use sampling distribution to get a better sense of what underlying population distribution may look like
  - Turns out, the sampling distribution has attractive properties
Three Distributions

Make sure you understand these distinctions:

- Population distribution: Underlying probability distribution that generates the data (unobserved)
- What we want to learn about
- Sample distribution: Distribution of your sample data (observed)
- Sampling distribution: Distribution of your estimates (usually only get one observation, so this hinges on repeated sampling)
Three Distributions

Make sure you understand these distinctions:

- **Population distribution**: Underlying probability distribution that generates the data (unobserved)
Three Distributions

Make sure you understand these distinctions:

- **Population distribution**: Underlying probability distribution that generates the data (unobserved)
  - What we want to learn about
Three Distributions

Make sure you understand these distinctions:

- **Population distribution**: Underlying probability distribution that generates the data (unobserved)
  - What we want to learn about

- **Sample distribution**: Distribution of your sample data (observed)
Three Distributions

Make sure you understand these distinctions:

- **Population distribution**: Underlying probability distribution that generates the data (unobserved)
  - What we want to learn about

- **Sample distribution**: Distribution of your sample data (observed)

- **Sampling distribution**: Distribution of your estimates (usually only get one observation, so this hinges on repeated sampling)
Three Distributions

Using weight of Ugandan women as an example

▶ Population distribution: Underlying probability distribution that generates the data (unobserved, but have reason to think this follows a Normal distribution)

▶ Sample distribution: Distribution of your sample data (observed)

▶ Sample 1000 women → what does histogram of data look like? What does mean look like?

▶ Sampling distribution: Distribution of your estimates (usually only get one observation)

▶ (1) Take a sample, calculate sample mean, and record, (2) take another sample and repeat, (3) take another sample and repeat...etc.

▶ What does a histogram of the recorded means look like?
Three Distributions

Using weight of Ugandan women as an example
Three Distributions

Using weight of Ugandan women as an example

- **Population distribution**: Underlying probability distribution that generates the data (unobserved, but have reason to think this follows a Normal distribution)

- **Sample distribution**: Distribution of your sample data (observed)

- **Sample 1000 women** → what does histogram of data look like? What does mean look like?

- **Sampling distribution**: Distribution of your estimates (usually only get one observation)

  - Take a sample, calculate sample mean, and record,
    - take another sample and repeat,
    - take another sample and repeat...

  - What does a histogram of the recorded means look like?
Three Distributions

Using weight of Ugandan women as an example

- **Population distribution**: Underlying probability distribution that generates the data (unobserved, but have reason to think this follows a Normal distribution)
- **Sample distribution**: Distribution of your sample data (observed)
Three Distributions

Using weight of Ugandan women as an example

- **Population distribution**: Underlying probability distribution that generates the data (unobserved, but have reason to think this follows a Normal distribution)
- **Sample distribution**: Distribution of your sample data (observed)
  - Sample 1000 women → what does histogram of data look like? What does mean look like?

- **Sampling distribution**: Distribution of your estimates (usually only get one observation)

  1. Take a sample, calculate sample mean, and record
  2. Take another sample and repeat
  3. Take another sample and repeat... etc.

  What does a histogram of the recorded means look like?
Three Distributions

Using weight of Ugandan women as an example

- **Population distribution**: Underlying probability distribution that generates the data (unobserved, but have reason to think this follows a Normal distribution)

- **Sample distribution**: Distribution of your sample data (observed)
  - Sample 1000 women → what does histogram of data look like? What does mean look like?

- **Sampling distribution**: Distribution of your estimates (usually only get one observation)
Three Distributions

Using weight of Ugandan women as an example

- **Population distribution**: Underlying probability distribution that generates the data (unobserved, but have reason to think this follows a Normal distribution)

- **Sample distribution**: Distribution of your sample data (observed)
  - Sample 1000 women → what does histogram of data look like? What does mean look like?

- **Sampling distribution**: Distribution of your estimates (usually only get one observation)
  - (1) Take a sample, calculate sample mean, and record, (2) take another sample and repeat, (3) take another sample and repeat...etc.
Three Distributions

Using weight of Ugandan women as an example

- **Population distribution**: Underlying probability distribution that generates the data (unobserved, but have reason to think this follows a Normal distribution)

- **Sample distribution**: Distribution of your sample data (observed)
  - Sample 1000 women → what does histogram of data look like? What does mean look like?

- **Sampling distribution**: Distribution of your estimates (usually only get one observation)
  - (1) Take a sample, calculate sample mean, and record, (2) take another sample and repeat, (3) take another sample and repeat...etc.
  - What does a histogram of the recorded means look like?
Three Distributions

Make sure you understand these distinctions:
Three Distributions

Make sure you understand these distinctions:

- Population distribution
Three Distributions

Make sure you understand these distinctions:

- Population distribution
- Sample distribution
Three Distributions

Make sure you understand these distinctions:

- Population distribution
- Sample distribution
- Sampling distribution
Three Distributions

Make sure you understand these distinctions:
Three Distributions

Make sure you understand these distinctions:

- Population mean
Three Distributions

Make sure you understand these distinctions:

- Population mean
- Sample mean
Three Distributions

Make sure you understand these distinctions:

- Population mean
- Sample mean
- Mean of the sampling distribution
If this isn’t clear...
If this isn’t clear...
Why Sampling Distributions?

We use sampling distributions because they have very nice properties.

How would we know what the sampling distribution of $\bar{X}$ would look like? (Seems hard!)

Fortunately, for many kinds of estimators (including sample means) we have a solution:

The Central Limit Theorem!
Why Sampling Distributions?

- We use sampling distributions because they have very nice properties
Why Sampling Distributions?

- We use sampling distributions because they have very nice properties.
- How would we know what the sampling distribution of $\bar{X}$ would look like? (Seems hard!)
Why Sampling Distributions?

- We use sampling distributions because they have very nice properties.
- How would we know what the sampling distribution of $\bar{X}$ would look like? (Seems hard!)
- Fortunately, for many kinds of estimators (including sample means) we have a solution:
Why Sampling Distributions?

- We use sampling distributions because they have very nice properties.
- How would we know what the sampling distribution of $\bar{X}$ would look like? (Seems hard!)
- Fortunately, for many kinds of estimators (including sample means) we have a solution:
  - The Central Limit Theorem!
Central Limit Theorem (CLT) has two parts:

1. The sums and means of independent random variables have an approximately normal distribution.
2. This distribution becomes "more and more normal" the more observations are included in the sum or the mean.

CLT big reason why Normal distributions so important!

(CLT implies Law of Large Numbers: As the number of observations in the sample increases, \( \bar{X} \) approaches the population mean, \( \mu \)).
Central Limit Theorem (CLT) has two parts:

1. The sums and means of independent random variables have an approximately normal distribution.

CLT is big reason why Normal distributions so important!

(CLT implies Law of Large Numbers: As number of observations in sample increases, $\bar{X}$ approaches population $\mu$.)
Central Limit Theorem (CLT) has two parts:

1. The sums and means of independent random variables have an approximately normal distribution.
2. This distribution becomes "more and more normal" the more observations are included in the sum or the mean.
Central Limit Theorem (CLT) has two parts:

1. The sums and means of independent random variables have an approximately normal distribution.
2. This distribution becomes “more and more normal” the more observations are included in the sum or the mean.

CLT big reason why Normal distributions so important!
Central Limit Theorem (CLT) has two parts:

1. The sums and means of independent random variables have an approximately normal distribution.
2. This distribution becomes “more and more normal” the more observations are included in the sum or the mean.

CLT big reason why Normal distributions so important!

(CLT implies Law of Large Numbers: As \( \# \) of observations in sample increases, \( \bar{X} \) approaches population \( \mu \)).
Central Limit Theorem

So, using CLT for $\bar{X}$

$E[\bar{X}] = \mu$

For $\text{Var}[\bar{X}]$:

$\text{Var}[\bar{X}] = \text{Var}[X_1 + X_2 + \ldots + X_n] = n\sigma^2/n^2 = \sigma^2/n$

(See proof in appendix)

And for $SD[\bar{X}]$:

$\sqrt{\text{Var}[\bar{X}]} = \sigma/\sqrt{n}$

Under CLT $\rightarrow$ As $n$ goes up, distribution of $\bar{X}$ approaches:

$\bar{X} \sim N(\mu, \sigma^2/n)$
Central Limit Theorem

So, using CLT for $\bar{X}$

- $E[\bar{X}] = \mu$

$\text{Var}[\bar{X}] = \text{Var}[X_1 + X_2 + \ldots + X_n] = \frac{\sigma^2}{n}$

Under CLT → As $n$ goes up, distribution of $\bar{X}$ approaches:

$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$
Central Limit Theorem

So, using CLT for $\bar{X}$

- $E[\bar{X}] = \mu$
- For $\text{Var}[\bar{X}]$:

\[
\text{Var}[\bar{X}] = \text{Var}[X_1 + X_2 + \ldots + X_n] = n \sigma^2/n^2 = \sigma^2/n
\]

(See proof in appendix)

And for $\text{SD}[\bar{X}]$:

\[
\sqrt{\text{Var}[\bar{X}]} = \frac{\sigma}{\sqrt{n}}
\]

Under CLT → As $n$ goes up, distribution of $\bar{X}$ approaches:

$\bar{X} \sim N(\mu, \sigma^2/n)$
Central Limit Theorem

So, using CLT for $\bar{X}$

- $E[\bar{X}] = \mu$
- For $\text{Var}[\bar{X}]$:

$$\text{Var}[\bar{X}] = \text{Var}\left[\frac{X_1 + X_2 + \ldots + X_n}{n}\right] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

(See proof in appendix)

And for $\text{SD}[\bar{X}]$:

$$\sqrt{\text{Var}[\bar{X}]} = \frac{\sigma}{\sqrt{n}}$$

Under CLT
→ As $n$ goes up, distribution of $\bar{X}$ approaches:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$
Central Limit Theorem

So, using CLT for $\bar{X}$

- $E[\bar{X}] = \mu$
- For $Var[\bar{X}]$:

$$Var[\bar{X}] = Var\left[\frac{X_1 + X_2 + \ldots + X_n}{n}\right] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

- (See proof in appendix)
Central Limit Theorem

So, using CLT for $\bar{X}$

- $E[\bar{X}] = \mu$
- For $Var[\bar{X}]$:

$$Var[\bar{X}] = Var\left[\frac{X_1 + X_2 + \ldots + X_n}{n}\right] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

- (See proof in appendix)
- And for $SD[\bar{X}]$:
Central Limit Theorem

So, using CLT for $\bar{X}$

- $E[\bar{X}] = \mu$
- For $\text{Var}[\bar{X}]$:

\[
\text{Var}[\bar{X}] = \text{Var}\left[\frac{X_1 + X_2 + \ldots + X_n}{n}\right] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}
\]

- (See proof in appendix)
- And for $\text{SD}[\bar{X}]$:

\[
\sqrt{\text{Var}[\bar{X}]} = \frac{\sigma}{\sqrt{n}}
\]
Central Limit Theorem

So, using CLT for $\bar{X}$

- $E[\bar{X}] = \mu$
- For $Var[\bar{X}]$:

$$Var[\bar{X}] = Var\left[\frac{X_1 + X_2 + \ldots + X_n}{n}\right] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

(See proof in appendix)

- And for $SD[\bar{X}]$:

$$\sqrt{Var[\bar{X}]} = \frac{\sigma}{\sqrt{n}}$$

- Under CLT $\rightarrow$ As $n$ goes up, distribution of $\bar{X}$ approaches:
Central Limit Theorem

So, using CLT for $\bar{X}$

- $E[\bar{X}] = \mu$
- For $Var[\bar{X}]$:

$$Var[\bar{X}] = Var\left[\frac{X_1 + X_2 + \ldots + X_n}{n}\right] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

- (See proof in appendix)
- And for $SD[\bar{X}]$:

$$\sqrt{Var[\bar{X}]} = \sigma/\sqrt{n}$$

- Under CLT $\rightarrow$ As $n$ goes up, distribution of $\bar{X}$ approaches:

$$\bar{X} \sim N(\mu, \sigma^2/n)$$
Central Limit Theorem

Two issues:

- We need a large sample size
- Sample sizes are often small

- $n > 30$ provides a good rule of thumb that CLT has kicked in
- A larger sample size required if the original distribution has skewness and/or non-normality (such as outliers)

- We need to know $\mu$ and $\sigma$
  - For $\mu$: We calculate $E[\bar{X}]$ and estimate it using the sample mean
  - For $\sigma$: We estimate using the sample standard deviation, $s$:
    
    $s / \sqrt{n}$

- This is known as the standard error
Central Limit Theorem

Two issues:

1. We need a large sample size, and sample sizes often small.
2. We’ll discuss later what happens in small sample sizes.

- $n > 30$ provides a good rule of thumb that the CLT has kicked in.
- A larger sample size is required if the original distribution has skewness and/or non-normality (such as outliers).

- We need to know $\mu$ and $\sigma$—the true mean and true standard deviation.
- For $\mu$: We calculate $E[\bar{X}]$ and estimate it using the sample mean.
- For $\sigma$: We estimate using sample standard deviation, $s$: $s = \frac{s}{\sqrt{n}}$.

This is known as the standard error.
Central Limit Theorem

Two issues:

- We need a large sample size → and sample sizes often small

- $n > 30$ provides good rule of thumb that CLT has kicked in

- A larger sample size required if original distribution has skewness and/or non-normality (such as outliers)

- We need to know $\mu$ and $\sigma$ → the true means and true standard deviation

- For $\mu$: We calculate $E[\bar{X}]$ and estimate it using sample mean

- For $\sigma$: We estimate using sample standard deviation, $s$:

$$s = \frac{s}{\sqrt{n}}$$

- This is known as the standard error
Two issues:

- We need a large sample size → and sample sizes often small
  - We’ll discuss later what happens in small sample sizes
Central Limit Theorem

Two issues:

▶ We need a large sample size → and sample sizes often small
  ▶ We’ll discuss later what happens in small sample sizes
  ▶ \( n > 30 \) provides good rule of thumb that CLT has kicked in

▶ A larger sample size required if original distribution has skewness and/or non-normality (such as outliers)

▶ We need to know \( \mu \) and \( \sigma \) → the true means and true standard deviation

▶ For \( \mu \): We calculate \( E[\bar{X}] \) and estimate it using sample mean

▶ For \( \sigma \): We estimate using sample standard deviation, \( s \):

\[
\sigma_{\text{SE}} = \frac{s}{\sqrt{n}}
\]

▶ This is known as the standard error
Central Limit Theorem

Two issues:

- We need a large sample size → and sample sizes often small
  - We’ll discuss later what happens in small sample sizes
  - \( n > 30 \) provides good rule of thumb that CLT has kicked in
  - A larger sample size required if original distribution has skewness and/or non-normality (such as outliers)

\[ E[\bar{X}] \] and \( \sigma \) → the true means and true standard deviation

\[ s = \frac{s}{\sqrt{n}} \]

This is known as the standard error
Central Limit Theorem

Two issues:

- We need a large sample size $\rightarrow$ and sample sizes often small
  - We’ll discuss later what happens in small sample sizes
  - $n > 30$ provides good rule of thumb that CLT has kicked in
  - A larger sample size required if original distribution has skewness and/or non-normality (such as outliers)

- We need to know $\mu$ and $\sigma$ $\rightarrow$ the true means and true standard deviation
Central Limit Theorem

Two issues:

▶ We need a large sample size → and sample sizes often small
  ▶ We’ll discuss later what happens in small sample sizes
  ▶ \( n > 30 \) provides good rule of thumb that CLT has kicked in
  ▶ A larger sample size required if original distribution has skewness and/or non-normality (such as outliers)

▶ We need to know \( \mu \) and \( \sigma \) → the true means and true standard deviation
  ▶ For \( \mu \): We calculate \( E[\bar{X}] \) and estimate it using sample mean
Central Limit Theorem

Two issues:

- We need a large sample size → and sample sizes often small
  - We’ll discuss later what happens in small sample sizes
  - $n > 30$ provides good rule of thumb that CLT has kicked in
  - A larger sample size required if original distribution has skewness and/or non-normality (such as outliers)

- We need to know $\mu$ and $\sigma$ → the true means and true standard deviation
  - For $\mu$: We calculate $E[\bar{X}]$ and estimate it using sample mean
  - For $\sigma$: We estimate using sample standard deviation, $s$:
    
    $$ = s/\sqrt{n} $$
Central Limit Theorem

Two issues:
▶ We need a large sample size → and sample sizes often small
  ▶ We’ll discuss later what happens in small sample sizes
  ▶ $n > 30$ provides good rule of thumb that CLT has kicked in
  ▶ A larger sample size required if original distribution has skewness and/or non-normality (such as outliers)
▶ We need to know $\mu$ and $\sigma$ → the true means and true standard deviation
  ▶ For $\mu$: We calculate $E[\bar{X}]$ and estimate it using sample mean
  ▶ For $\sigma$: We estimate using sample standard deviation, $s$:

$$s/\sqrt{n}$$

▶ This is known as the standard error
Standard Error

To clarify this terminology:

- Standard deviation ($\sigma$): relates to population
- Sample standard deviation ($s$): relates to the sample that we have drawn from the population
- Standard error (SE): relates to the standard deviation for sampling distribution of our estimator

It is a commonly used measure of how spread out the sampling distribution of our estimator is

For $\bar{X}$, $SE[\bar{X}] = \frac{s}{\sqrt{n}}$
To clarify this terminology:

- **Standard deviation ($\sigma$):** relates to the population.
- **Sample standard deviation ($s$):** relates to the sample that we have drawn from the population.
- **Standard error (SE):** relates to the standard deviation for the sampling distribution of our estimator. It is a commonly used measure of how spread out the sampling distribution of our estimator is.

For $\bar{X}$, $SE[\bar{X}] = \frac{s}{\sqrt{n}}$
Standard Error

To clarify this terminology:

- **Standard deviation** (σ): relates to population

- **Sample standard deviation** (s): relates to the sample that we have drawn from the population

- **Standard error (SE)**: relates to the standard deviation for sampling distribution of our estimator
  
  It is a commonly used measure of how spread out the sampling distribution of our estimator is

For $\bar{X}$, $SE(\bar{X}) = \frac{s}{\sqrt{n}}$
Standard Error

To clarify this terminology:

- **Standard deviation \((\sigma)\):** relates to population

- **Sample standard deviation \((s)\):** Relates to the sample that we have drawn from the population
Standard Error

To clarify this terminology:

- **Standard deviation (\(\sigma\))**: relates to population
- **Sample standard deviation (\(s\))**: Relates to the sample that we have drawn from the population
- **Standard error (SE)**: Relates to the standard deviation for sampling distribution of our estimator
Standard Error

To clarify this terminology:

- **Standard deviation (σ):** relates to population
- **Sample standard deviation (s):** Relates to the sample that we have drawn from the population
- **Standard error (SE):** Relates to the standard deviation for sampling distribution of our estimator
  - It is a commonly used measure of how spread out the sampling distribution of our estimator is
Standard Error

To clarify this terminology:

- **Standard deviation (σ)**: relates to population
- **Sample standard deviation (s)**: Relates to the sample that we have drawn from the population
- **Standard error (SE)**: Relates to the standard deviation for sampling distribution of our estimator
  - It is a commonly used measure of how spread out the sampling distribution of our estimator is
  - For $\bar{X}$, $SE[\bar{X}] = s/\sqrt{n}$
Central Limit Theorem Example

Want to understand body weights in Uganda

Weights of Ugandan women $\sim N(135, 20^2)$

But suppose don't know this (perhaps don't have enough research $)$

Do have enough to analyze weights of samples drawn from the population
Central Limit Theorem Example

- Want to understand body weights in Uganda
Central Limit Theorem Example

- Want to understand body weights in Uganda
- Weights of Ugandan women \( \sim N(135, 20^2) \)
Central Limit Theorem Example

- Want to understand body weights in Uganda
- Weights of Ugandan women $\sim N(135, 20^2)$
- But suppose don’t know this (perhaps don’t have enough research $\$$)
Central Limit Theorem Example

- Want to understand body weights in Uganda
- Weights of Ugandan women $\sim N(135, 20^2)$
- But suppose don’t know this (perhaps don’t have enough research $\$$)
- Do have enough to analyze weights of *samples* drawn from the population
Central Limit Theorem Example
Central Limit Theorem Example

Population (truth) looks like this:
Central Limit Theorem Example

Population (truth) looks like this:

![Distribution of Weight of Ugandan Women](image_url)
Central Limit Theorem Example

Take 500 samples of $n = 2$, average, and plot:
Central Limit Theorem Example

Take 500 samples of $n = 2$, average, and plot:

Distribution of the Sample Mean

Pounds

Density

0.00 0.05 0.10 0.15 0.20

100 120 140 160 180

Pounds
Central Limit Theorem Example

Take 500 samples of $n = 10$, average, and plot:
Central Limit Theorem Example

Take 500 samples of $n = 30$, average, and plot:
Central Limit Theorem Example

Take 500 samples of $n = 60$, average, and plot:
Central Limit Theorem Example

Take 500 samples of \( n = 100 \), average, and plot:
Central Limit Theorem Example

Take 500 samples of $n = 500$, average, and plot:
Central Limit Theorem Example

Take 500 samples of $n = 1000$, average, and plot:

![Distribution of the Sample Mean](image)

- Pounds
- Density
- 100 120 140 160 180
- 0.0 0.1 0.2 0.3 0.4 0.5 0.6
Central Limit Theorem Example

Take 500 samples of \( n = 10000 \), average, and plot:
Central Limit Theorem Example

Does not matter on the number of samples:
Ex) 500 samples of $n = 60$
Central Limit Theorem Example

Does not matter on the number of samples:
Ex) 250 samples of \( n = 60 \)
Central Limit Theorem Example

Does not matter on the number of samples:
Ex) 1000 samples of $n = 60$
Central Limit Theorem

This example used a Normal distribution
But Central Limit Theorem amazingly also holds for any other kind of distribution!
That is, $X_1, X_2, \ldots, X_n$ can be drawn from any kind of distribution
→ $\bar{X}$ will still approach normal as sample size $n$ goes up
(Again, big reason why Normal distribution so important!)
Central Limit Theorem

- This example used a Normal distribution

But Central Limit Theorem amazingly also holds for any other kind of distribution! That is, $X_1, X_2, \ldots, X_n$ can be drawn from any kind of distribution includes discrete distributions, includes non-normal continuous distributions. $\bar{X}$ will still approach normal as sample size $n$ goes up. (Again, big reason why Normal distribution so important!)
Central Limit Theorem

- This example used a Normal distribution
- But Central Limit Theorem amazingly also holds for any other kind of distribution!
Central Limit Theorem

- This example used a Normal distribution
- But Central Limit Theorem amazingly also holds for any other kind of distribution!
- That is, $X_1, X_2, ..., X_3$ can be drawn from any kind of distribution
Central Limit Theorem

- This example used a Normal distribution
- But Central Limit Theorem amazingly also holds for any other kind of distribution!
- That is, $X_1, X_2, \ldots, X_3$ can be drawn from any kind of distribution
  - Includes discrete distributions
  - Includes non-normal continuous distributions
  - $\bar{X}$ will still approach normal as sample size $n$ goes up
  - (Again, big reason why Normal distribution so important!)
Central Limit Theorem

- This example used a Normal distribution
- But Central Limit Theorem amazingly also holds for any other kind of distribution!
- That is, $X_1, X_2, ..., X_3$ can be drawn from any kind of distribution
  - Includes discrete distributions
  - Includes non-normal continuous distributions

$\bar{X}$ will still approach normal as sample size $n$ goes up

(Again, big reason why Normal distribution so important!)
Central Limit Theorem

- This example used a Normal distribution
- But Central Limit Theorem amazingly also holds for any other kind of distribution!
- That is, $X_1, X_2, ..., X_3$ can be drawn from any kind of distribution
  - Includes discrete distributions
  - Includes non-normal continuous distributions
- $\bar{X}$ will still approach normal as sample size $n$ goes up
Central Limit Theorem

- This example used a Normal distribution
- But Central Limit Theorem amazingly also holds for any other kind of distribution!
- That is, $X_1, X_2, ..., X_3$ can be drawn from any kind of distribution
  - Includes discrete distributions
  - Includes non-normal continuous distributions
- $\bar{X}$ will still approach normal as sample size $n$ goes up
- (Again, big reason why Normal distribution so important!)
Central Limit Theorem Example

We are interested in the true proportion (share) of the US population that turns out to vote.

For each person, $X$ is a random variable (RV) that follows a Bernoulli process.

$X$ takes on two values, 1 (vote) or 0 (no vote).

Again, don't have enough money to ask all voters if they voted.

Instead, analyze a sample of voters.

Note: Though we are interested in proportions, we will use the formula for sample mean:

$$\bar{X} = \frac{\sum x_i}{n}$$

For example, if $n = 1$, then $\bar{X}$ must equal? (0 or 1)
Central Limit Theorem Example

- Proportion example

We are interested in true proportion (share) of US population that turns out to vote.

For each person, $X$ is a RV that follows a Bernoulli process.

$X$ takes on two values, 1 (vote) or 0 (no vote).

Again, don’t have enough $ to ask all voters if they voted.

Instead, analyze sample of voters.

Note: Though we are interested in proportions, we will use formula for sample mean:

$$ \bar{X} = \frac{\sum x_i}{n} $$

For example, if $n = 1$, then $\bar{X}$ must equal? (0 or 1)
Central Limit Theorem Example

- Proportion example
- We are interested in true proportion (share) of US population that turns out to vote
Central Limit Theorem Example

- Proportion example
- We are interested in true proportion (share) of US population that turns out to vote
- For each person, $X$ is a RV that follows a Bernoulli process
Central Limit Theorem Example

- Proportion example
- We are interested in true proportion (share) of US population that turns out to vote
- For each person, $X$ is a RV that follows a Bernoulli process
- $X$ takes on two values, 1 (vote) or 0 (no vote)
Central Limit Theorem Example

- Proportion example
- We are interested in true proportion (share) of US population that turns out to vote
- For each person, $X$ is a RV that follows a Bernoulli process
- $X$ takes on two values, 1 (vote) or 0 (no vote)
- Again, don’t have enough $ to ask all voters if they voted
Central Limit Theorem Example

- Proportion example
- We are interested in true proportion (share) of US population that turns out to vote
- For each person, $X$ is a RV that follows a Bernoulli process
- $X$ takes on two values, 1 (vote) or 0 (no vote)
- Again, don’t have enough $ to ask all voters if they voted
- Instead, analyze sample of voters
Central Limit Theorem Example

- Proportion example
- We are interested in true proportion (share) of US population that turns out to vote
- For each person, $X$ is a RV that follows a Bernoulli process
- $X$ takes on two values, 1 (vote) or 0 (no vote)
- Again, don’t have enough $ to ask all voters if they voted
- Instead, analyze sample of voters
- Note: Though we are interested in proportions, we will use formula for sample mean:
Central Limit Theorem Example

- Proportion example
- We are interested in true proportion (share) of US population that turns out to vote
- For each person, $X$ is a RV that follows a Bernoulli process
- $X$ takes on two values, 1 (vote) or 0 (no vote)
- Again, don’t have enough $ to ask all voters if they voted
- Instead, analyze sample of voters
- Note: Though we are interested in proportions, we will use formula for sample mean:

$$\bar{X} = \frac{\sum x_i}{n}$$
Central Limit Theorem Example

- Proportion example
- We are interested in true proportion (share) of US population that turns out to vote
- For each person, $X$ is a RV that follows a Bernoulli process
- $X$ takes on two values, 1 (vote) or 0 (no vote)
- Again, don’t have enough $ to ask all voters if they voted
- Instead, analyze sample of voters
- Note: Though we are interested in proportions, we will use formula for sample mean:

$$\bar{X} = \frac{\sum x_i}{n}$$

- For example, if $n = 1$, then $\bar{X}$ must equal? (0 or 1)
Central Limit Theorem Example
Central Limit Theorem Example

Take 500 samples of $n = 1$ individuals, average, and plot:
Central Limit Theorem Example

Take 500 samples of $n = 1$ individuals, average, and plot:
Central Limit Theorem Example

Take 500 samples of $n = 5$ individuals, average, and plot:

![Distribution of the Sample Mean](image_url)
Central Limit Theorem Example

Take 500 samples of \( n = 10 \) individuals, average, and plot:

![Distribution of the Sample Mean](image)
Central Limit Theorem Example

Take 500 samples of $n = 50$ individuals, average, and plot:

![Distribution of the Sample Mean](image)
Central Limit Theorem Example

Take 500 samples of $n = 100$ individuals, average, and plot:

Distribution of the Sample Mean

Sample Means

Density
Central Limit Theorem Example

Take 500 samples of $n = 500$ individuals, average, and plot:

![Distribution of the Sample Mean](image-url)
Central Limit Theorem Example

Take 500 samples of $n = 1000$ individuals, average, and plot:
Central Limit Theorem

Voter turnout example shows that CLT applies more generally to sample proportions.

Extremely helpful in variety of contexts in which the underlying distribution is Bernoulli (0 or 1 outcome):

- Voter turnout
- Share of population that is female
- Share of population that has some trait (or disease)
- Ex) Share of smokers
Central Limit Theorem

- Voter turnout example shows that CLT applies more generally to sample proportions
Central Limit Theorem

- Voter turnout example shows that CLT applies more generally to sample proportions
- Extremely helpful in variety of contexts in which the underlying distribution is Bernoulli (0 or 1 outcome):
Central Limit Theorem

- Voter turnout example shows that CLT applies more generally to sample proportions
- Extremely helpful in variety of contexts in which the underlying distribution is Bernoulli (0 or 1 outcome):
  - Voter turnout
Central Limit Theorem

- Voter turnout example shows that CLT applies more generally to sample proportions
- Extremely helpful in variety of contexts in which the underlying distribution is Bernoulli (0 or 1 outcome):
  - Voter turnout
  - Share of population that is female
Central Limit Theorem

- Voter turnout example shows that CLT applies more generally to sample proportions.
- Extremely helpful in variety of contexts in which the underlying distribution is Bernoulli (0 or 1 outcome):
  - Voter turnout
  - Share of population that is female
  - Share of population that has some trait (or disease)
Central Limit Theorem

- Voter turnout example shows that CLT applies more generally to sample proportions
- Extremely helpful in variety of contexts in which the underlying distribution is Bernoulli (0 or 1 outcome):
  - Voter turnout
  - Share of population that is female
  - Share of population that has some trait (or disease)
    - Ex) Share of smokers
Use CLT for Binomials

Note: Remember that Binomials are actually the sums of Bernoullis.

Under CLT, with a large sample, sums of random variables tend to be Normally distributed.

So: you have many trials in the binomial distribution, what distribution should those take? Normal!

Called the Normal approximation for the binomial.

Good b/c calculating things using binomial PMF tedious with large $n$. 
Use CLT for Binomials

▶ Note: Remember that Binomials are actually the sums of Bernoullis
Use CLT for Binomials

- Note: Remember that Binomials are actually the sums of Bernoullis
- Under CLT, with a large sample, sums of random variables tend to be Normally distributed
Note: Remember that Binomials are actually the sums of Bernoullis

Under CLT, with a large sample, sums of random variables tend to be Normally distributed

So: you have many trials in the binomial distribution, what distribution should those take?
Use CLT for Binomials

- Note: Remember that Binomials are actually the sums of Bernoullis
- Under CLT, with a large sample, sums of random variables tend to be Normally distributed
- So: you have many trials in the binomial distribution, what distribution should those take? Normal!
Use CLT for Binomials

- Note: Remember that Binomials are actually the sums of Bernoullis
- Under CLT, with a large sample, sums of random variables tend to be Normally distributed
- So: you have many trials in the binomial distribution, what distribution should those take? Normal!
- Called the Normal approximation for the binomial
Use CLT for Binomials

- Note: Remember that Binomials are actually the sums of Bernoullis
- Under CLT, with a large sample, sums of random variables tend to be Normally distributed
- So: you have many trials in the binomial distribution, what distribution should those take? Normal!
- Called the Normal approximation for the binomial
- Good b/c calculating things using binomial PMF tedious with large $n$
Normal Approximation for the Binomial
Normal Approximation for the Binomial

For $n = 25$, $p = 0.5$.

The diagram shows the frequency distribution for $n = 25$, $p = 0.5$. The x-axis represents the number of successes, ranging from 5 to 20, and the y-axis represents the frequency, ranging from 0 to 300. The graph indicates the expected distribution under the normal approximation.
Normal Approximation for the Binomial

$n = 30, p = 0.5$

Frequency

0 50 100 150 200 250

$n = 30, p = 0.5$

Frequency

0 50 100 150 200 250
Normal Approximation for the Binomial

$n = 1000$, $p = 0.5$

Frequency

<table>
<thead>
<tr>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
<tr>
<td>50</td>
</tr>
<tr>
<td>100</td>
</tr>
<tr>
<td>150</td>
</tr>
<tr>
<td>200</td>
</tr>
<tr>
<td>250</td>
</tr>
</tbody>
</table>

$n = 1000$, $p = 0.5$
Normal Approximation for the Binomial

$n = 10000, p = 0.5$

Frequency

<table>
<thead>
<tr>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
<tr>
<td>50</td>
</tr>
<tr>
<td>100</td>
</tr>
<tr>
<td>150</td>
</tr>
</tbody>
</table>

$n = 10000, p = 0.5$

Frequency

<table>
<thead>
<tr>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
<tr>
<td>50</td>
</tr>
<tr>
<td>100</td>
</tr>
<tr>
<td>150</td>
</tr>
</tbody>
</table>
Normal Approximation for the Binomial

In large samples:

\[ X \sim \text{Bin}(n, p) \approx Y \sim N(np, np(1-p)) \]

- Small note: Binomial discrete, while normal continuous
- Often include continuity correction by adding or subtracting 0.5 from each discrete \( x \) value

Ex) if \( P(X > 6) \), then ask \( P(Y > 6.5) \)

Ex) if \( P(X \leq 6) \), then ask \( P(Y < 6.5) \)
Normal Approximation for the Binomial

- In large samples:

\[ X \sim \text{Bin}(n, p) \approx Y \sim N(np, np(1-p)) \]

- Small note: Binomial discrete, while normal continuous

- Often include continuity correction by adding or subtracting 0.5 from each discrete \( x \) value

Ex) if \( P(X > 6) \), then ask \( P(Y > 6.5) \)

Ex) if \( P(X \leq 6) \), then ask \( P(Y < 6.5) \)
Normal Approximation for the Binomial

In large samples:

\[ X \sim Bin(n, p) \]
\[ \approx Y \sim N(np, np(1 - p)) \]
Normal Approximation for the Binomial

- In large samples:

  \[ X \sim Bin(n, p) \]
  \[ \approx Y \sim N(np, np(1 - p)) \]

- Small note: Binomial discrete, while normal continuous
Normal Approximation for the Binomial

- In large samples:
  
  \[ X \sim Bin(n, p) \]
  \[ \approx Y \sim N(np, np(1 - p)) \]

- Small note: Binomial discrete, while normal continuous
- Often include continuity correction by adding or subtracting 0.5 from each discrete \( x \) value
Normal Approximation for the Binomial

▶ In large samples:

\[ X \sim Bin(n, p) \]
\[ \approx Y \sim N(np, np(1 - p)) \]

▶ Small note: Binomial discrete, while normal continuous

▶ Often include **continuity correction** by adding or subtracting 0.5 from each discrete \( x \) value

▶ Ex) if \( P(X > 6) \), then ask \( P(Y > 6.5) \)
Normal Approximation for the Binomial

In large samples:

\[ X \sim Bin(n, p) \]
\[ \approx Y \sim N(np, np(1 - p)) \]

Small note: Binomial discrete, while normal continuous

Often include continuity correction by adding or subtracting 0.5 from each discrete \( x \) value

Ex) if \( P(X > 6) \), then ask \( P(Y > 6.5) \)

Ex) if \( P(X \leq 6) \), then ask \( P(Y < 6.5) \)
Central Limit Theorem

Again, CLT works for $x_1, x_2, \ldots, x_k$ drawn from any kind of underlying distribution.

Note: If your individual observations come from a normal distribution, then sample mean has an exact (rather than approximate) normal distribution.

Why? Linear combination of normally distributed random variables is also normally distributed.

Also: A sample mean is simply a linear combination of individual observations.
Central Limit Theorem

- Again, CLT works for $x_1, x_2, \ldots, x_k$ drawn from any kind of underlying distribution.
Central Limit Theorem

- Again, CLT works for $x_1, x_2, ..., x_k$ drawn from any kind of underlying distribution
- Note: If your individual observations come from a normal distribution, then sample mean has an exact (rather than approximate) normal distribution
Central Limit Theorem

▶ Again, CLT works for $x_1, x_2, ..., x_k$ drawn from any kind of underlying distribution

▶ Note: If your individual observations come from a normal distribution, then sample mean has an exact (rather than approximate) normal distribution

▶ Why? Linear combination of normally distributed random variables is also normally distributed
Central Limit Theorem

- Again, CLT works for $x_1, x_2, ..., x_k$ drawn from any kind of underlying distribution.
- Note: If your individual observations come from a normal distribution, then sample mean has an exact (rather than approximate) normal distribution.
- Why? Linear combination of normally distributed random variables is also normally distributed.
- Also: A sample mean is simply a linear combination of individual observations.
Central Limit Theorem

If in the real world we only have access to one sample, how do we use the CLT to understand what one sample tells us about the world?
Central Limit Theorem

- If in the real world we only have access to one sample
Central Limit Theorem

- If in the real world we only have access to one sample
- How do we use the CLT to understand what one sample tells us about the world?
CLT in practice

Question: “In general, do you believe that the country is on the right track?”

In September, poll of 1,000 Americans:

- 27% say “Yes” (1)
- 73% say “No” (0)

We don’t care about these 1,000 people per se → we care about what these 1,000 people tell us about the truth in the population.

We use these 1,000 people in tandem with CLT to address our question.
CLT in practice

- Question: “In general, do you believe that the country is on the right track?”
Question: “In general, do you believe that the country is on the right track?”
In September, poll of 1,000 Americans:
CLT in practice

- Question: “In general, do you believe that the country is on the right track?”
- In September, poll of 1,000 Americans:
  - 27% say “Yes” (1)
CLT in practice

- Question: “In general, do you believe that the country is on the right track?”
- In September, poll of 1,000 Americans:
  - 27% say “Yes” (1)
  - 73% say “No” (0)
Question: “In general, do you believe that the country is on the right track?”

In September, poll of 1,000 Americans:
- 27% say “Yes” (1)
- 73% say “No” (0)

We don’t care about these 1,000 people per se → we care about what these 1,000 people tell us about the truth in the population
Question: “In general, do you believe that the country is on the right track?”

In September, poll of 1,000 Americans:
- 27% say “Yes” (1)
- 73% say “No” (0)

We don’t care about these 1,000 people per se → we care about what these 1,000 people tell us about the truth in the population

We use these 1,000 people in tandem with CLT to address our question
CLT in practice

Start by calculating the sample mean ($\bar{x}$):

$$\bar{x} = \frac{\sum X_i}{n} = \frac{270 \times 1 + 730 \times 0}{1000} = 0.27$$
CLT in practice

- Start by calculating the sample mean ($\bar{x}$):
Start by calculating the sample mean ($\bar{x}$):

$$\bar{x} = \frac{\sum X_i}{n}$$

$$= \frac{\sum (270 \times 1) + (730 \times 0)}{1000}$$

$$= 0.27$$
And then calculate the sample standard deviation (s):

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{n-1}}$$

$$= \sqrt{\frac{270(1 - 0.27)^2 + 730(0 - 0.27)^2}{999}}$$

$$\approx 0.444$$
CLT in practice

- And then calculate the sample standard deviation ($s$):

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}} = \sqrt{\frac{270(1 - 0.27)^2 + 730(0 - 0.27)^2}{999}} \approx 0.444$$
And then calculate the sample standard deviation ($s$):

\[
s = \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}} = \sqrt{\frac{270(1 - 0.27)^2 + 730(0 - 0.27)^2}{999}} = \sqrt{\frac{197.1}{999}} \approx 0.444
\]
CLT in practice

What can we say about the population?

With \( n = 1000 \) CLT kicks in

It would tell us that the means of these polls

\[ \bar{X} \sim N(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2) \]

Our best guess for \( \mu \) is sample mean, 0.27

Our best guess for \( \sigma \) is sample standard deviation, 0.444

Which simply leaves calculating the standard error:

\[ 0.44 \sqrt{1000} \approx 0.013 \]
So what can we say about the population?

With $n = 1000$ CLT kicks in.

It would tell us that the means of these polls
\[ \bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})^2 \]

Our best guess for $\mu$ is sample mean, 0.27.

Our best guess for $\sigma$ is sample standard deviation, 0.44.

Which simply leaves calculating the standard error:

$0.44 \sqrt{1000} \approx 0.013$
CLT in practice

- So what can we say about the population?
- With $n = 1000$ CLT kicks in

\[ \bar{X} \sim N(\mu, \sigma^2/n^{1/2}) \]

Our best guess for $\mu$ is sample mean, 0.27
Our best guess for $\sigma$ is sample standard deviation, 0.44

Which simply leaves calculating the standard error:

\[ 0.44 \sqrt{1000} \approx 0.013 \]
CLT in practice

- So what can we say about the population?
- With \( n = 1000 \) CLT kicks in
- It would tell us that the means of these polls

\[
\bar{X} \sim N(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2)
\]

- Our best guess for \( \mu \) is sample mean, 0.27
- Our best guess for \( \sigma \) is sample standard deviation, 0.444
- Which simply leaves calculating the standard error:

\[
0.44 \sqrt{1000} \approx 0.013
\]
CLT in practice

- So what can we say about the population?
- With $n = 1000$ CLT kicks in
- It would tell us that the means of these polls

\[ \bar{X} \sim N(\mu, (\sigma/\sqrt{n})^2) \]
CLT in practice

- So what can we say about the population?
- With $n = 1000$ CLT kicks in
- It would tell us that the means of these polls

$$\bar{X} \sim N(\mu, (\sigma/\sqrt{n})^2)$$

- Our best guess for $\mu$ is sample mean, 0.27
So what can we say about the population? With $n = 1000$ CLT kicks in. It would tell us that the means of these polls

$$\bar{X} \sim N(\mu, (\sigma/\sqrt{n})^2)$$

Our best guess for $\mu$ is sample mean, 0.27
Our best guess for $\sigma$ is sample standard deviation, 0.444
CLT in practice

- So what can we say about the population?
- With $n = 1000$ CLT kicks in
- It would tell us that the means of these polls

$$\bar{X} \sim N(\mu, (\sigma/\sqrt{n})^2)$$

- Our best guess for $\mu$ is sample mean, 0.27
- Our best guess for $\sigma$ is sample standard deviation, 0.444
- Which simply leaves calculating the standard error:

$$\frac{0.44}{\sqrt{1000}} \approx 0.013$$
CLT in practice

So, CLT tells us a good approximation for the sampling distribution is $\sim \mathcal{N}(0.27, 0.013^2)$

Can convert to standard normal, and then calculate all sorts of things
So, CLT tells us a good approximation for the sampling distribution is

\[ \sim N(0.27, 0.013^2) \]
So, CLT tells us a good approximation for the sampling distribution is

\[ \sim N(0.27, 0.013^2) \]

Can convert to standard normal, and then calculate all sorts of things
Next Time

- Note: Proof of CLT found here: http://mathworld.wolfram.com/CentralLimitTheorem.html
- Continue with statistical inference
- Introduce concept of Hypothesis Testing
Sampling Distribution Variance

Note that $X_1, X_2, \ldots, X_n$ are independent. Therefore:

\[
\text{var}[\bar{X}] = \text{var}\left[\frac{X_1 + X_2 + \ldots + X_n}{n}\right]
\]

\[
= \frac{1}{n^2} \text{var}[X_1 + X_2 + \ldots + X_n]
\]

\[
= \frac{1}{n^2} (\sigma^2 + \sigma^2 + \ldots + \sigma^2)
\]

\[
= \frac{1}{n^2} n\sigma^2
\]

\[
= \frac{\sigma}{n}
\]