

# CONFIDENCE INTERVALS

---

Analytics Primer

# Parameters vs. Statistics

- **Parameter**
  - Measures computed from a population.
- **Statistic**
  - Measures computed from a sample.
  - Sample statistics is the **point estimate** of the population parameter.

# Sampling Error

- When the expected value of the point estimator is equal to the population parameter, the point estimator is said to be **unbiased**.
- The absolute value of the difference between an unbiased point estimate and the corresponding population parameter is called the **sampling error**.
- Sampling error is the result of the sample being only a subset of the population.

# Point vs. Interval


- A point estimator cannot be expected to provide the exact value of the population parameter.
- The purpose of an interval estimate is to provide information about how close the point estimate is to the value of the parameter.

# Confidence Intervals

- **Confidence Intervals** are interval estimates where we say we have a certain **level of confidence** in the interval.
- For example, we are **95% confident** that the population mean is between 20 and 30.

# Confidence Intervals

- **Confidence Intervals** are interval estimates where we say we have a certain **level of confidence** in the interval.
- For example, we are **95% confident** that the population mean is between 20 and 30.



If we were to take many samples (same size) that each produced different confidence intervals, then 95% of them would contain the true parameter.

# Confidence Intervals

- **Confidence Intervals** are interval estimates where we say we have a certain **level of confidence** in the interval.
- For example, we are **95% confident** that the population mean is between 20 and 30.



95% of the time, our confidence intervals would contain the true parameter of interest.

# Confidence Intervals

- **Confidence Intervals** are interval estimates where we say we have a certain **level of confidence** in the interval.
- For example, we are **95% confident** that the population mean is between 20 and 30.

**NOT** 95% chance the population parameter falls inside our one confidence interval.



# Margin of Error

- A confidence interval can be computed by adding and subtracting a **margin of error** to the point estimate:

$$\text{Point Estimate} \pm \text{Margin of Error}$$

- A **margin of error** is the largest possible sampling error at the specified level of confidence:

$$\text{Margin of Error} = \text{Critical Value} \pm \text{Standard Error}$$

# Margin of Error

- A confidence interval can be computed by adding and subtracting a **margin of error** to the point estimate:

$$\text{Point Estimate} \pm \text{Margin of Error}$$

- A **margin of error** is the largest possible sampling error at the specified level of confidence:

$$\text{Margin of Error} = \text{Critical Value} \pm \text{Standard Error}$$

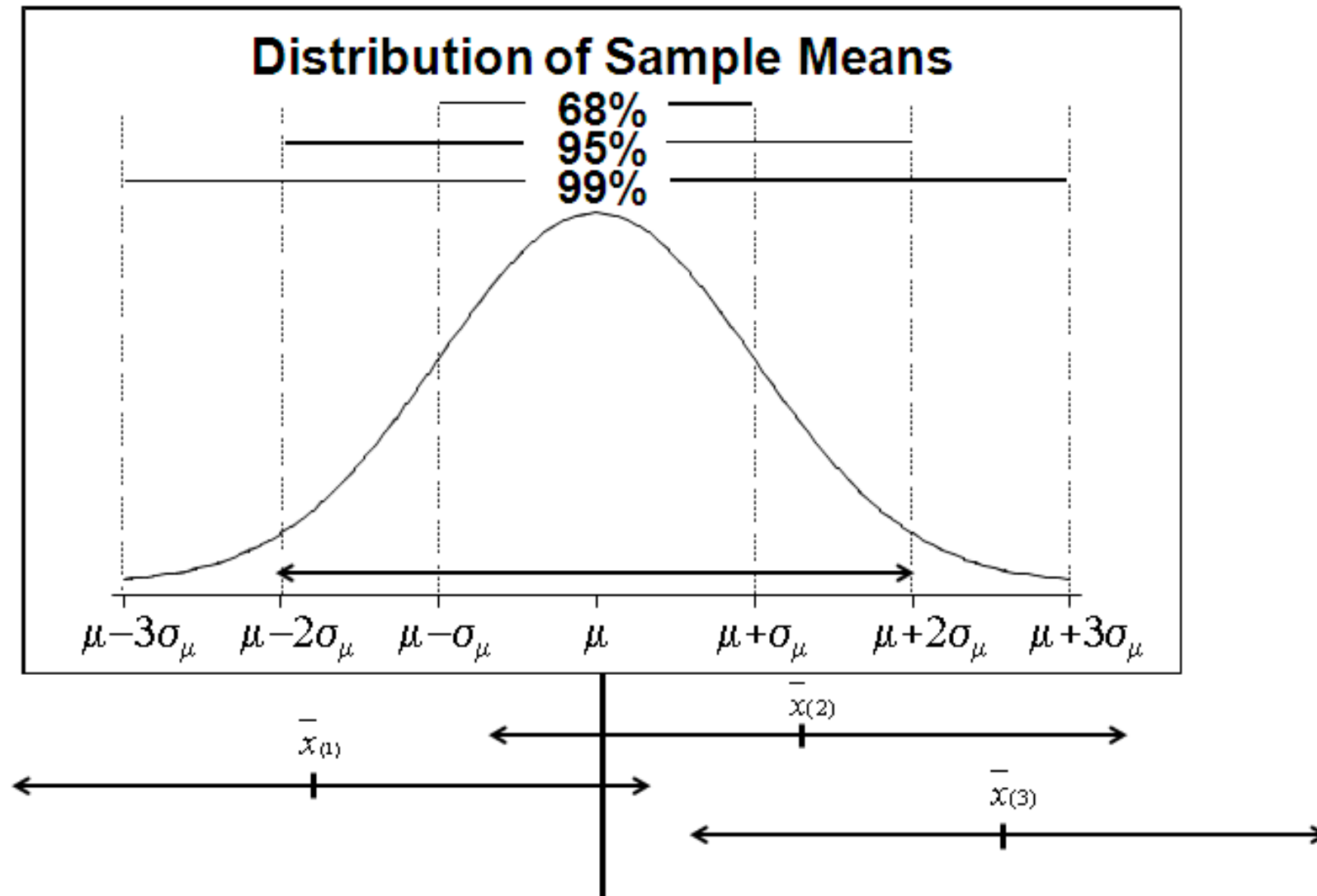


**Estimated** value of the standard deviation of the sampling error.

# Margin of Error

- The purpose of an interval estimate is to provide information about how close the point estimate is to the value of the parameter.
- This **does not mean** that your interval estimates will always contain the population parameter.

# Confidence Intervals



# INTERVAL ESTIMATION OF $\hat{p}$

---

# Margin of Error

- An **interval estimate** can be computed by adding and subtracting a **margin or error** to the point estimate:

$$\hat{p} \pm \text{Margin of Error}$$

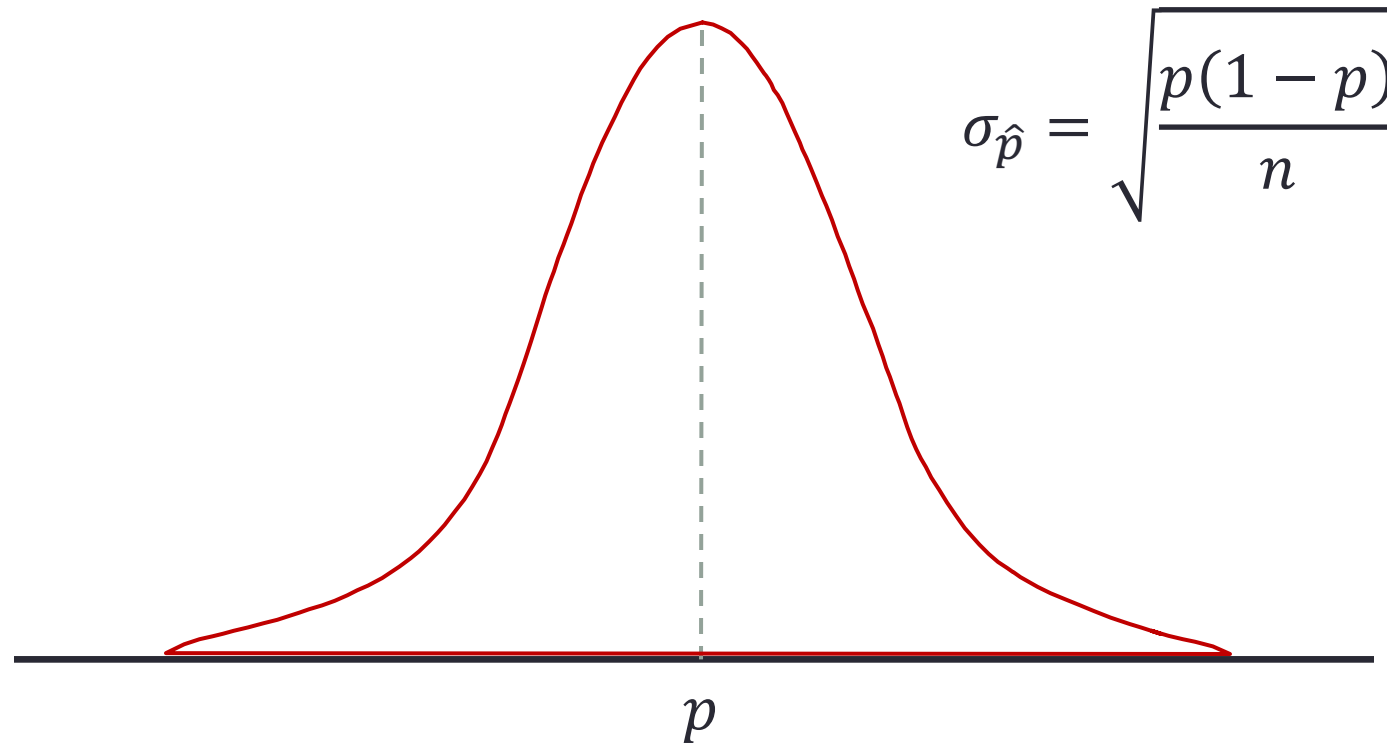
- The purpose of an interval estimate is to provide information about how close the point estimate is to the value of the parameter.

# Sampling Distribution of $\hat{p}$

- The sampling distribution of  $\hat{p}$  plays a key role in computing the margin of error for this interval estimate.
- The **sampling distribution of  $\hat{p}$**  is approximately the **Normal distribution** whenever  $np \geq 5$  and  $n(1 - p) \geq 5$ .

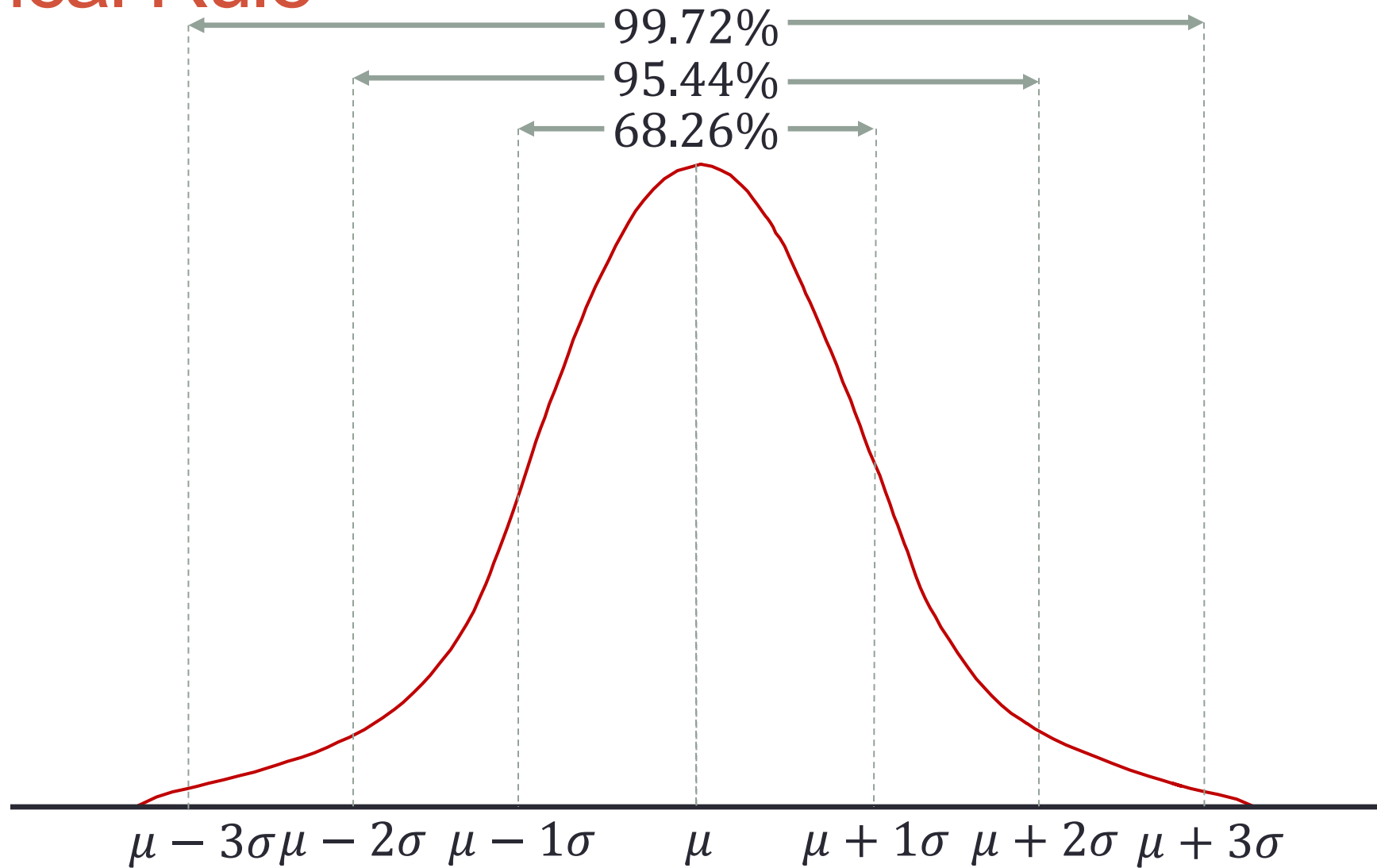
# Sampling Distribution of $\hat{p}$

- The **sampling distribution of  $\hat{p}$**  is approximately the **Normal distribution** whenever  $np \geq 5$  and  $n(1 - p) \geq 5$ .

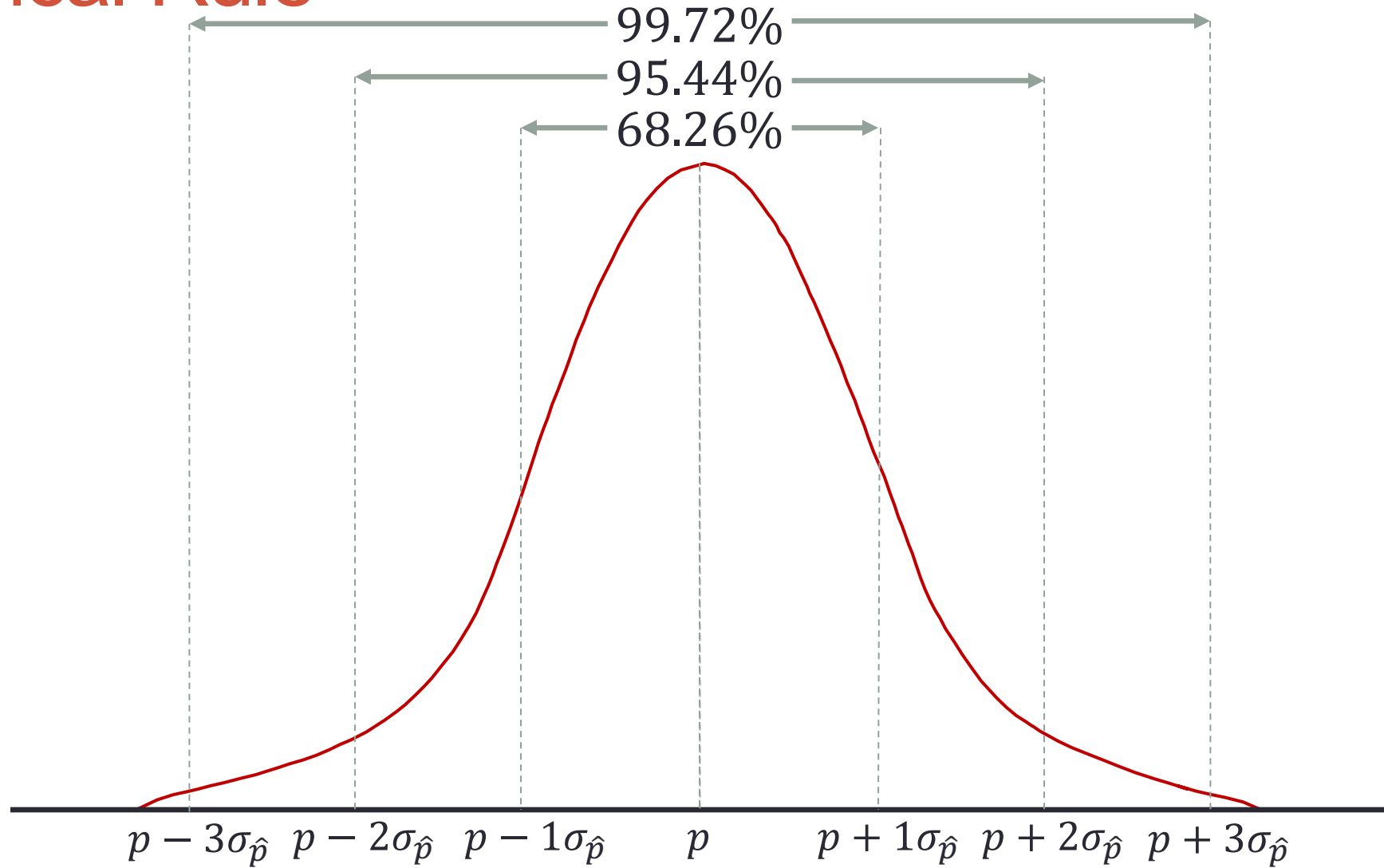




# Empirical Rule

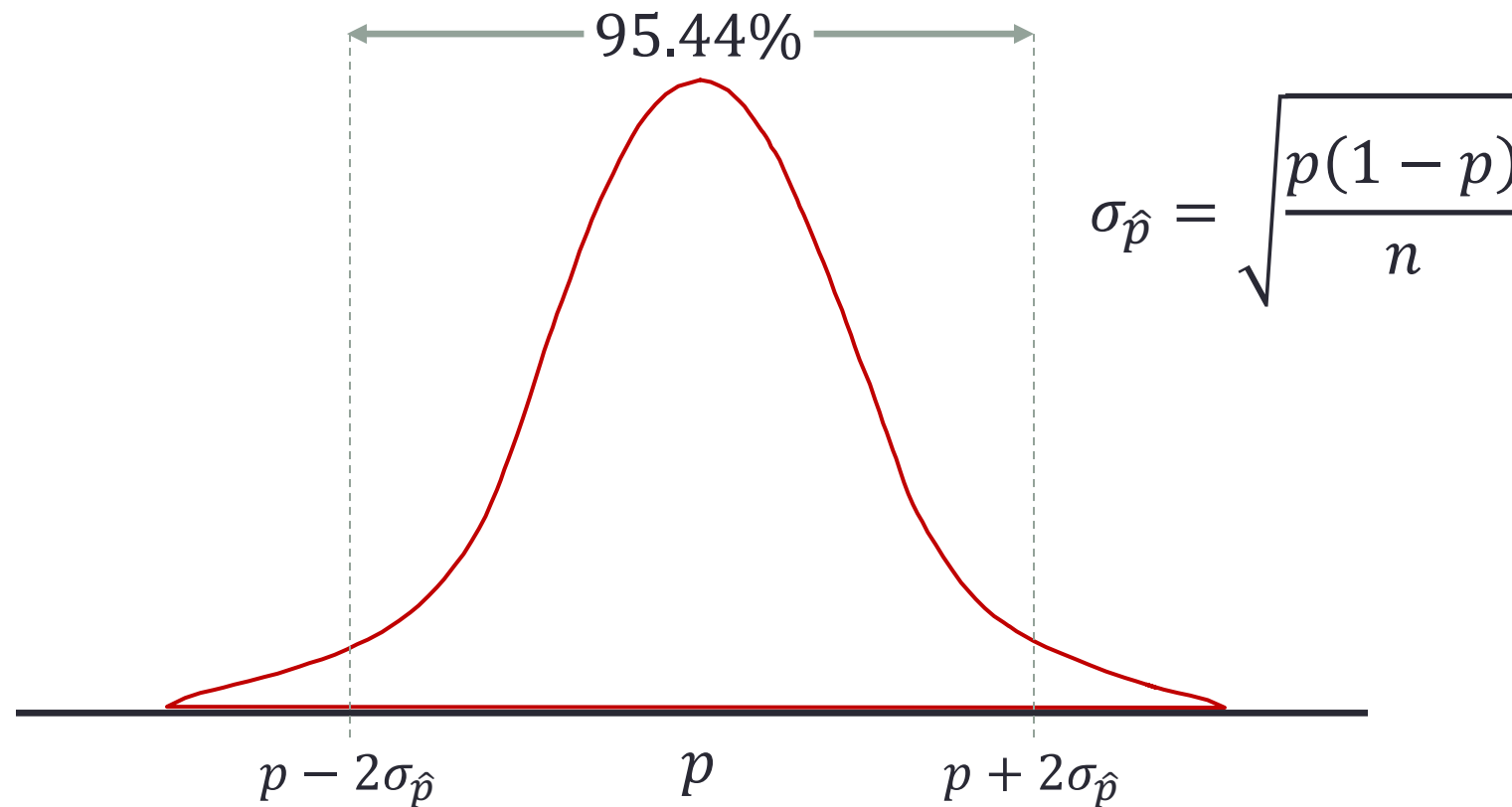


# Empirical Rule



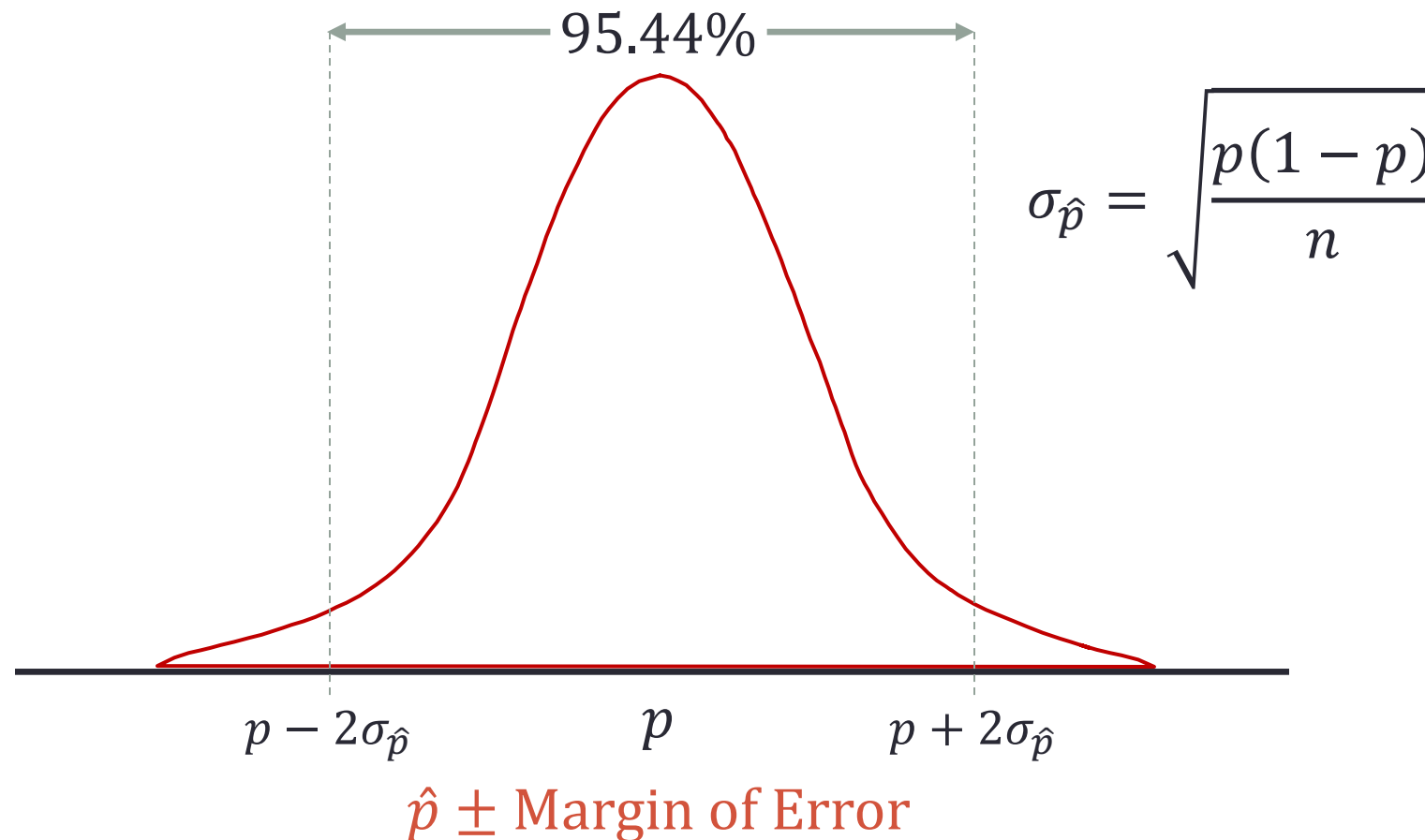
# Sampling Distribution of $\hat{p}$

- The **sampling distribution of  $\hat{p}$**  is approximately the **Normal distribution** whenever  $np \geq 5$  and  $n(1 - p) \geq 5$ .



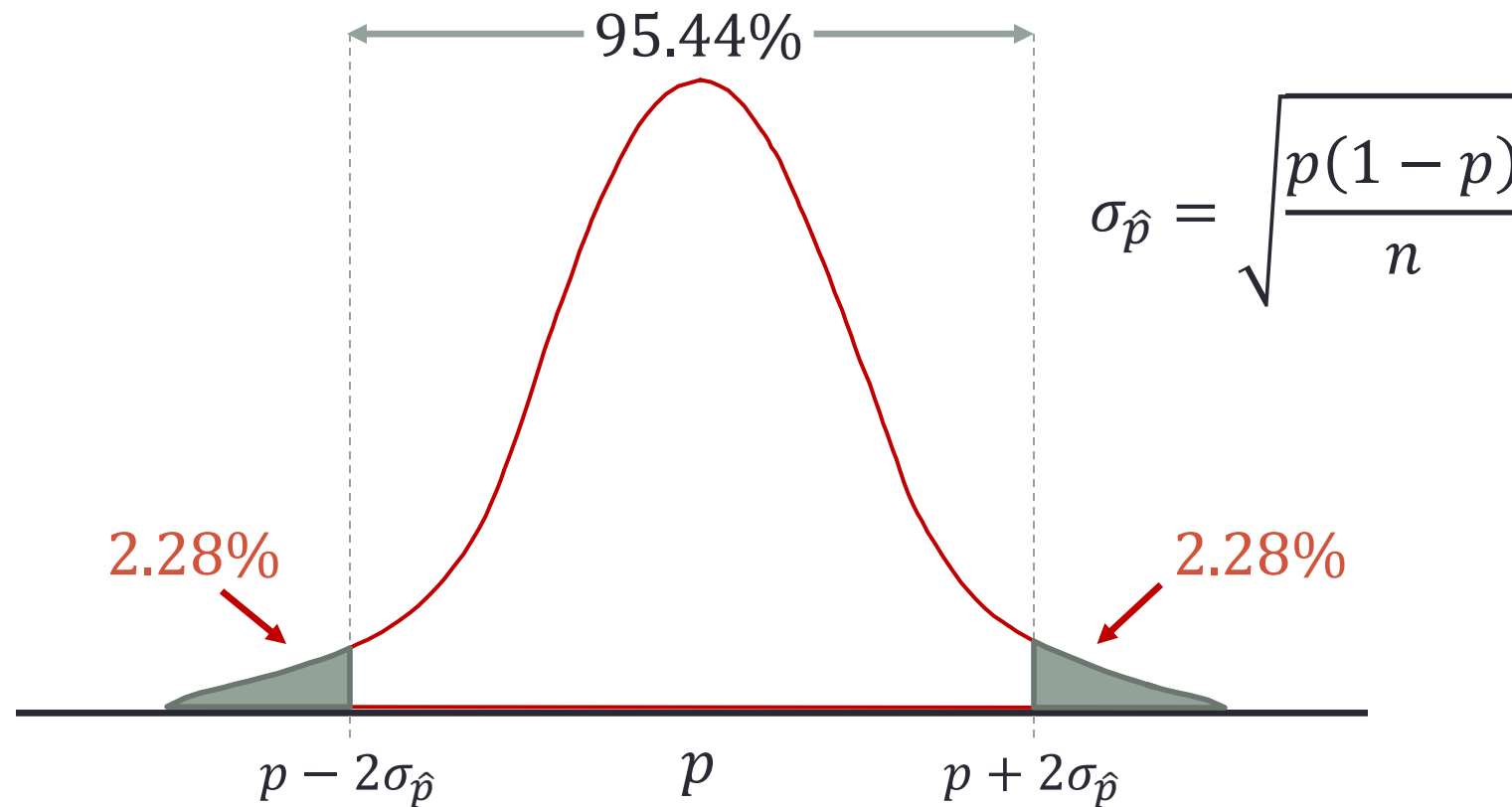
# Sampling Distribution of $\hat{p}$

- The **sampling distribution of  $\hat{p}$**  is approximately the **Normal distribution** whenever  $np \geq 5$  and  $n(1 - p) \geq 5$ .



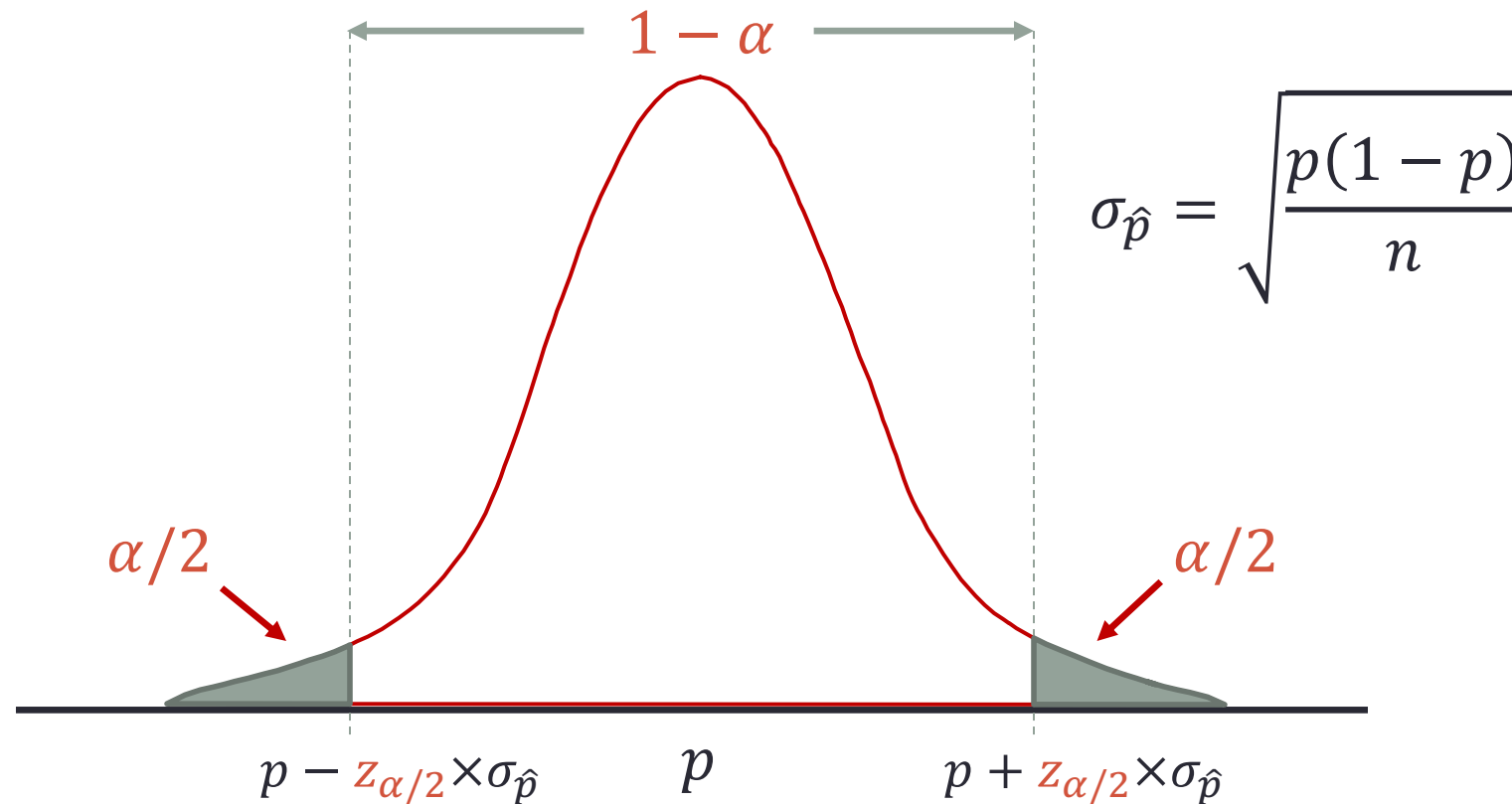
# Sampling Distribution of $\hat{p}$

- The **sampling distribution of  $\hat{p}$**  is approximately the **Normal distribution** whenever  $np \geq 5$  and  $n(1 - p) \geq 5$ .



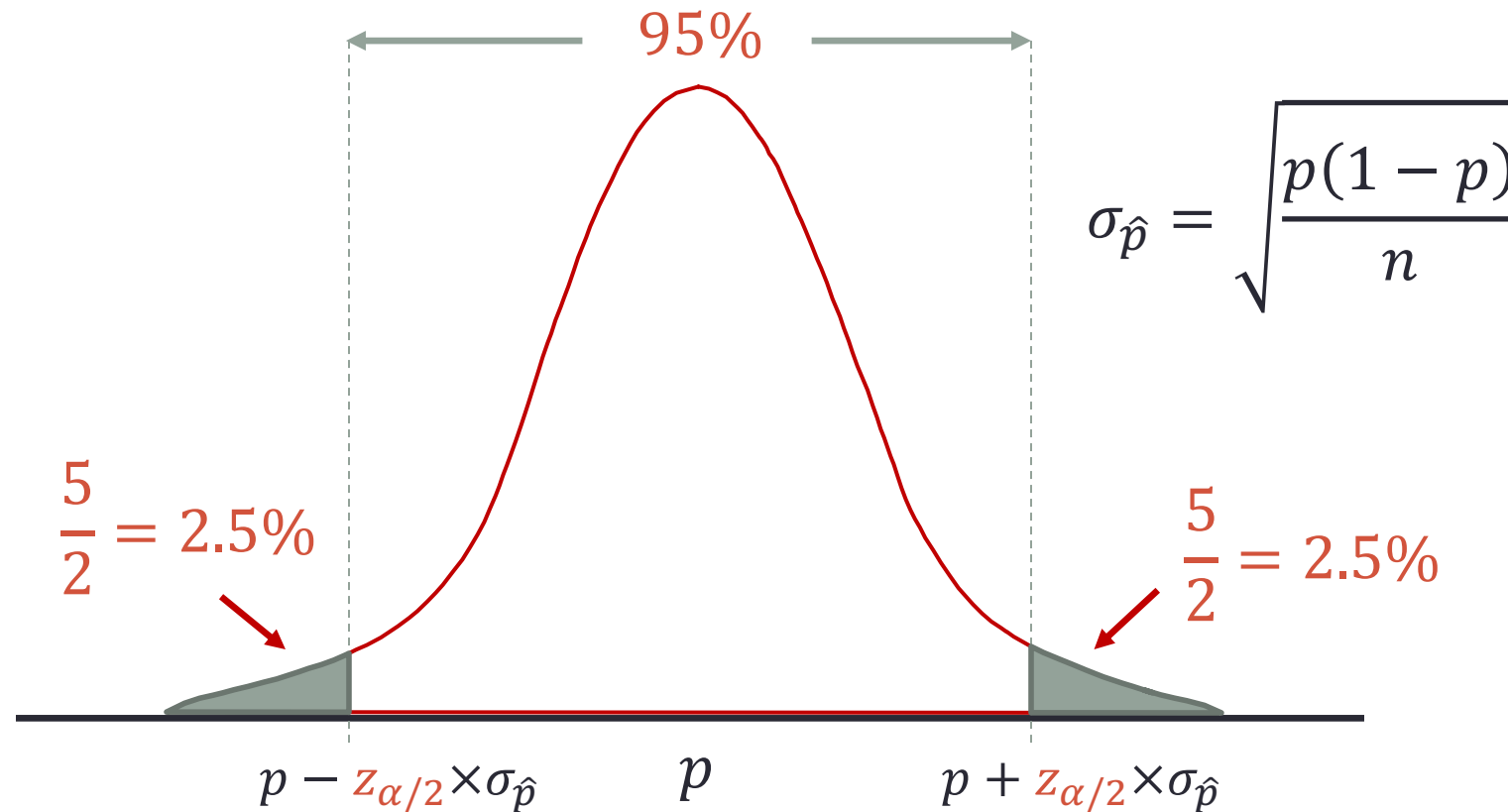
# Sampling Distribution of $\hat{p}$

- The **sampling distribution of  $\hat{p}$**  is approximately the **Normal distribution** whenever  $np \geq 5$  and  $n(1 - p) \geq 5$ .



# Sampling Distribution of $\hat{p}$

- The **sampling distribution of  $\hat{p}$**  is approximately the **Normal distribution** whenever  $np \geq 5$  and  $n(1 - p) \geq 5$ .



What is  $z_{\alpha/2}$ ?

# How to Calculate $z_{\alpha/2}$

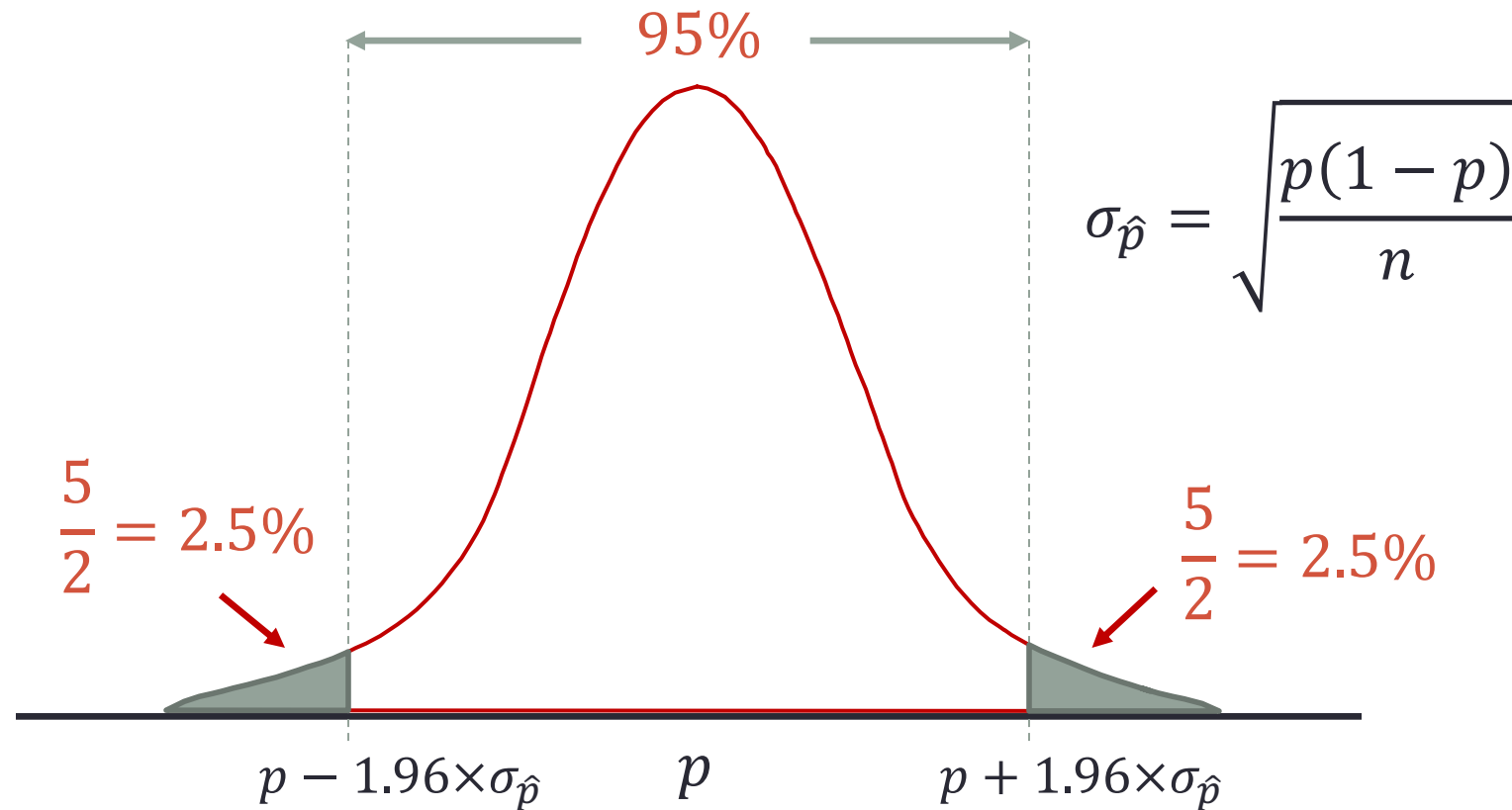
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
	.	.	.	.	.	.	.	.	.	.
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
.	.	.	.	.	.	.	.	.	.	.

$P(z \leq ?) = 0.025$



# Sampling Distribution of $\hat{p}$

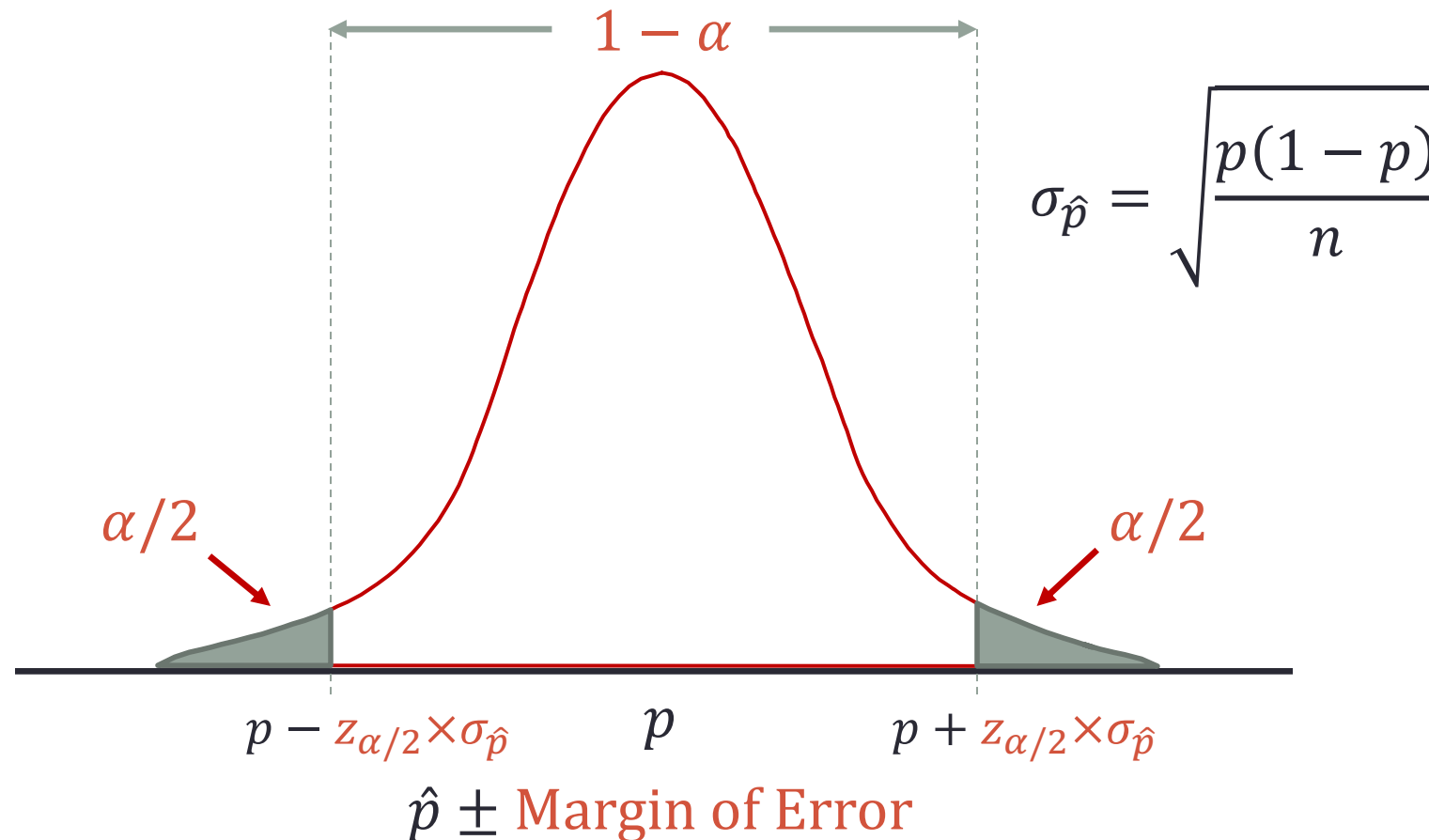
- The **sampling distribution of  $\hat{p}$**  is approximately the **Normal distribution** whenever  $np \geq 5$  and  $n(1 - p) \geq 5$ .



What is  $z_{\alpha/2}$ ?

# Sampling Distribution of $\hat{p}$

- The **sampling distribution of  $\hat{p}$**  is approximately the **Normal distribution** whenever  $np \geq 5$  and  $n(1 - p) \geq 5$ .



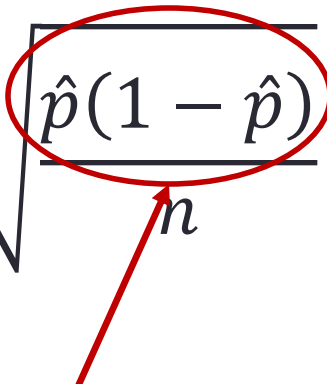
# Confidence Interval for $\hat{p}$

- The **confidence interval for  $\hat{p}$**  with a **confidence coefficient** of  $1 - \alpha$  (error of  $\alpha$ ) is the following:

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

# Confidence Interval for $\hat{p}$

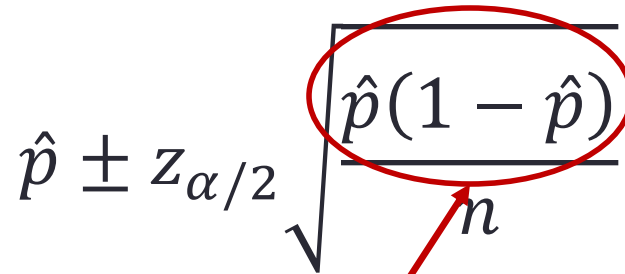
- The **confidence interval for  $\hat{p}$**  with a **confidence coefficient** of  $1 - \alpha$  (error of  $\alpha$ ) is the following:

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$


DO NOT KNOW  $p$ !

# Confidence Interval for $\hat{p}$

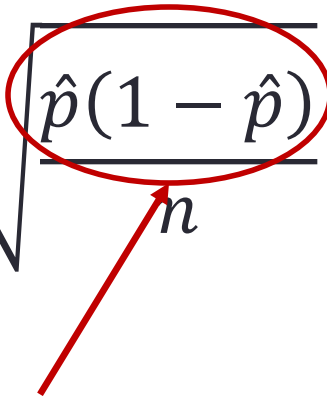
- The **confidence interval for  $\hat{p}$**  with a **confidence coefficient** of  $1 - \alpha$  (error of  $\alpha$ ) is the following:

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$


Estimate with  $\hat{p}$

# Confidence Interval for $\hat{p}$

- The **confidence interval** for  $\hat{p}$  with a **confidence coefficient** of  $1 - \alpha$  (error of  $\alpha$ ) is the following:

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$


This is why it is called  
the **standard error** and  
not the standard deviation!

# Example Confidence Interval for $\hat{p}$

- You are interested in hair color and eye color across 2 different regions of the country. You have a sample of 762 people.
- The sample has the following distribution of eye color:

Eye Color	Frequency	Percent
Blue	222	29.13%
Brown	341	44.75%
Green	199	26.12%
	762	100.00%

# Example Confidence Interval for $\hat{p}$

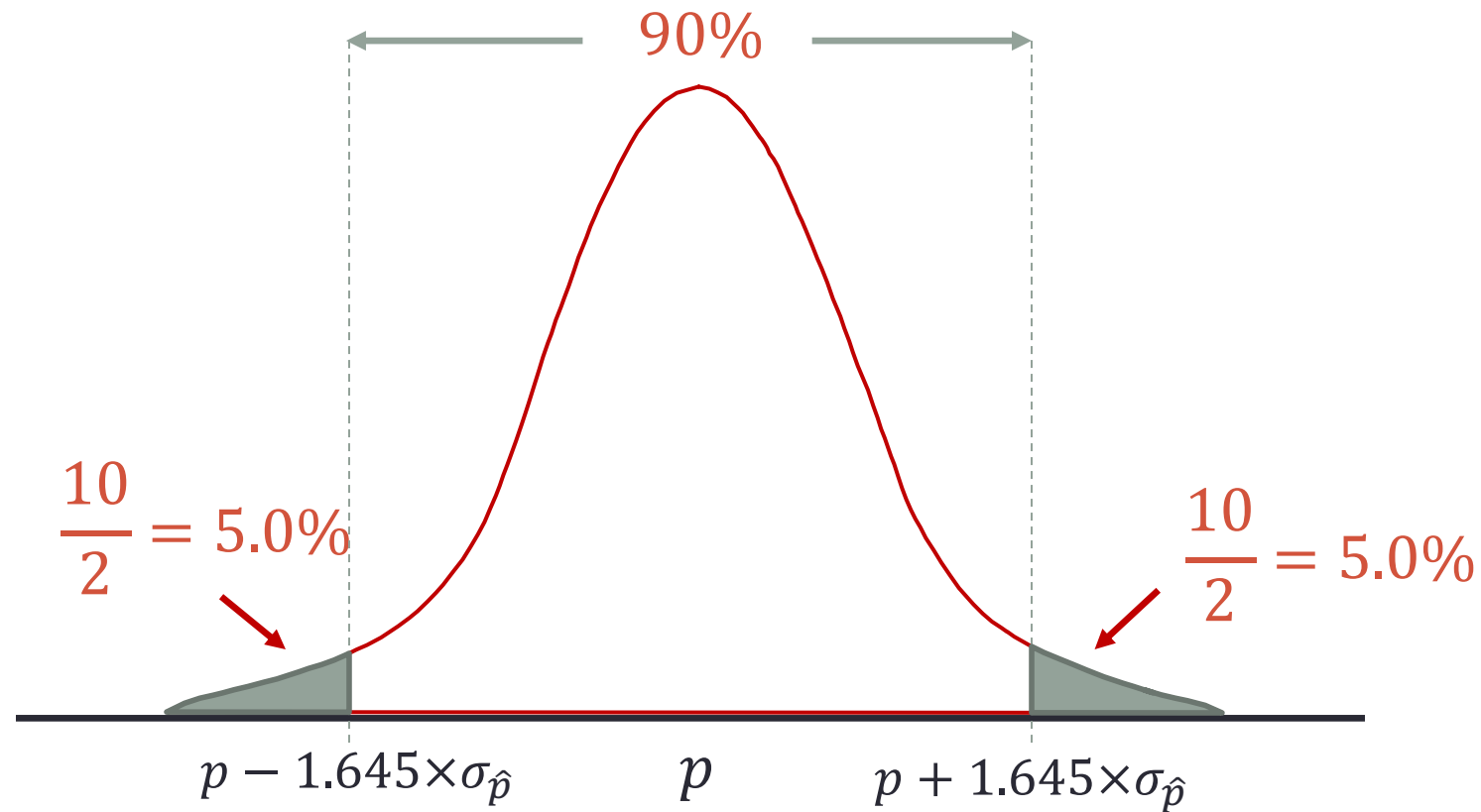
- The sample has the following distribution of eye color.
- Build a 90% Confidence Interval for the true proportion of people with blue eyes.

Eye Color	Frequency	Percent
Blue	222	29.13%
Brown	341	44.75%
Green	199	26.12%
	762	100.00%



# Example Confidence Interval for $\hat{p}$

- Build a 90% Confidence Interval for the true proportion of people with blue eyes.



What is  $z_{\alpha/2}$ ?

# Example Confidence Interval for $\hat{p}$

- Build a 90% Confidence Interval for the true proportion of people with blue eyes.

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

$$0.2913 \pm 1.645 \sqrt{\frac{0.2913(1 - 0.2913)}{762}}$$

$$0.2913 \pm 1.645 \times 0.0165$$

$$0.2913 \pm 0.027$$

# Example Confidence Interval for $\hat{p}$

- Build a 90% Confidence Interval for the true proportion of people with blue eyes.

$$0.2913 \pm 0.027$$

OR

$$(0.2643, 0.3184)$$

# Change of Things

- What happens to confidence intervals with a change of sample size?
- What happens with a change of confidence level?

# Change of Things

- What happens to confidence intervals with a change of sample size?

$$n \uparrow \Rightarrow \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \downarrow \Rightarrow z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \downarrow \Rightarrow \text{width} \downarrow$$

- What happens with a change of confidence level?

# Change of Things

- What happens to confidence intervals with a change of sample size?

$$n \uparrow \Rightarrow \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \downarrow \Rightarrow z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \downarrow \Rightarrow \text{width} \downarrow$$

- What happens with a change of confidence level?

$$C \uparrow \Rightarrow z^* \uparrow \Rightarrow z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \uparrow \Rightarrow \text{width} \uparrow$$

# Example

- An electronics manufacturer provides a full warranty on a certain type of television they make. The company will replace the television if any problems occur in the first year of use. The manager in charge of the warranty division wants to determine the proportion of warranties that are claimed. The manager samples 150 customer records and found that 17 of the customers used their warranty. Create a 95% confidence interval for the estimate of the proportion of customers who use their warranties.

# Example

- An electronics manufacturer provides a full warranty on a certain type of television they make. The company will replace the television if any problems occur in the first year of use. The manager in charge of the warranty division wants to determine the proportion of warranties that are claimed. The manager samples 150 customer records and found that 17 of the customers used their warranty. Create a 95% confidence interval for the estimate of the proportion of customers who use their warranties.

$$\hat{p} = \frac{17}{150} = 0.113$$

$$\hat{p} \pm 1.96 \times \sqrt{\frac{0.113(1 - 0.113)}{150}}$$
$$\mathbf{0.113 \pm 0.0507}$$



# INTERVAL ESTIMATION OF $\bar{x}$

---

# Margin of Error

- An **interval estimate** can be computed by adding and subtracting a **margin or error** to the point estimate:

$$\bar{x} \pm \text{Margin of Error}$$

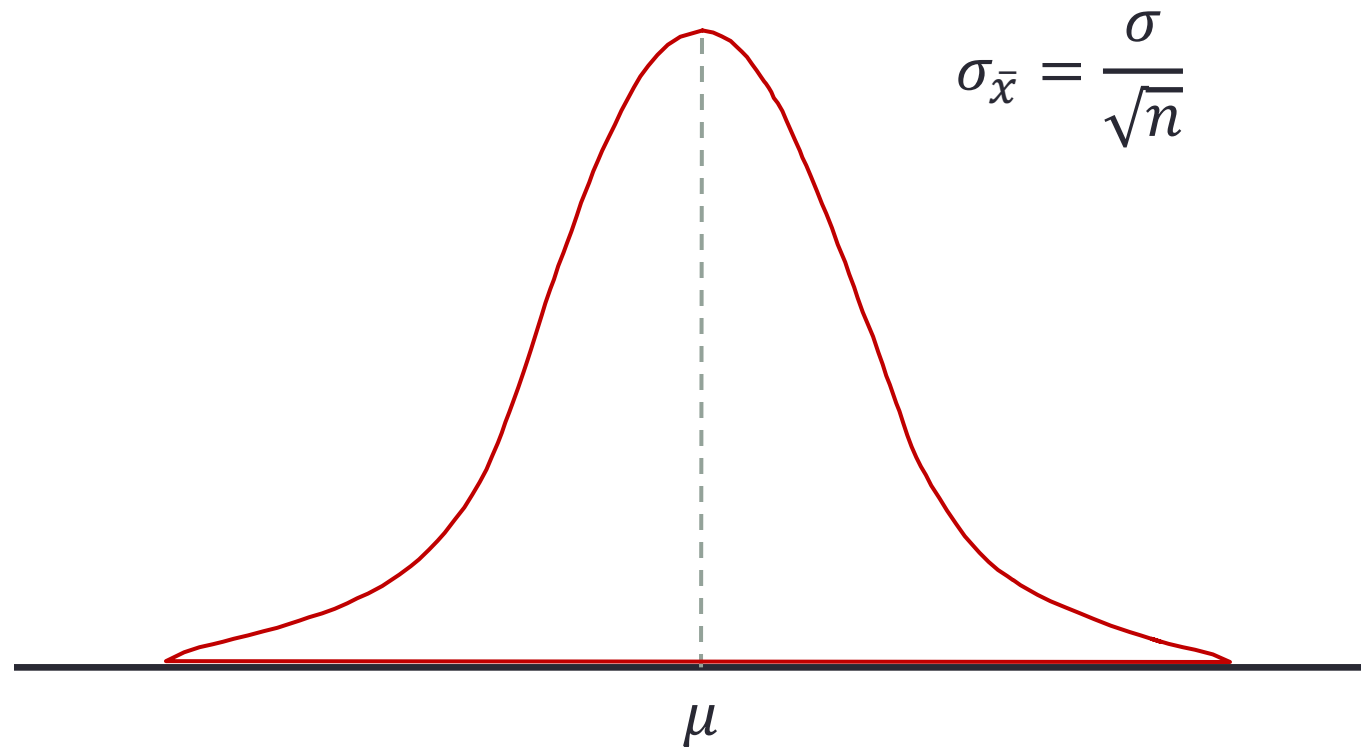
- The purpose of an interval estimate is to provide information about how close the point estimate is to the value of the parameter.

# Sampling Distribution of $\bar{x}$

- The sampling distribution of  $\bar{x}$  plays a key role in computing the margin of error for this interval estimate.
- The **sampling distribution of  $\bar{x}$**  is approximately the **Normal distribution** whenever  $n \geq 50$ .

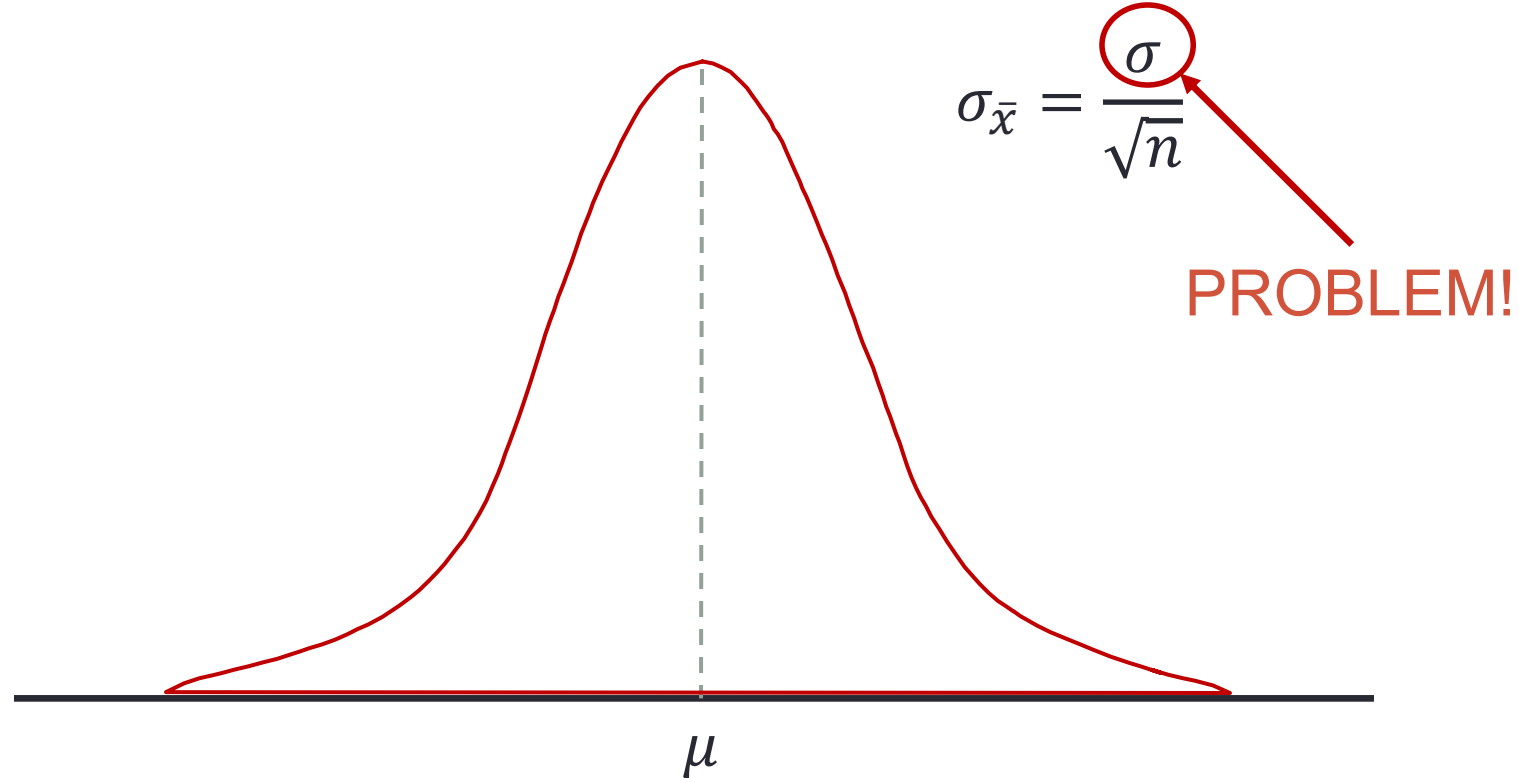
# Sampling Distribution of $\bar{x}$

- The **sampling distribution of  $\bar{x}$**  is approximately the **Normal distribution** whenever  $n \geq 50$ .



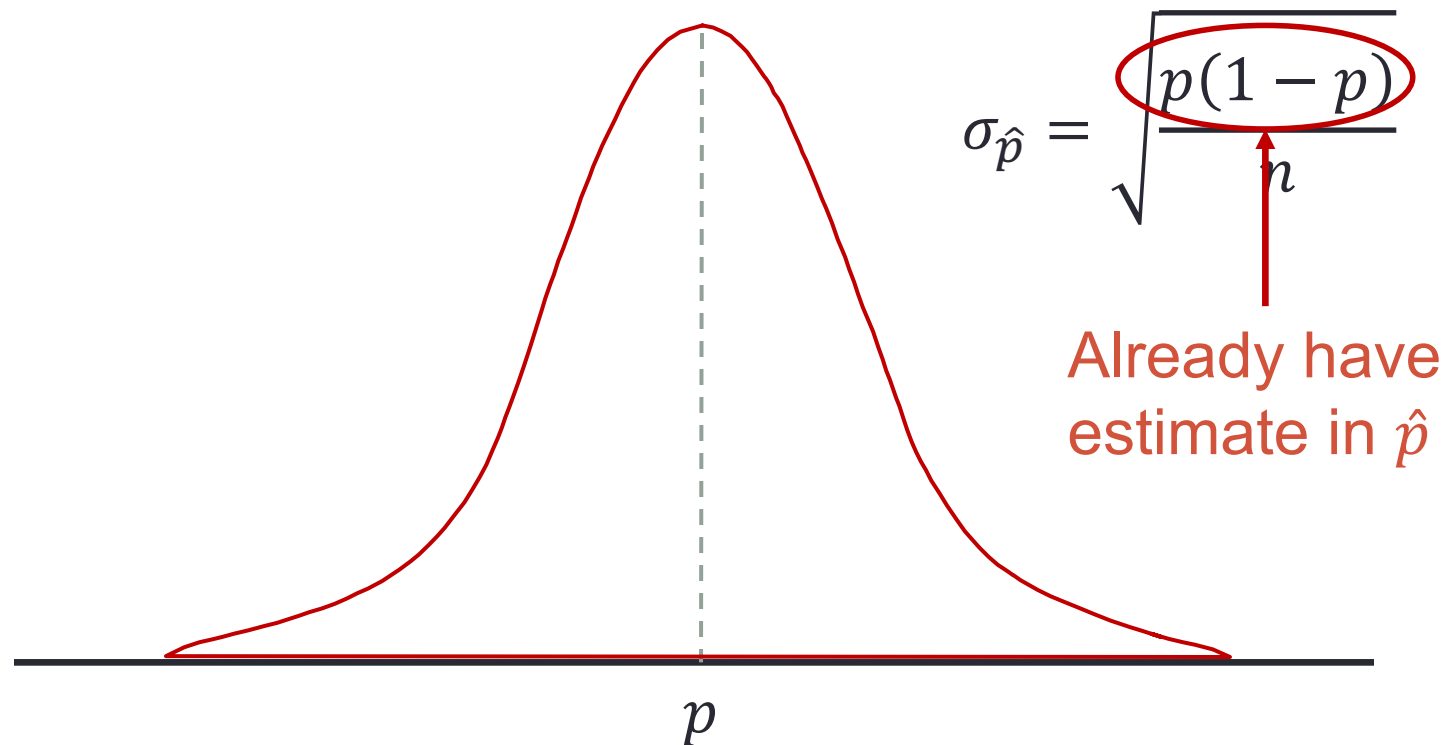
# Sampling Distribution of $\bar{x}$

- The **sampling distribution of  $\bar{x}$**  is approximately the **Normal distribution** whenever  $n \geq 50$ .



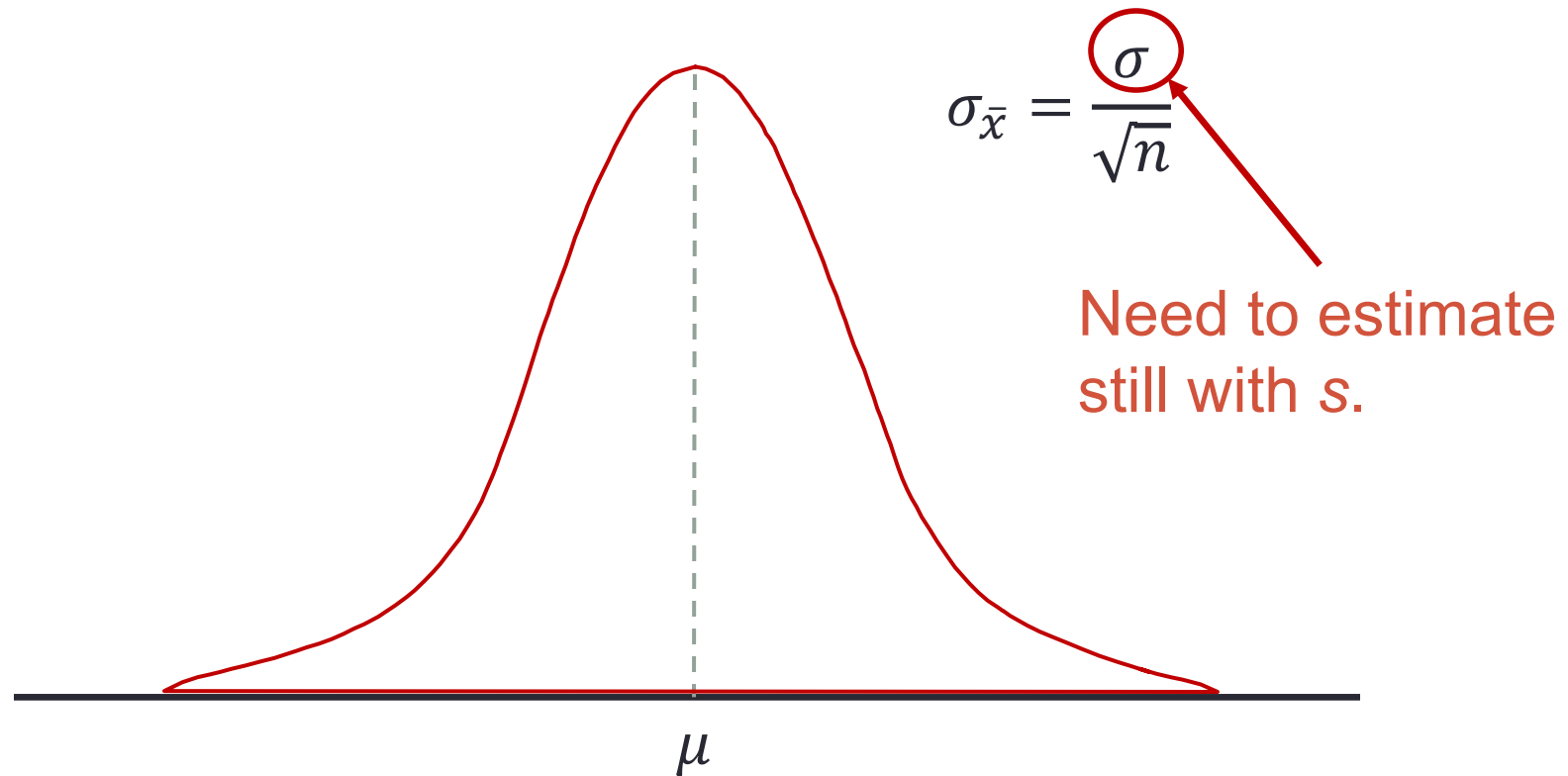
# Sampling Distribution of $\hat{p}$

- The **sampling distribution of  $\hat{p}$**  is approximately the **Normal distribution** whenever  $np \geq 5$  and  $n(1 - p) \geq 5$ .



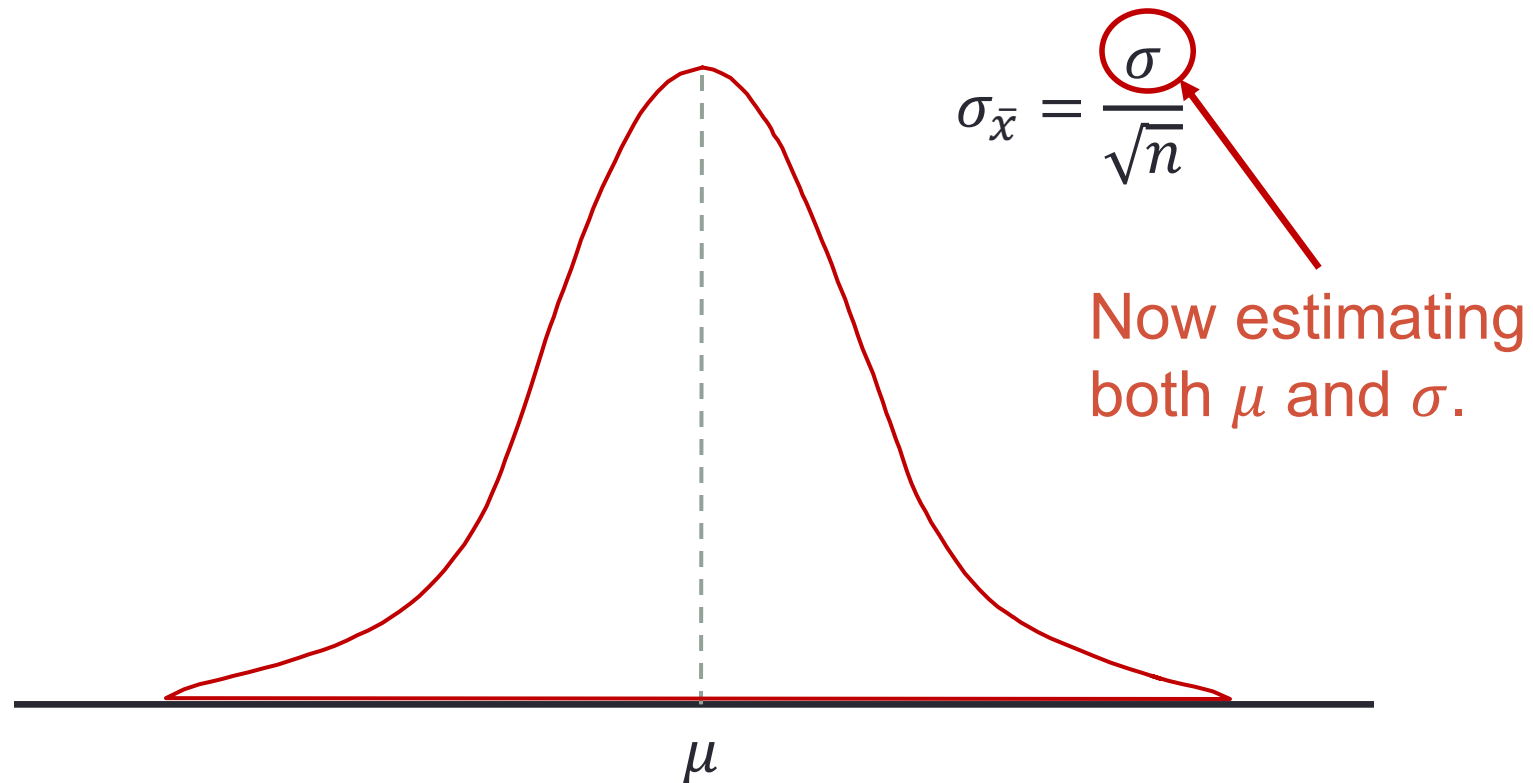
# Sampling Distribution of $\bar{x}$

- The **sampling distribution of  $\bar{x}$**  is approximately the **Normal distribution** whenever  $n \geq 50$ .



# Sampling Distribution of $\bar{x}$

- The **sampling distribution of  $\bar{x}$**  is approximately the **Normal distribution** whenever  $n \geq 50$ .





# Unknown $\sigma$

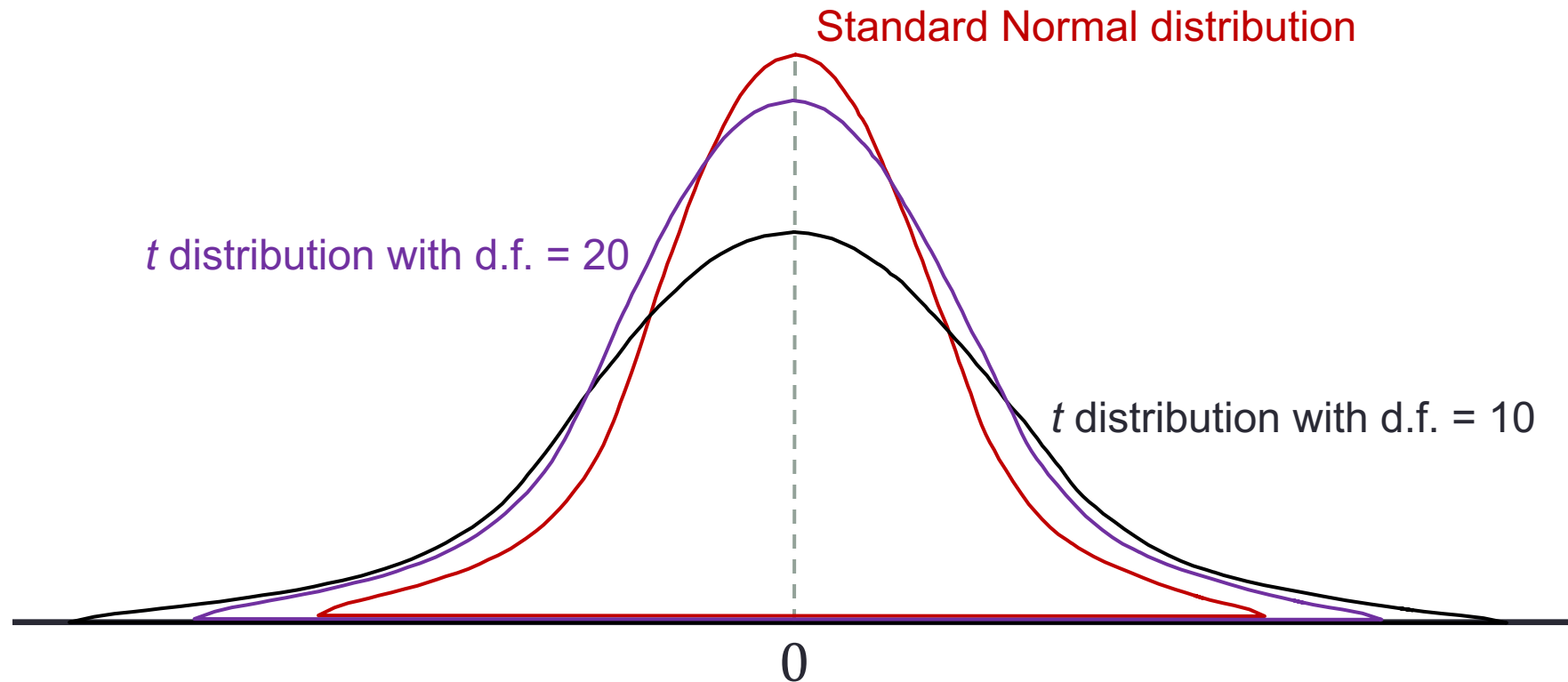
- Since we do not know the population standard deviation and need to estimate it with the sample standard deviation, we have added extra error into our calculations.
- Estimating two statistics has more error than just estimating one.
- Normal distribution is no longer a good approximation for the sampling distribution of  $\bar{x}$  because it doesn't account for this extra error.
- Need to use another distribution.

# Student $t$ - Distribution

- The  **$t$  distribution** is a family of similar probability distributions.
- The  $t$  distribution is symmetric, but has thicker tails than the Normal distribution.
- The  $t$  distribution has degrees of freedom:  $d.f. = n - 1$
- Degrees of freedom are the number of independent pieces of information that go into the computation of  $s$ .
- More degrees of freedom leads to less dispersion in the distribution.
- For larger samples, the  $t$  distribution is **approximately** the standard Normal distribution.

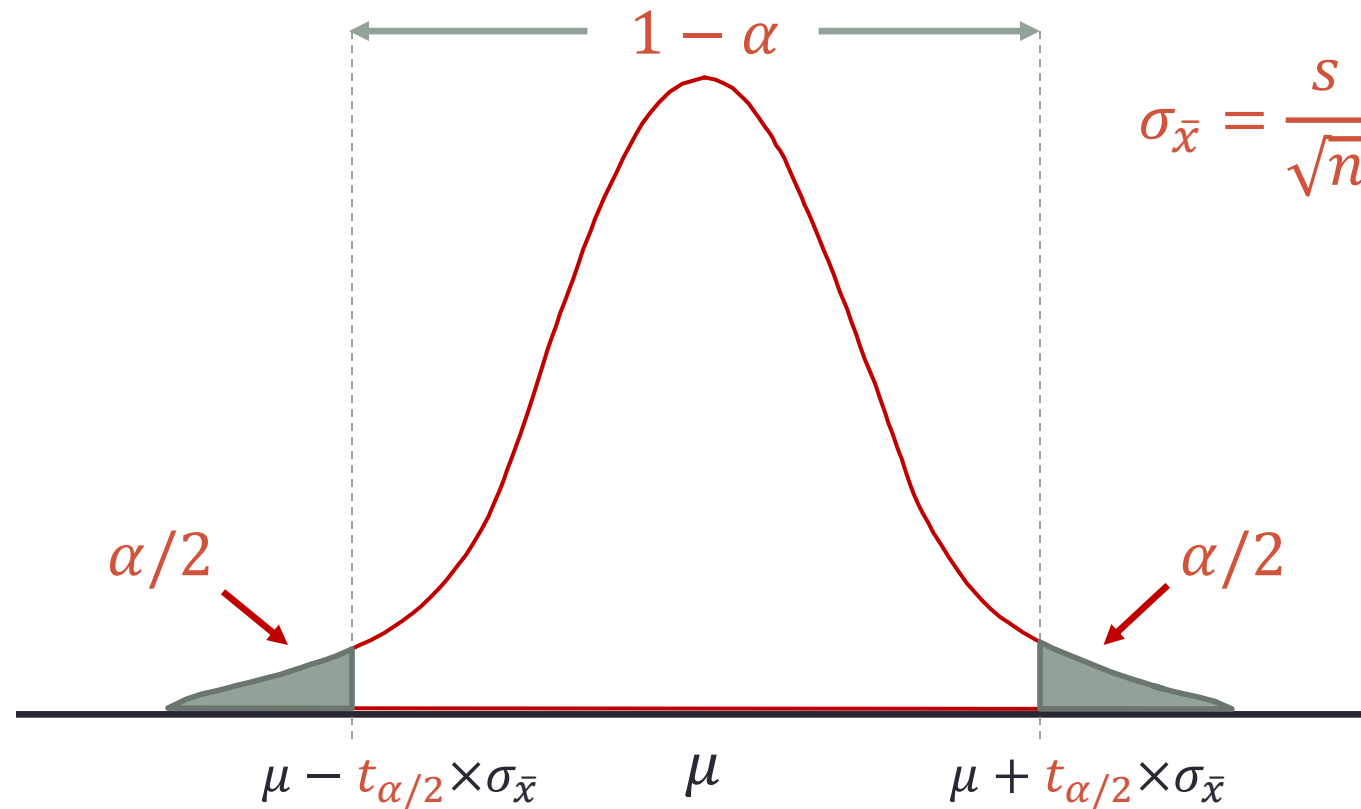
# Standard Normal vs. $t$ Distribution

- For larger samples, the  $t$  distribution is **approximately** the standard Normal distribution.



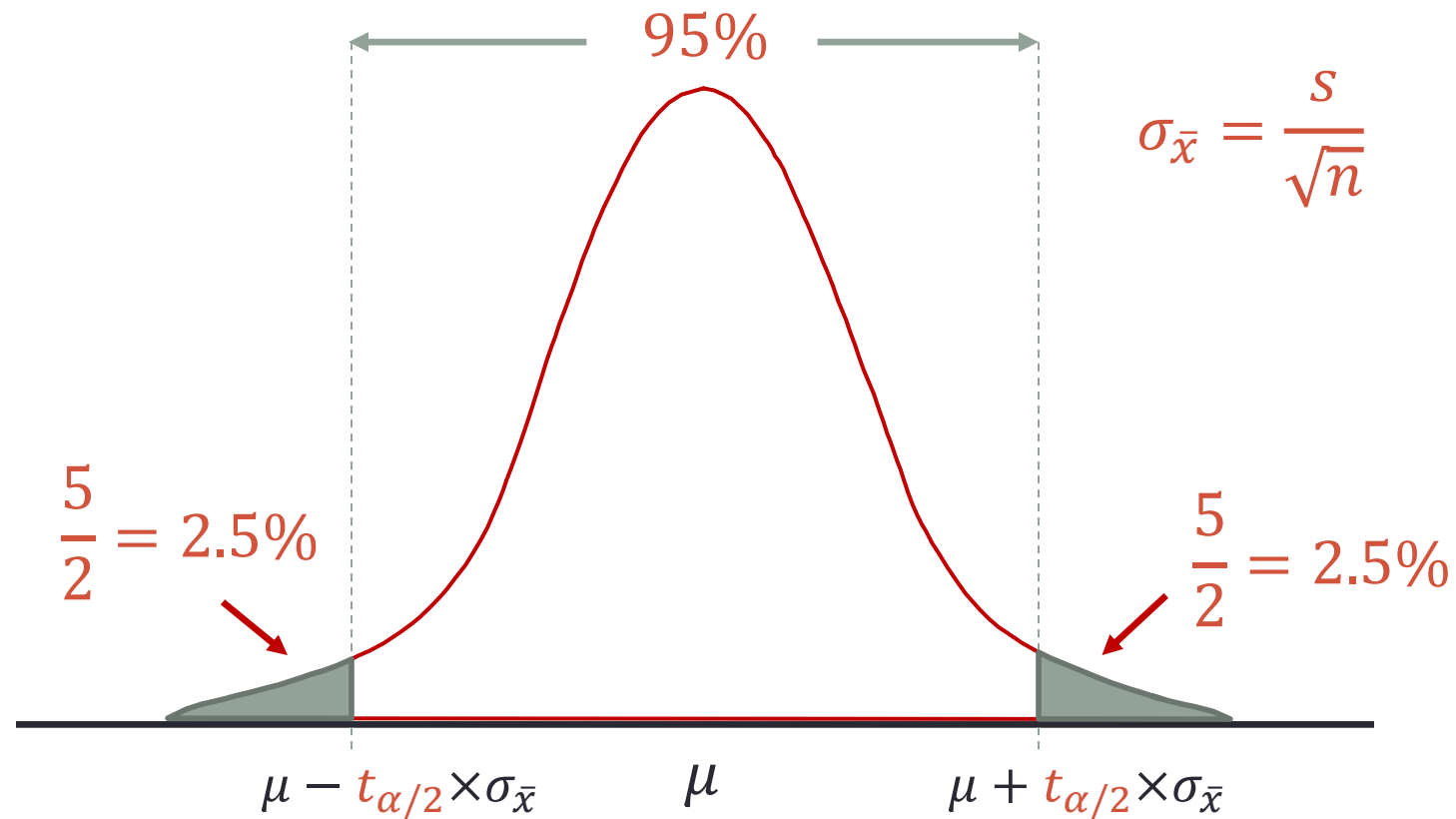
# Sampling Distribution of $\bar{x}$

- Need to use the  $t$  distribution instead for the confidence intervals of  $\bar{x}$ .



# Sampling Distribution of $\bar{x}$

- Need to use the  $t$  distribution instead for the confidence intervals of  $\bar{x}$ .



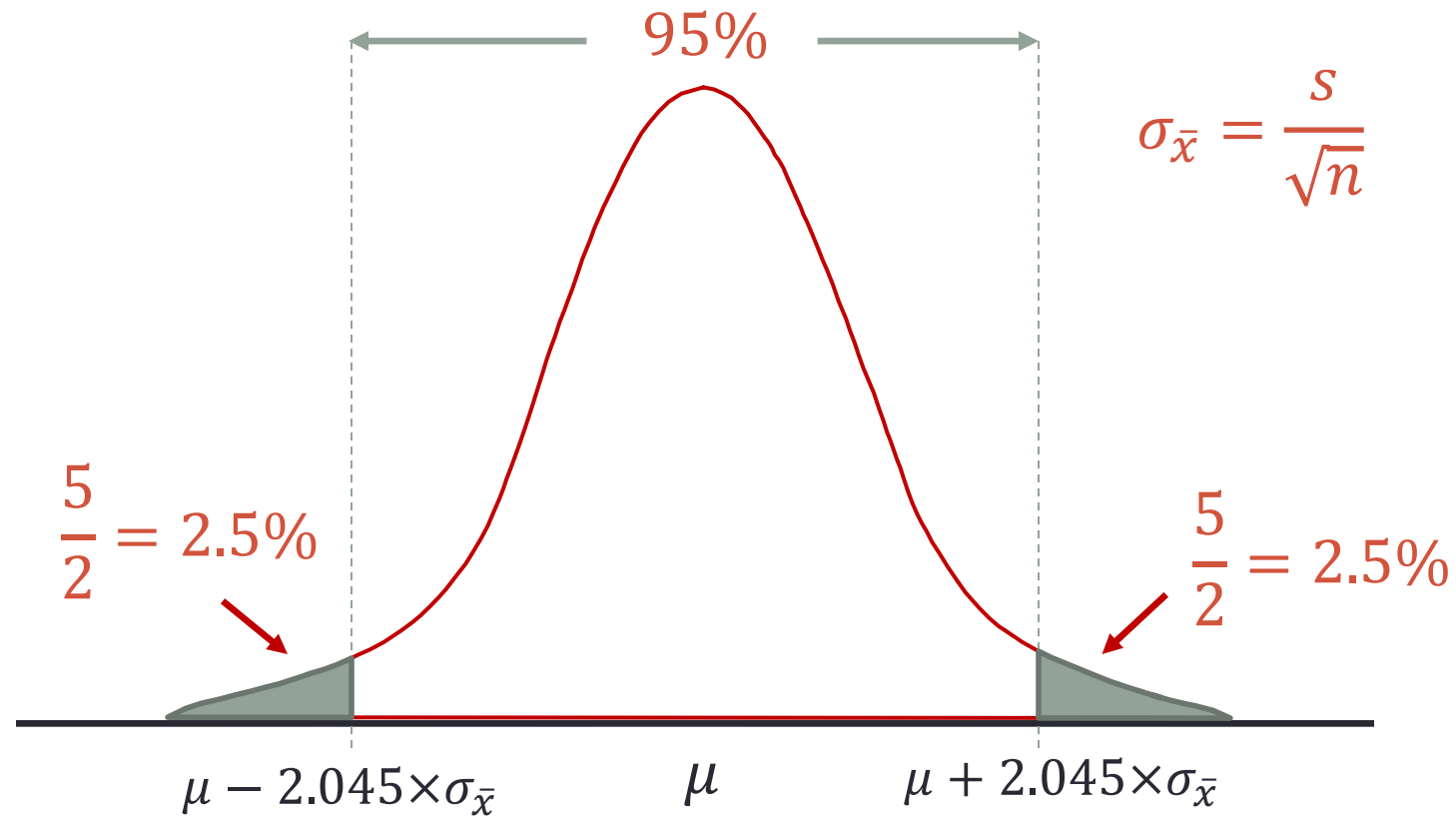
What is  $t_{\alpha/2}$  for  $n = 30$ ?

# How to Calculate $t_{\alpha/2}$

[illegible]

# Sampling Distribution of $\bar{x}$

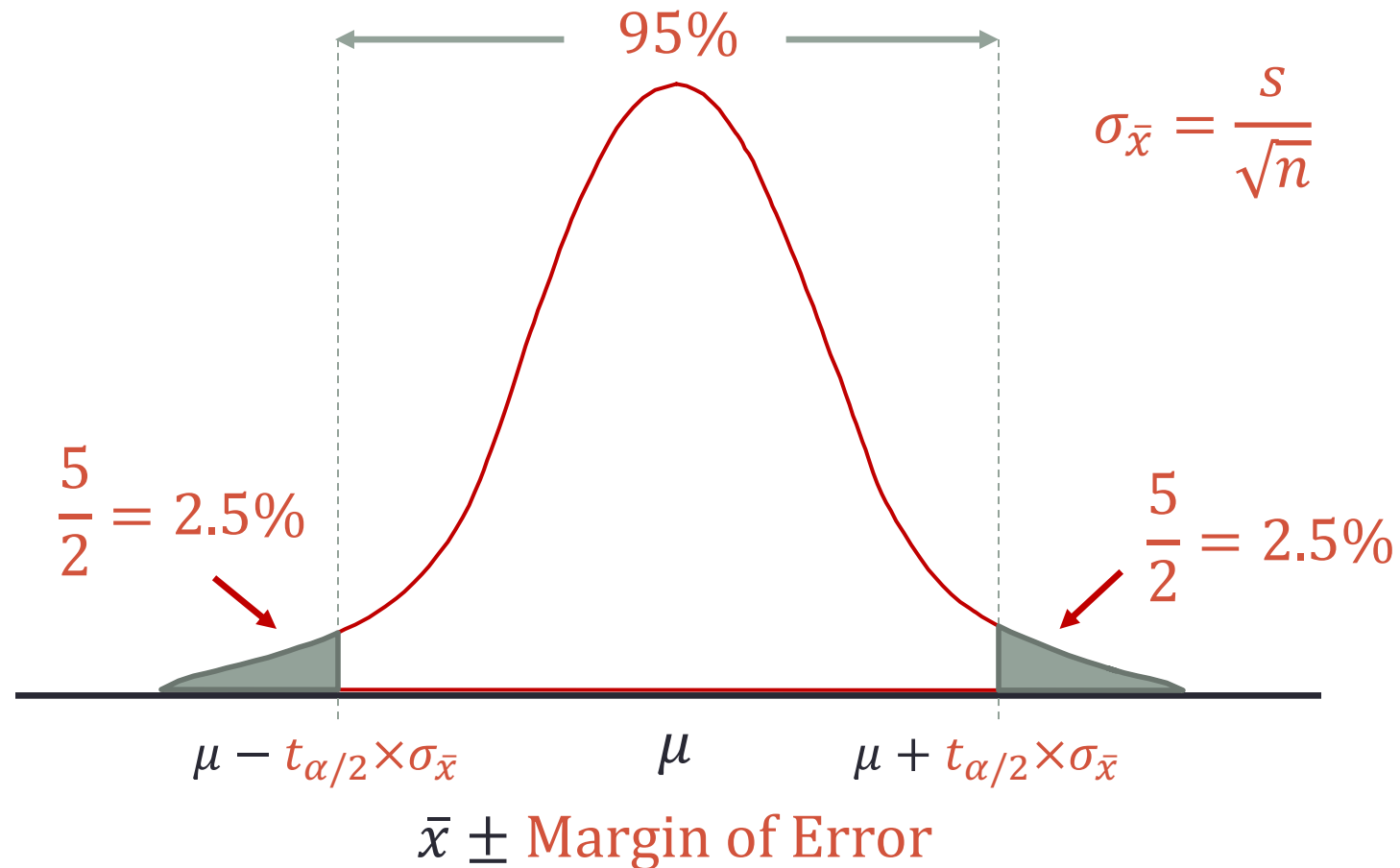
- Need to use the  $t$  distribution instead for the confidence intervals of  $\bar{x}$ .



What is  $t_{\alpha/2}$  for  $n = 30$ ?

# Sampling Distribution of $\bar{x}$

- Need to use the  $t$  distribution instead for the confidence intervals of  $\bar{x}$ .





# Confidence Interval for $\bar{x}$

- The **confidence interval for  $\bar{x}$**  with a **confidence coefficient** of  $1 - \alpha$  (error of  $\alpha$ ) is the following:

$$\bar{x} \pm t_{\alpha/2} \times \frac{s}{\sqrt{n}}$$

# Additional Assumptions

- For large samples ( $n \geq 50$ ), you can calculate the confidence interval for the mean **from any population**.
- For small samples ( $n < 50$ ), you need to assume that the **population follows a Normal distribution**.

# Change of Things

- What happens to confidence intervals with a change of sample size?
- What happens with a change of confidence level?

# Change of Things

- What happens to confidence intervals with a change of sample size?

$$n \uparrow \Rightarrow \frac{s}{\sqrt{n}} \downarrow \text{ \& } t^* \downarrow \Rightarrow t^* \left( \frac{s}{\sqrt{n}} \right) \downarrow \Rightarrow \text{width} \downarrow$$

- What happens with a change of confidence level?

# Change of Things

- What happens to confidence intervals with a change of sample size?

$$n \uparrow \Rightarrow \frac{s}{\sqrt{n}} \downarrow \ \& \ t^* \downarrow \Rightarrow t^* \left( \frac{s}{\sqrt{n}} \right) \downarrow \Rightarrow \text{width} \downarrow$$

- What happens with a change of confidence level?

$$C \uparrow \Rightarrow t^* \uparrow \Rightarrow t^* \left( \frac{s}{\sqrt{n}} \right) \uparrow \Rightarrow \text{width} \uparrow$$

# Example Confidence Interval for $\bar{x}$

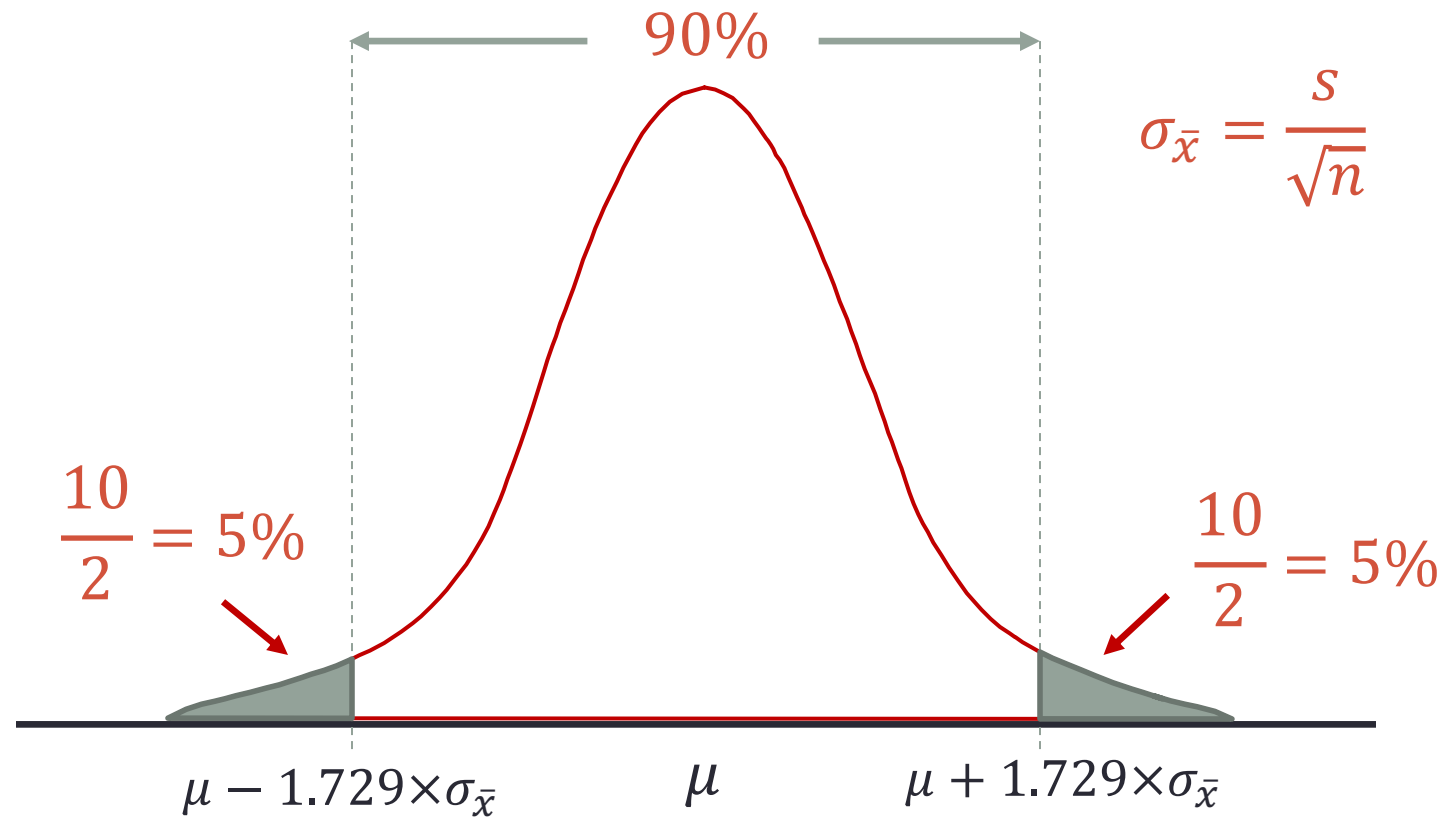
- Build a 90% confidence interval for the true average daily sales in thousands of dollars for stores.
- A sample of 20 stores was taken with a sample mean of \$17.06K/day with a sample standard deviation of \$5.12K.
- What additional assumption do we need?

# Example Confidence Interval for $\bar{x}$

- Build a 90% confidence interval for the true average daily sales in thousands of dollars for stores.
- A sample of 20 stores was taken with a sample mean of \$17.06K/day with a sample standard deviation of \$5.12K.
- Assume the population follows a Normal distribution.

# Sampling Distribution of $\bar{x}$

- Need to use the  $t$  distribution instead for the confidence intervals of  $\bar{x}$ .



What is  $t_{\alpha/2}$  for  $n = 20$ ?



# Example Confidence Interval for $\bar{x}$

- Build a 90% confidence interval for the true average daily sales in thousands of dollars for stores.
- A sample of 20 stores was taken with a sample mean of \$17.06K/day with a sample standard deviation of \$5.12K.
- Assume the population follows a Normal distribution.

$$\bar{x} \pm t_{\alpha/2} \times \frac{s}{\sqrt{n}}$$

$$17.06 \pm 1.729 \times \frac{5.12}{\sqrt{20}}$$

$$17.06 \pm 1.98 \quad \text{OR} \quad (15.08, 19.04)$$

# Example Confidence Interval for $\bar{x}$

- Build a 90% confidence interval for the true average daily sales in thousands of dollars for stores.
- A sample of 100 stores was taken with a sample mean of \$17.06K/day with a sample standard deviation of \$5.12K.
- What additional assumption do we need?

# Example Confidence Interval for $\bar{x}$

- Build a 90% confidence interval for the true average daily sales in thousands of dollars for stores.
- A sample of 100 stores was taken with a sample mean of \$17.06K/day with a sample standard deviation of \$5.12K.
- **NONE!**

# Example Confidence Interval for $\bar{x}$

- Build a 90% confidence interval for the true average daily sales in thousands of dollars for stores.
- A sample of 100 stores was taken with a sample mean of \$17.06K/day with a sample standard deviation of \$5.12K.
- Assume the population follows a Normal distribution.

$$\bar{x} \pm t_{\alpha/2} \times \frac{s}{\sqrt{n}}$$

$$17.06 \pm 1.662 \times \frac{5.12}{\sqrt{100}}$$

$$17.06 \pm 0.85 \quad \text{OR} \quad (16.21, 17.91)$$

# SAMPLE SIZE CALCULATION

---

# Reversing the Problem

- What if we wanted to know what sample size  $n$  would need to collect to get a desired margin of error?
- Instead of calculating a confidence interval (or margin of error) after a sample is taken, we can look at the problem in reverse.
- For example, your boss allows a margin of error of  $E$ , but wants you to take as small of a sample as needed to have at least that margin of error.

# Sample Size Needed for $\hat{p}$

- Take the margin of error:

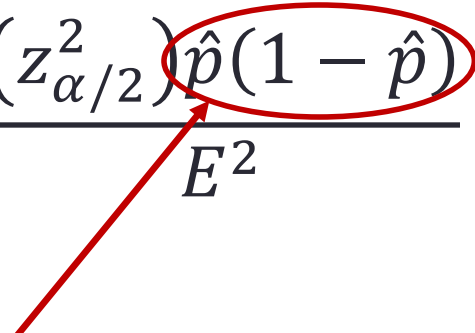
$$E = z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

- Solve for sample size:

$$n = \frac{(z_{\alpha/2}^2) \hat{p}(1 - \hat{p})}{E^2}$$

# Sample Size Needed for $\hat{p}$

- Solve for sample size:

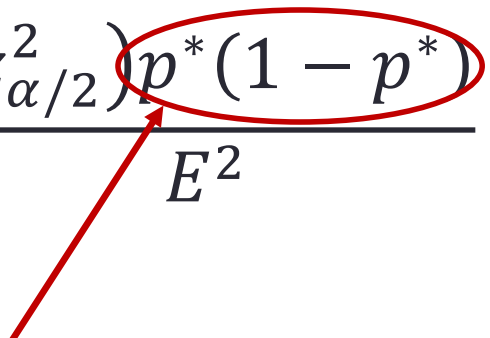
$$n = \frac{(z_{\alpha/2}^2) \hat{p}(1 - \hat{p})}{E^2}$$


Don't know  $\hat{p}$  ahead of sampling!



# Sample Size Needed for $\hat{p}$

- Solve for sample size:

$$n = \frac{(z_{\alpha/2}^2) p^* (1 - p^*)}{E^2}$$


$p^*$  is best guess ahead of sampling

# Sample Size Needed for $\hat{p}$

- Sample size calculation:

$$n = \frac{(z_{\alpha/2}^2)p^*(1 - p^*)}{E^2}$$

- In practice, typically we use  $p^* = 0.5$  as that will provide us the largest sample size for any true value of  $\hat{p}$ .
- You can put any value of  $p^*$  into  $p^*(1 - p^*)$  and no value will be larger than if  $p^* = 0.5$ .

## Example Sample Size Needed for $\hat{p}$

- You are interested in hair color and eye color across 2 different regions of the country. You want to estimate the proportion of people with blue eyes with a 90% CI. You are willing to have a margin of error of 3%.

## Example Sample Size Needed for $\hat{p}$

- You are interested in hair color and eye color across 2 different regions of the country. You want to estimate the proportion of people with blue eyes with a 90% CI. You are willing to have a margin of error of 3%.

$$n = \frac{(z_{\alpha/2}^2)p^*(1 - p^*)}{E^2}$$

$$n = \frac{(1.645^2)0.5(1 - 0.5)}{0.03^2}$$

$$n = 751.67$$

$$n \approx 752$$

# Sample Size Needed for $\bar{x}$

- Take the margin of error:

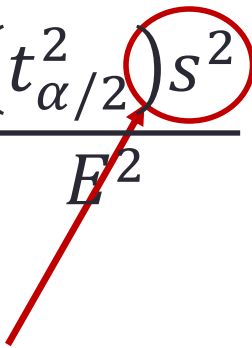
$$E = t_{\alpha/2} \times \frac{s}{\sqrt{n}}$$

- Solve for sample size:

$$n = \frac{(t_{\alpha/2}^2) s^2}{E^2}$$

# Sample Size Needed for $\bar{x}$

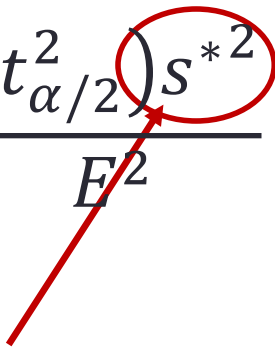
- Solve for sample size:

$$n = \frac{(t_{\alpha/2}^2) s^2}{E^2}$$


Don't know  $s^2$  ahead of sampling!

# Sample Size Needed for $\bar{x}$


- Solve for sample size:

$$n = \frac{(t_{\alpha/2}^2) s^{*2}}{E^2}$$


$s^*$  is best guess ahead of sampling

# Sample Size Needed for $\bar{x}$

- Solve for sample size:

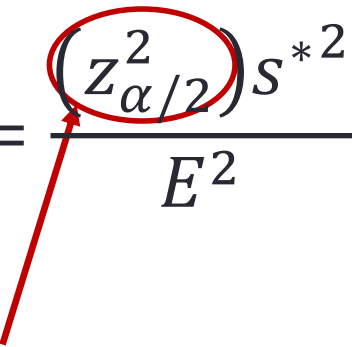
$$n = \frac{(t_{\alpha/2}^2) s^{*2}}{E^2}$$


Don't know ahead of sampling because  
it depends on sample size!



# Sample Size Needed for $\bar{x}$

- Solve for sample size:

$$n = \frac{(z_{\alpha/2}^2) s^{*2}}{E^2}$$


Typically use Normal distribution approximation

# Sample Size Needed for $\bar{x}$

- Sample size calculation:

$$n = \frac{(z_{\alpha/2}^2) s^{*2}}{E^2}$$

- In practice, typically we take a pilot sample or use previous information to get  $s^{*2}$ .