INTERVAL ESTIMATION WITH DATA

ST101 – DR. ARIC LABARR

MARGIN OF ERROR

- A point estimator cannot be expected to provide the exact value of the population parameter.
- An **interval estimate** can be computed by adding and subtracting a **margin or error** to the point estimate:

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

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

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

- **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 average daily number of total users of the bike rental company is between 4,000 and 5,000.

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 average daily number of total users of the bike rental company is between 4,000 and 5,000.

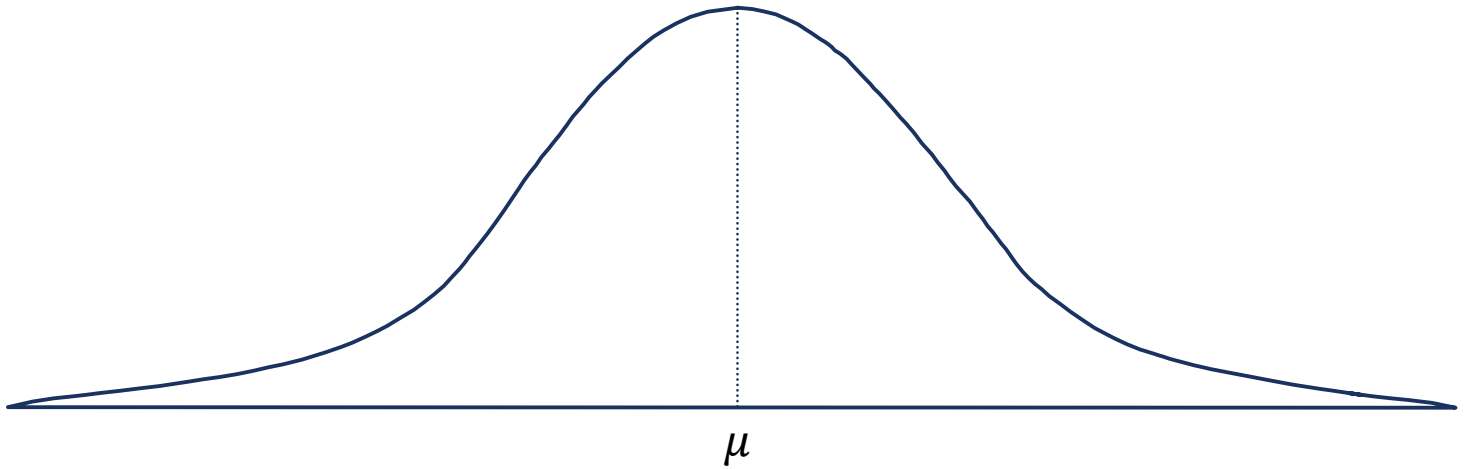
If we were to take many samples (same size) that each produced different confidence intervals, then 95% of them would contain the true parameter.

CONFIDENCE INTERVALS

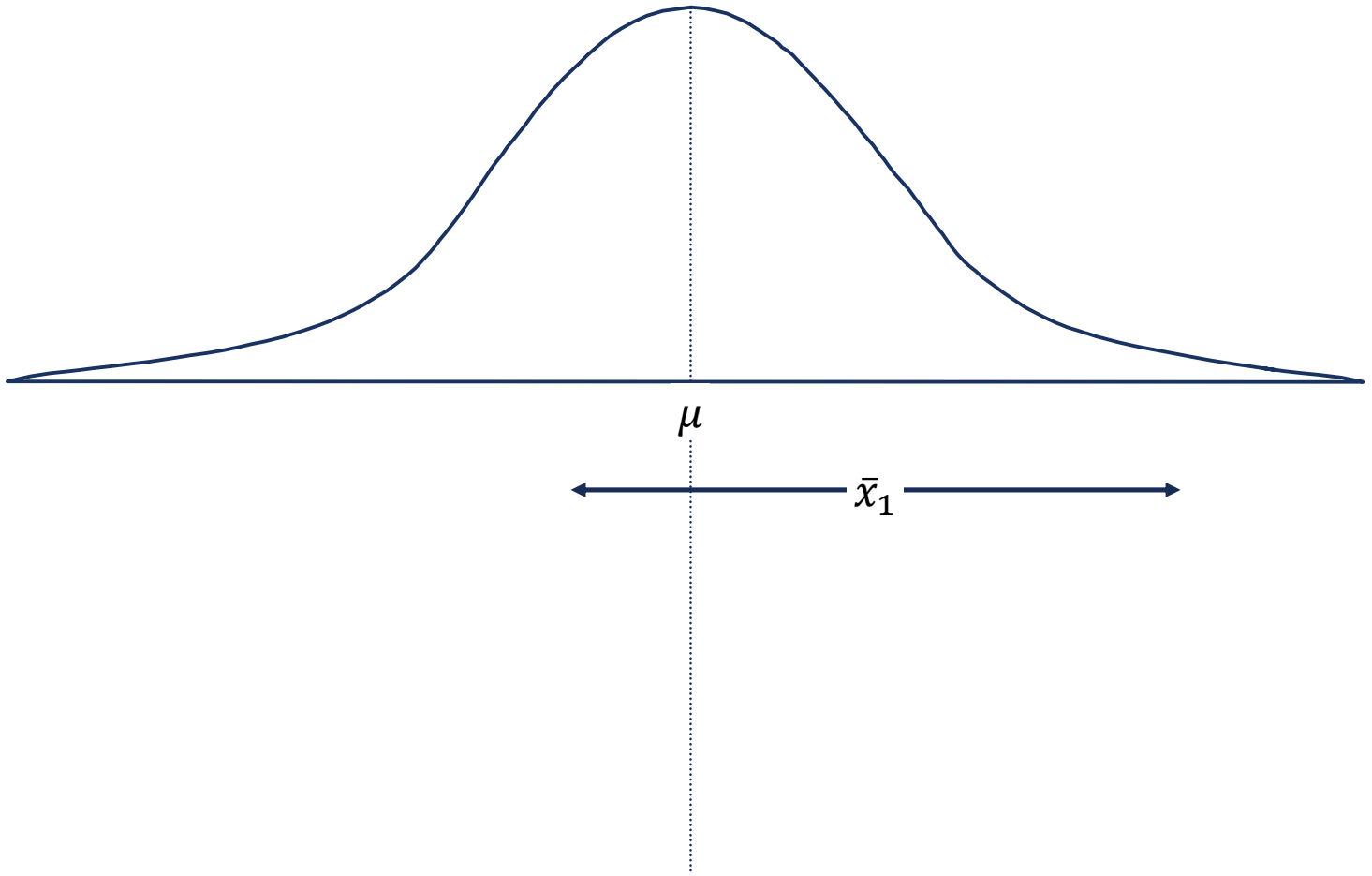
- **Confidence Intervals** are interval estimates where we say we have a certain level of **confidence** in the interval.
- For example, we are **95% confident** that the population average daily number of total users of the bike rental company is between 4,000 and 5,000.

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

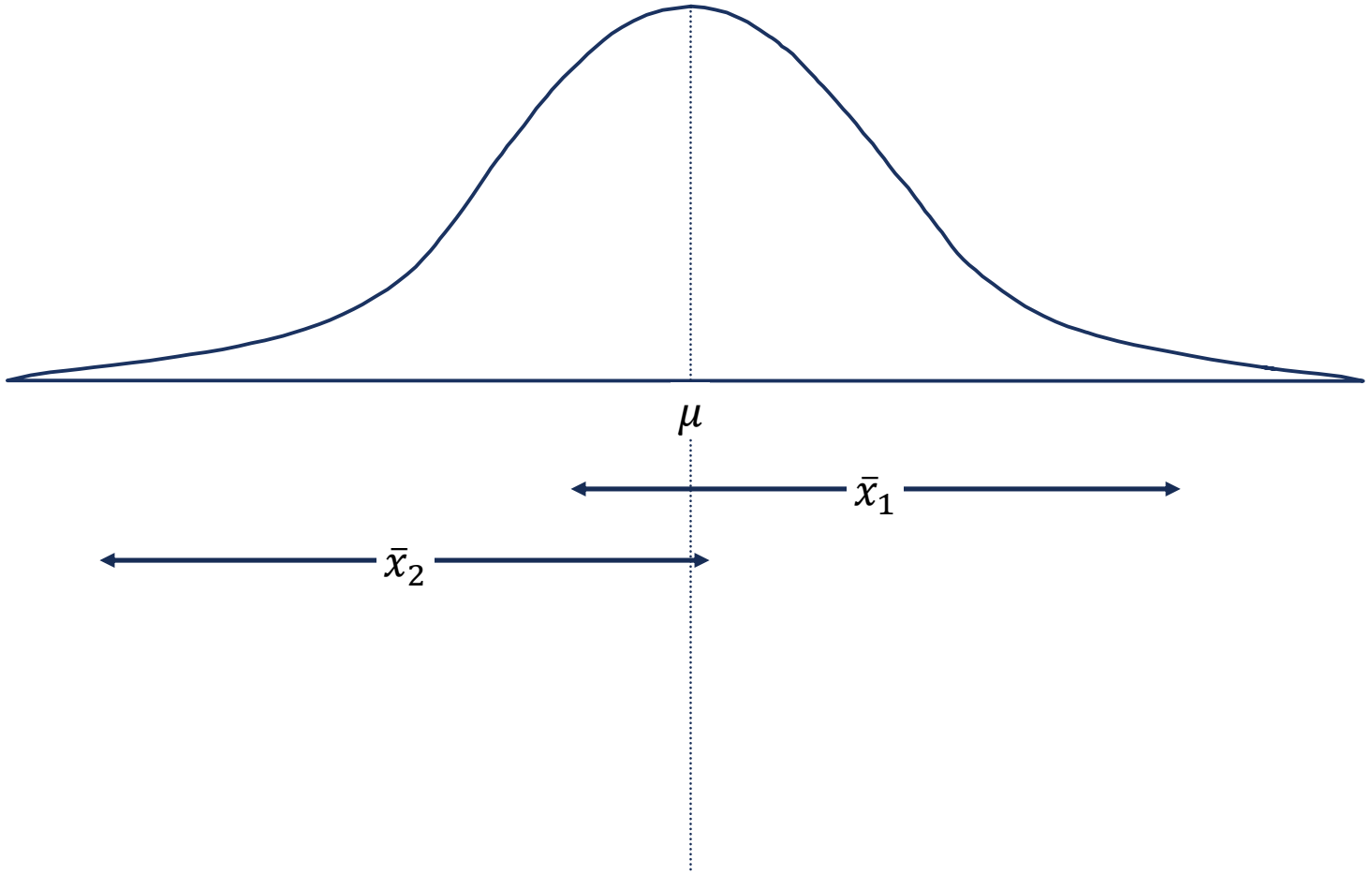
CONFIDENCE INTERVALS EXAMPLE



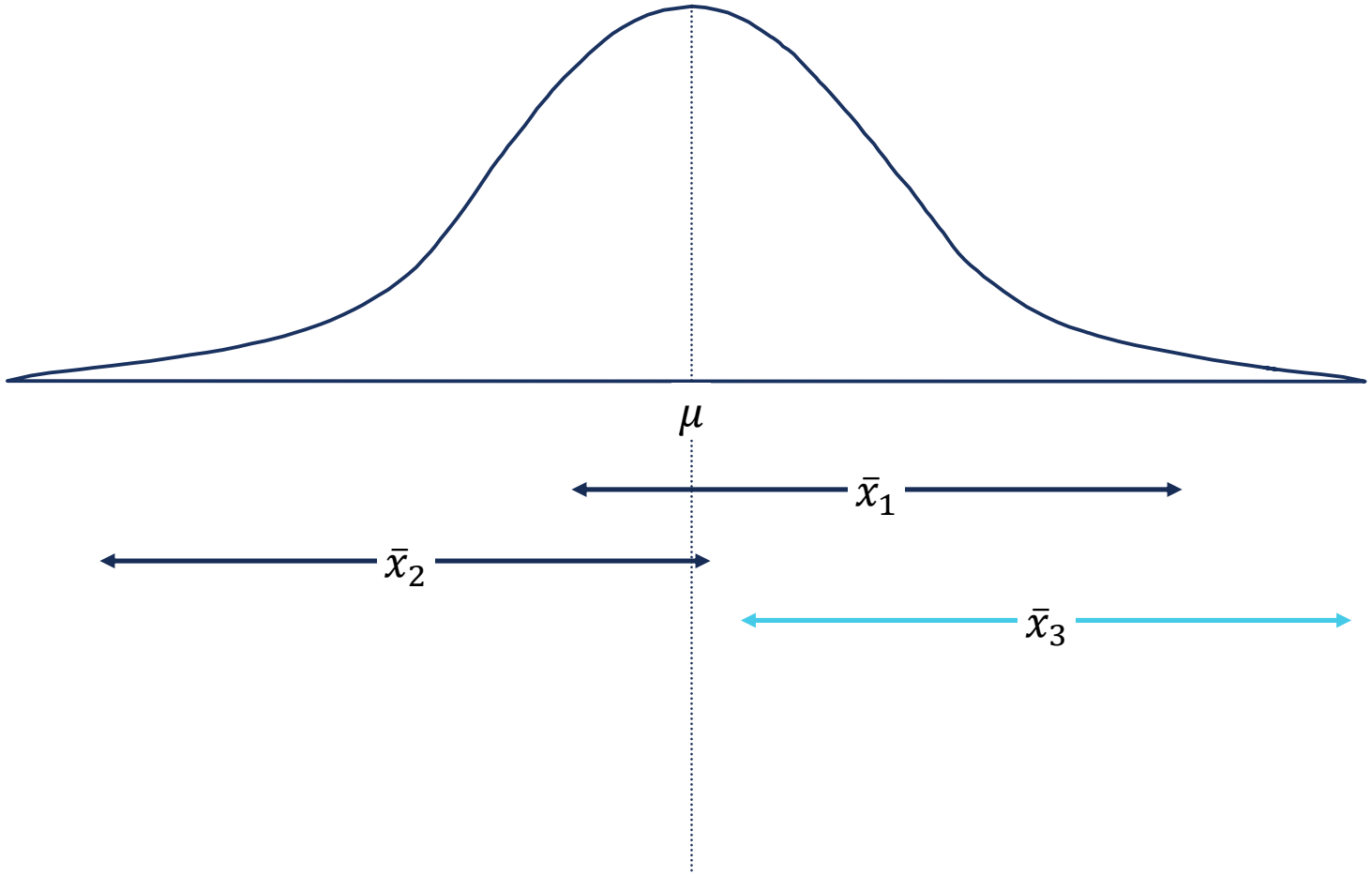
CONFIDENCE INTERVALS EXAMPLE



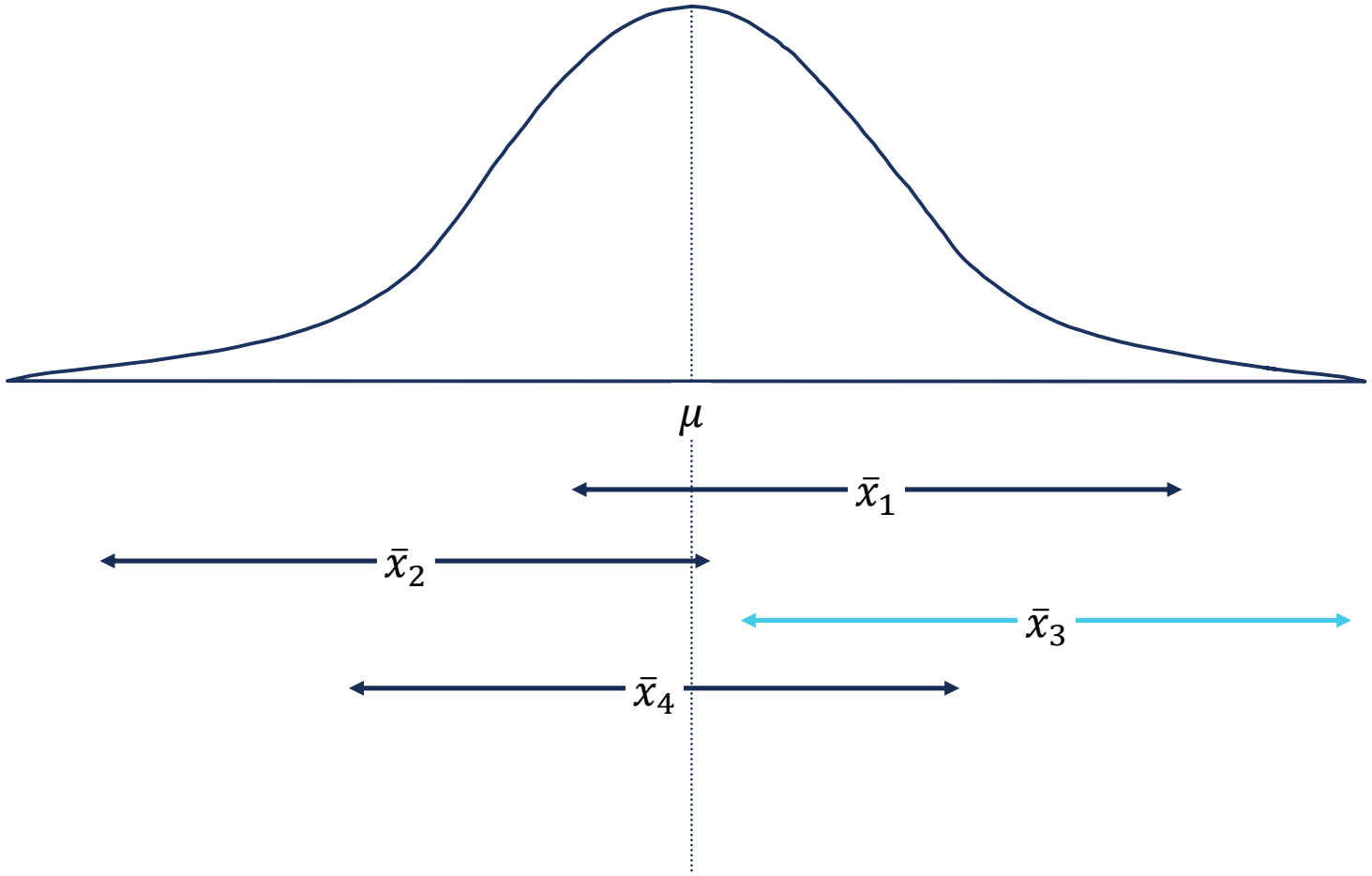
CONFIDENCE INTERVALS EXAMPLE



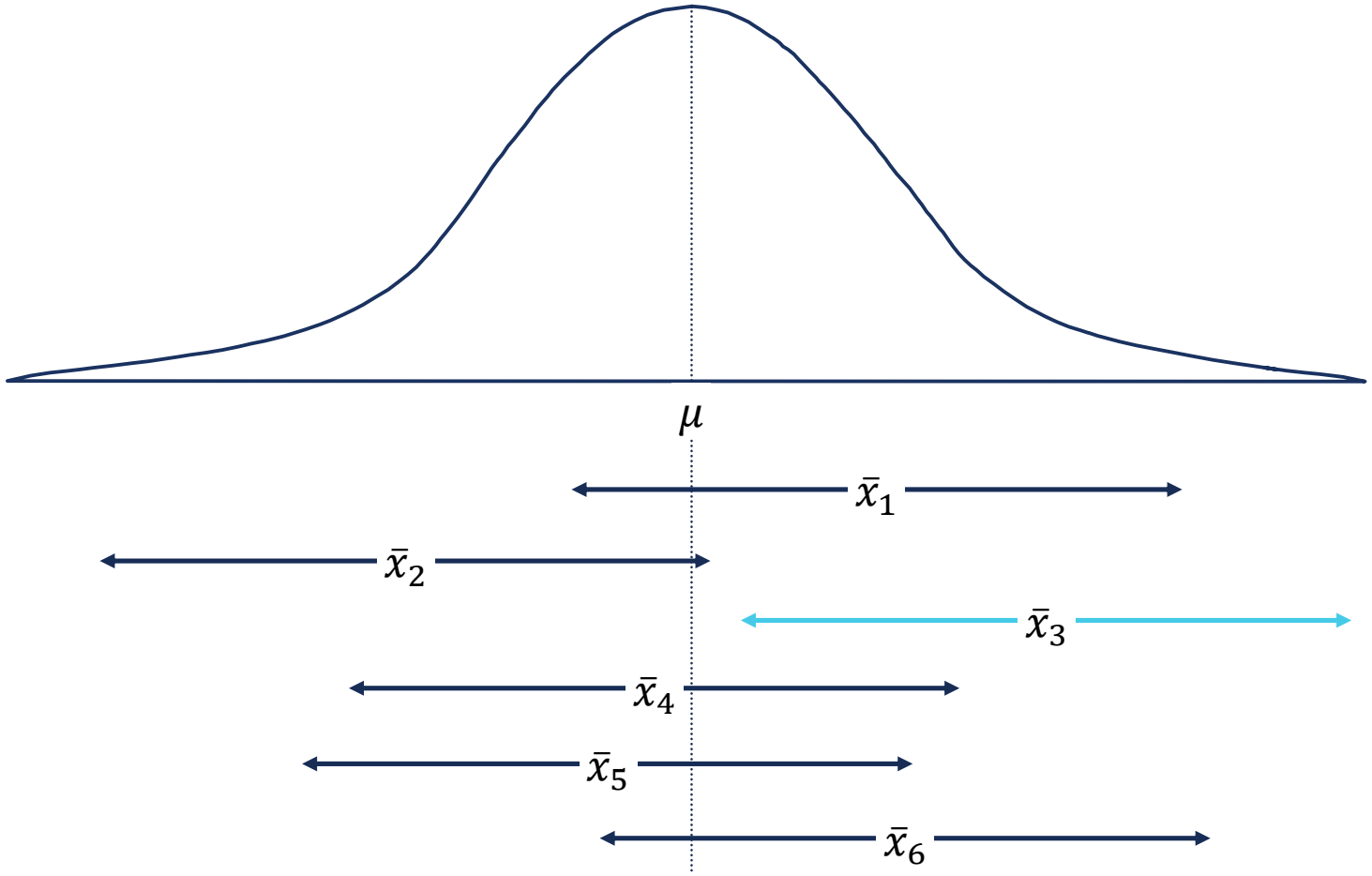
CONFIDENCE INTERVALS EXAMPLE



CONFIDENCE INTERVALS EXAMPLE



CONFIDENCE INTERVALS EXAMPLE



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 average daily number of total users of the bike rental company is between 4,000 and 5,000.

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

SUMMARY

- Confidence Intervals are interval estimates where we say we have a certain level of confidence in the interval.
- Confidence implies 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 is **NOT** the chance the population parameter falls inside our one confidence interval.

INTERVAL ESTIMATION OF $\hat{p}$

INTERVAL ESTIMATION WITH DATA

MARGIN OF ERROR

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

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

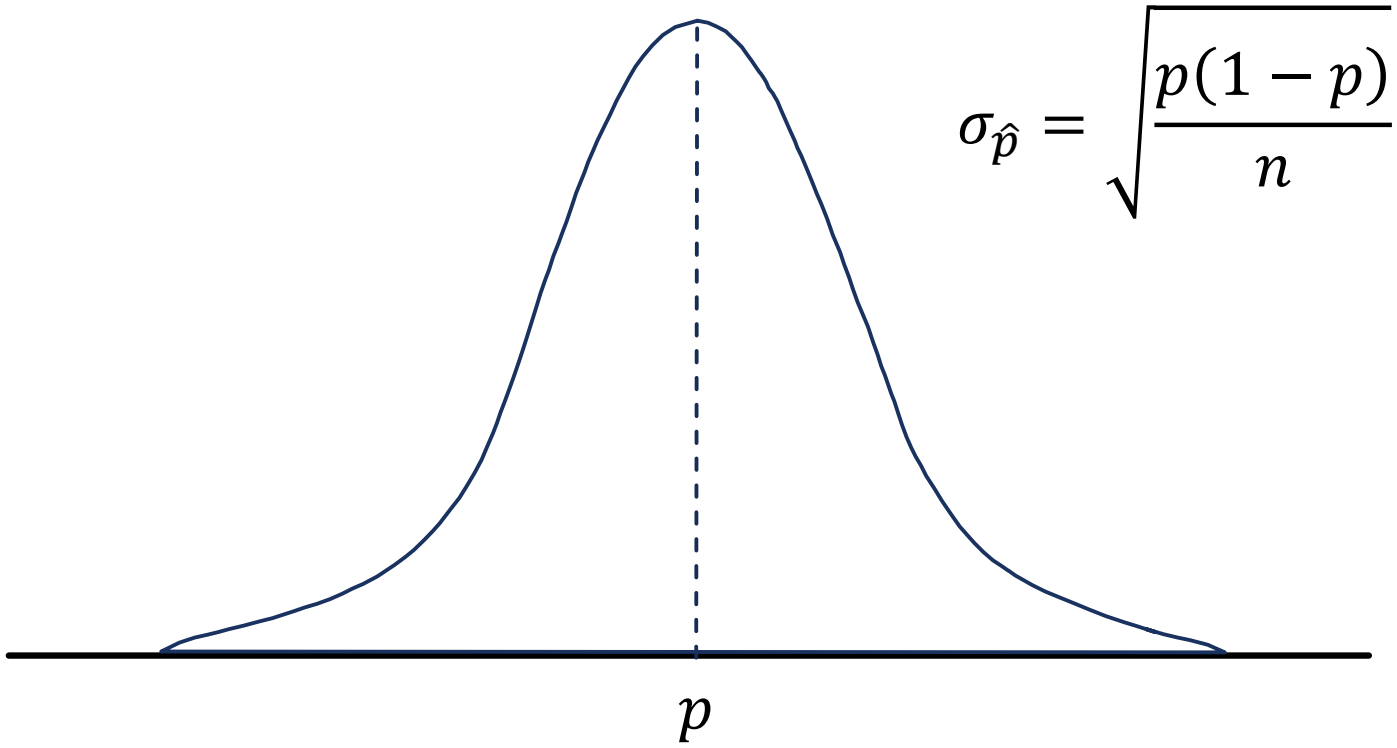
- The purpose of an interval estimate is to provide information about how close the point estimate is to the value of the parameter.

SAMPLING DISTRIBUTION OF $\hat{p}$

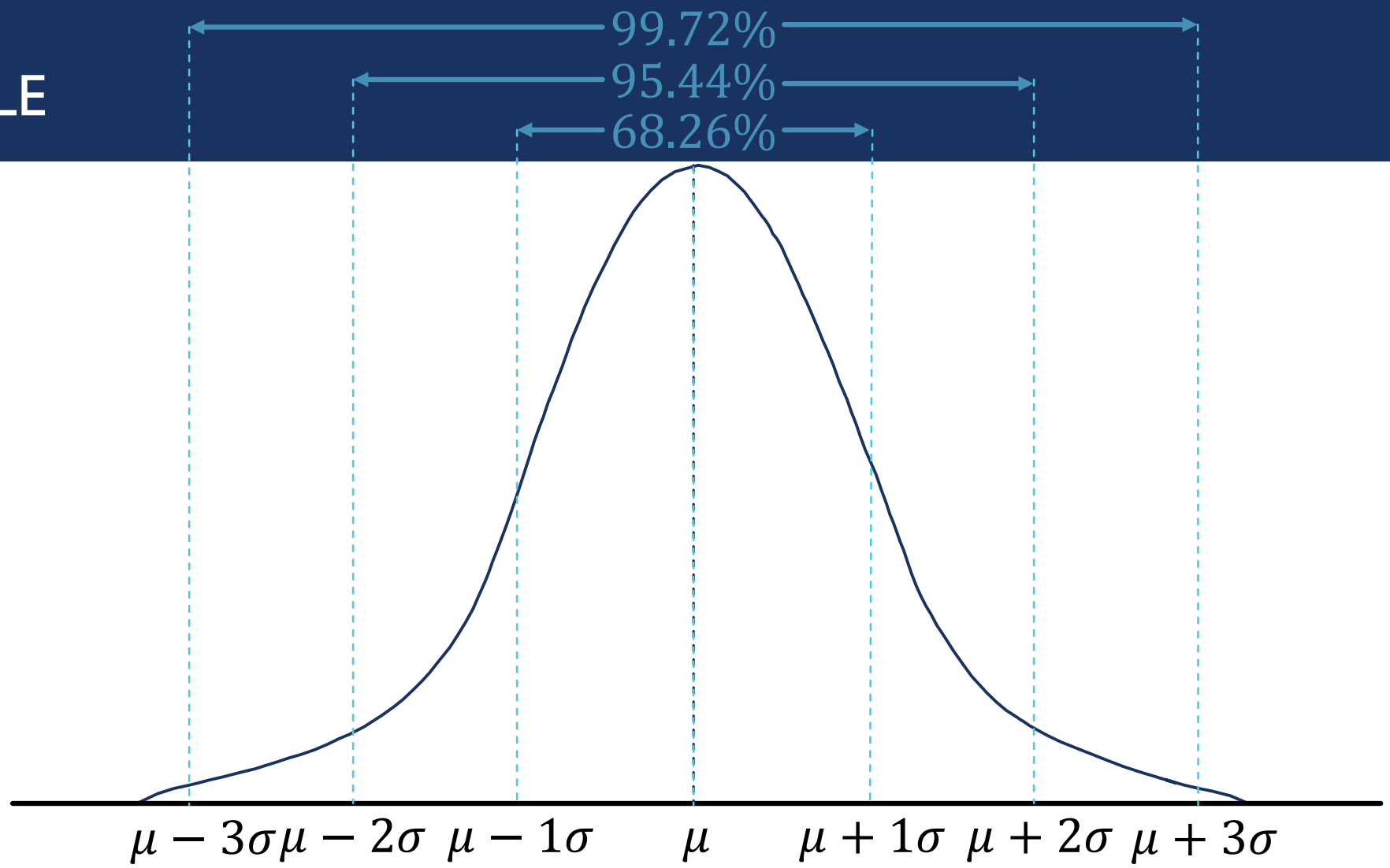
- The sampling distribution of $\hat{p}$ plays a key role in computing the margin of error for this interval estimate.
- The **sampling distribution of $\hat{p}$** is approximately the **Normal distribution** whenever $np \geq 5$ and $n(1-p) \geq 5$.

SAMPLING DISTRIBUTION OF $\hat{p}$

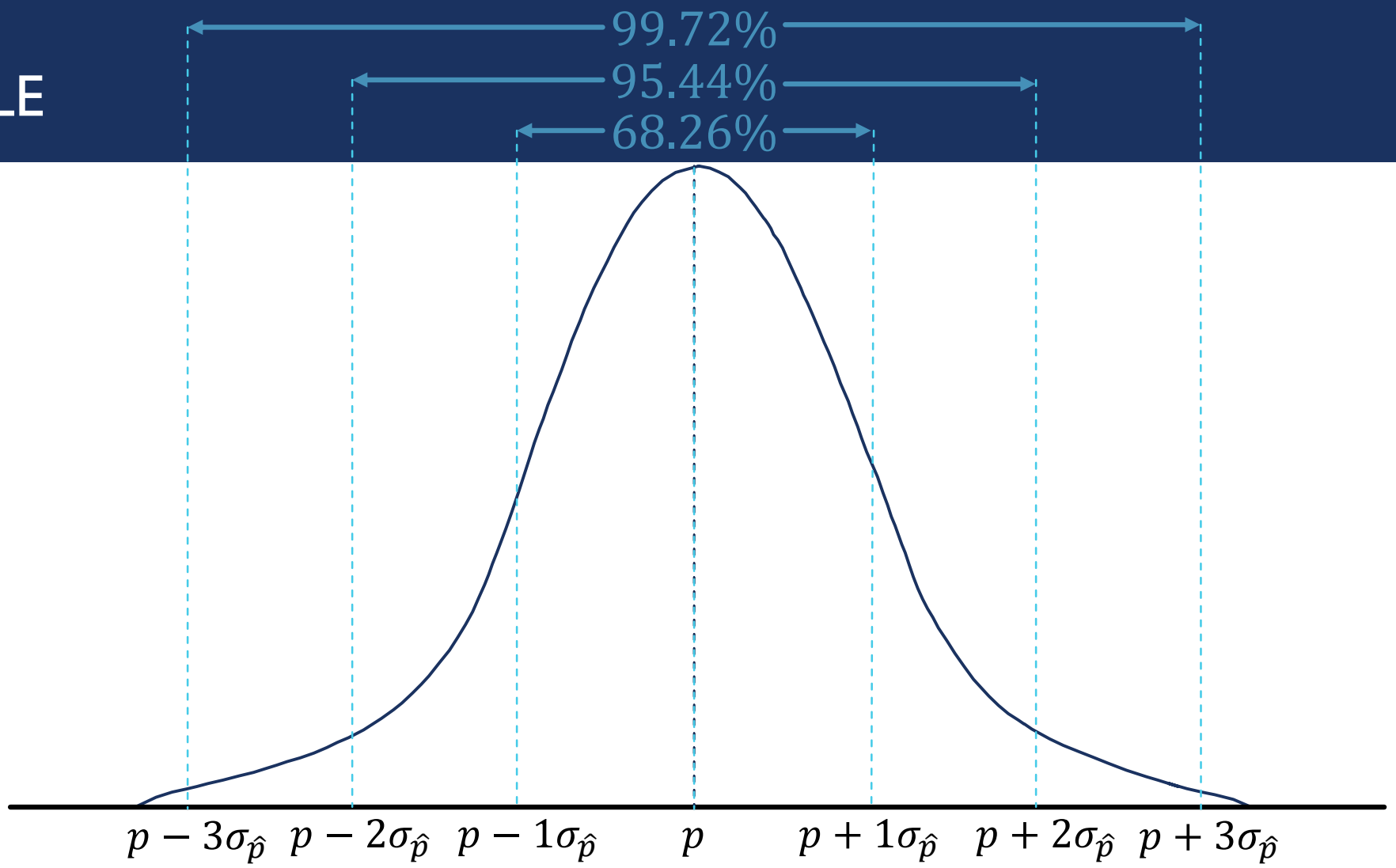
- The **sampling distribution of $\hat{p}$** is approximately the **Normal distribution** whenever $np \geq 5$ and $n(1-p) \geq 5$.



EMPIRICAL RULE

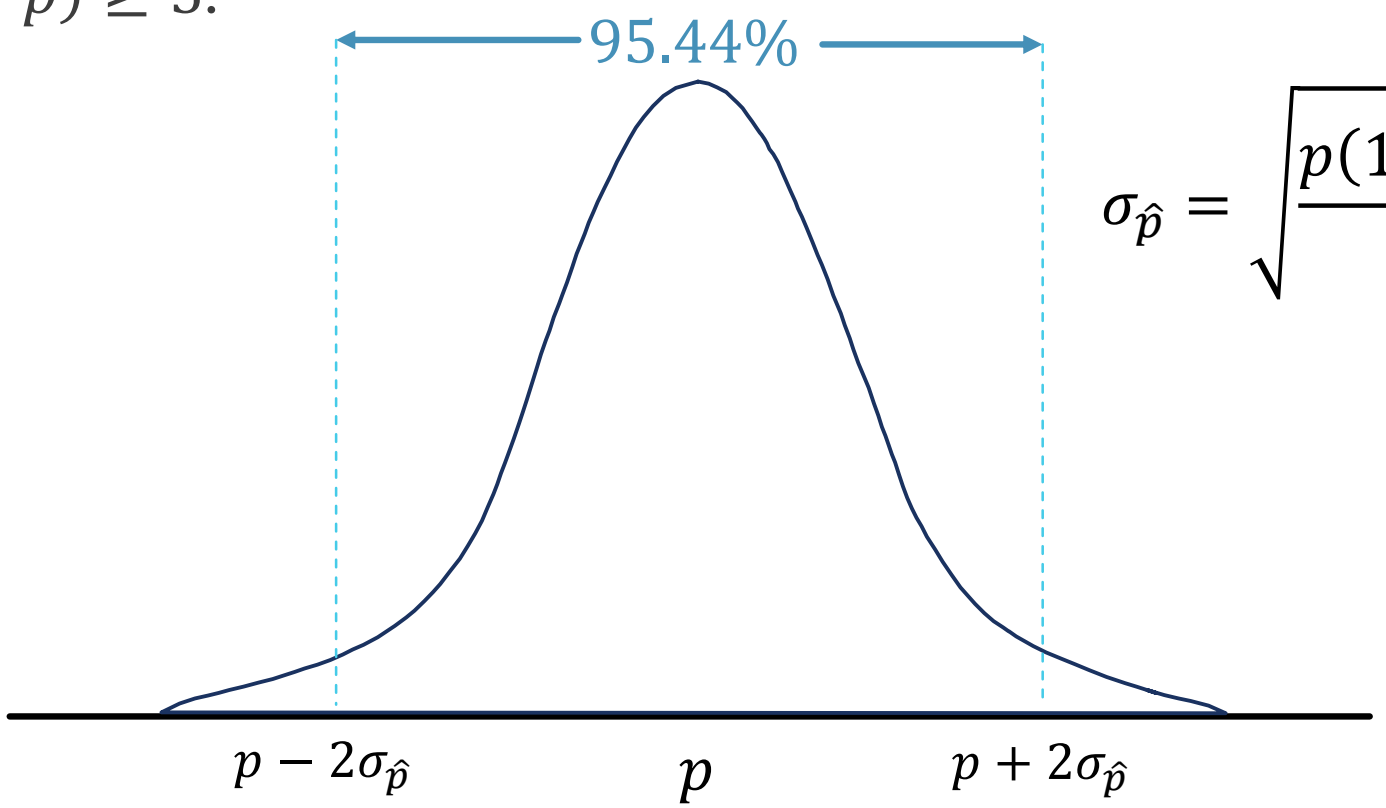


EMPIRICAL RULE



SAMPLING DISTRIBUTION OF $\hat{p}$

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

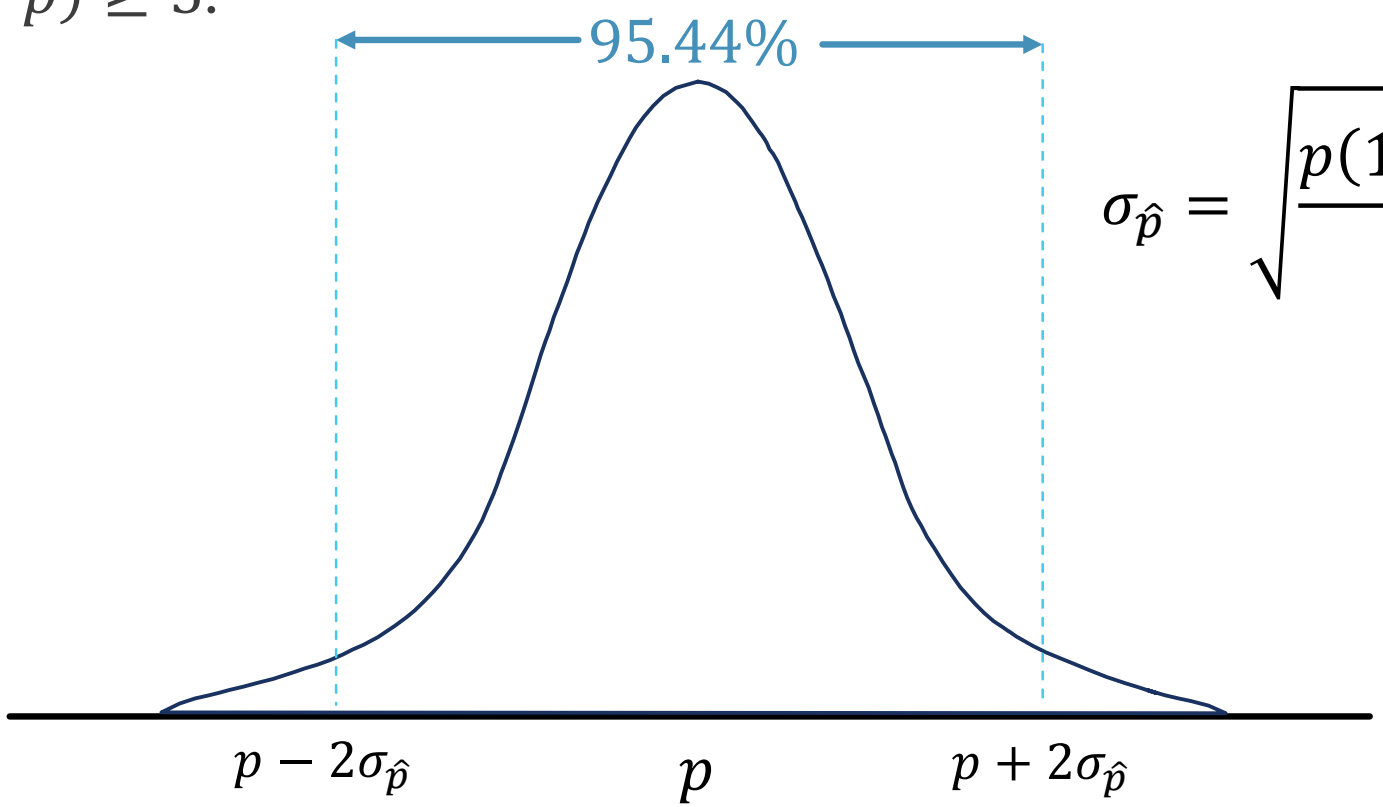


$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

SAMPLING DISTRIBUTION OF $\hat{p}$

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

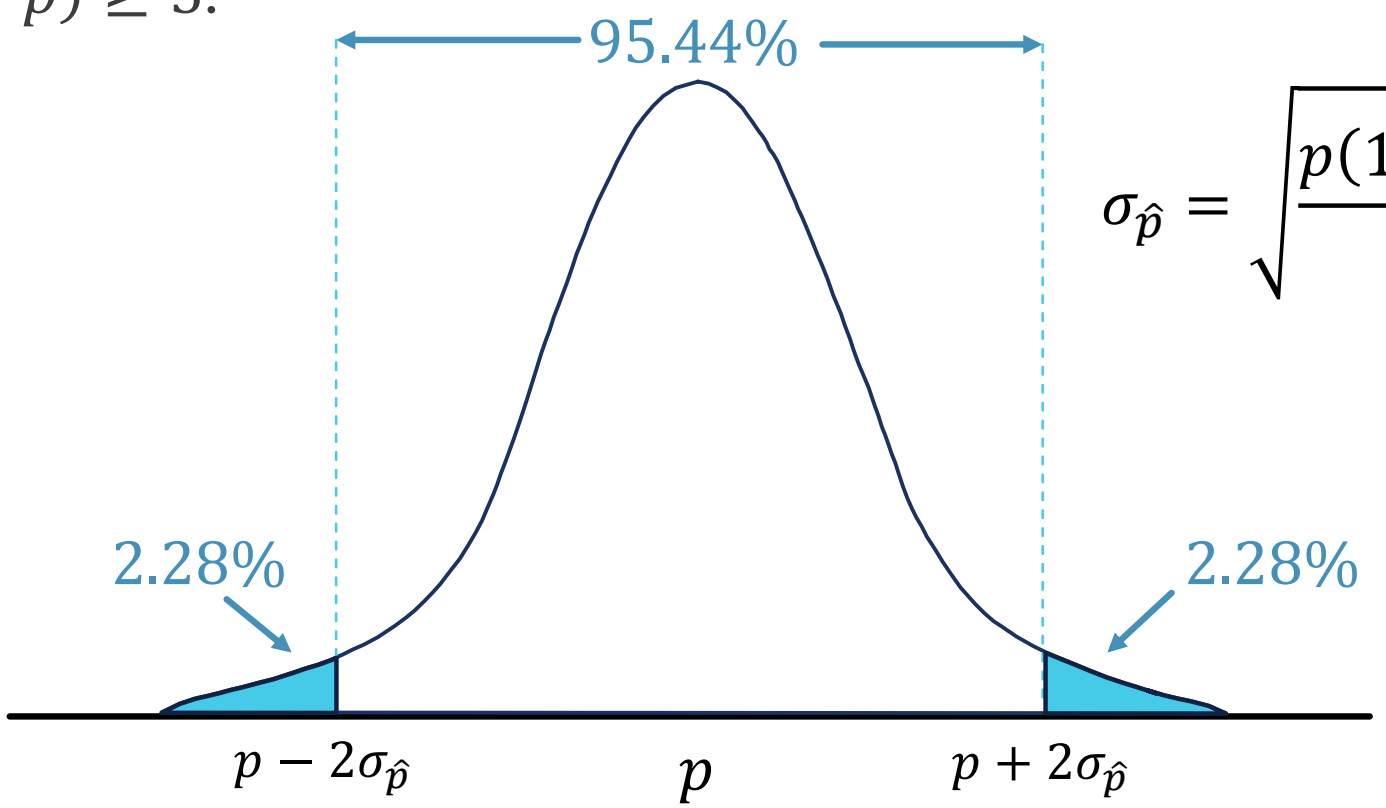
$$\hat{p} \pm \text{Margin of Error}$$



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

SAMPLING DISTRIBUTION OF $\hat{p}$

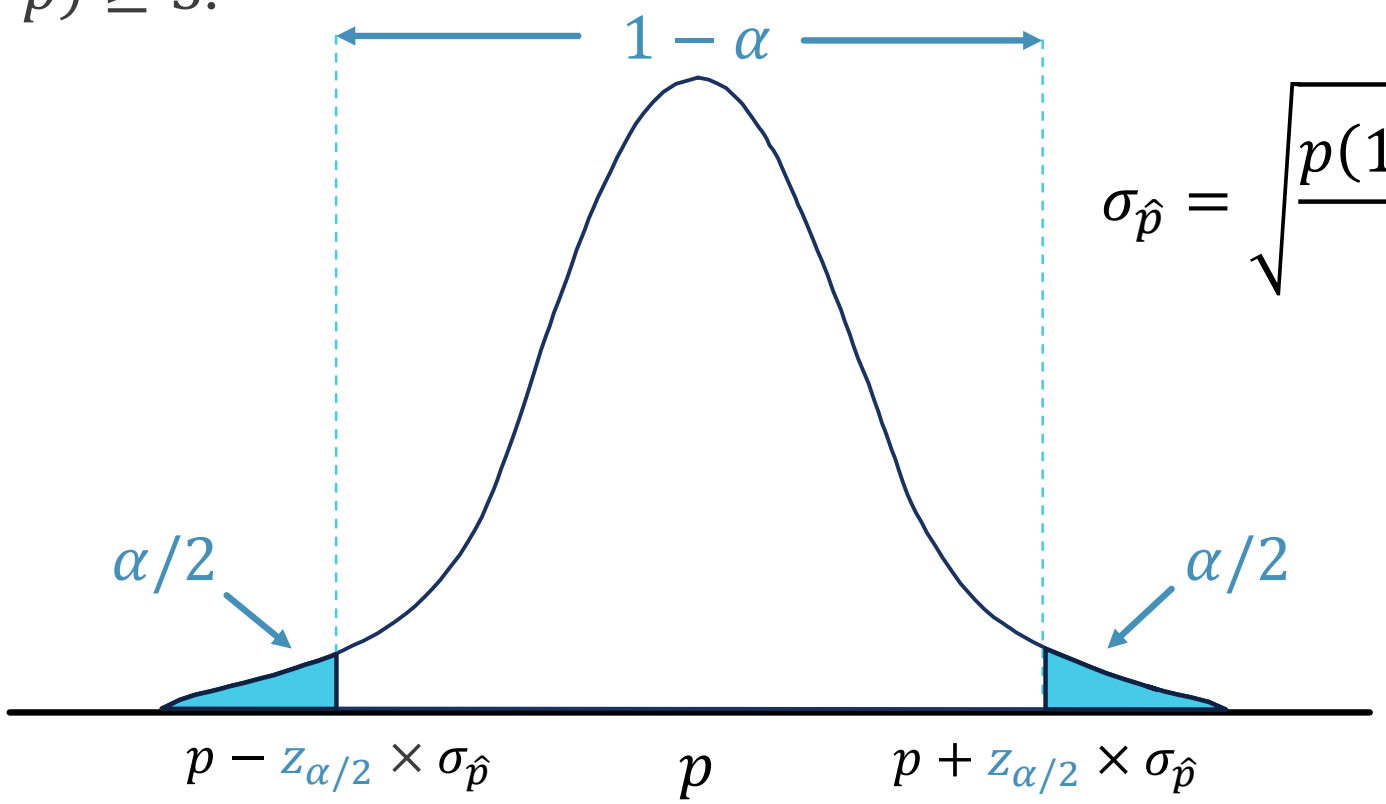
- The **sampling distribution of $\hat{p}$** is approximately the **Normal distribution** whenever $np \geq 5$ and $n(1-p) \geq 5$.



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

SAMPLING DISTRIBUTION OF $\hat{p}$

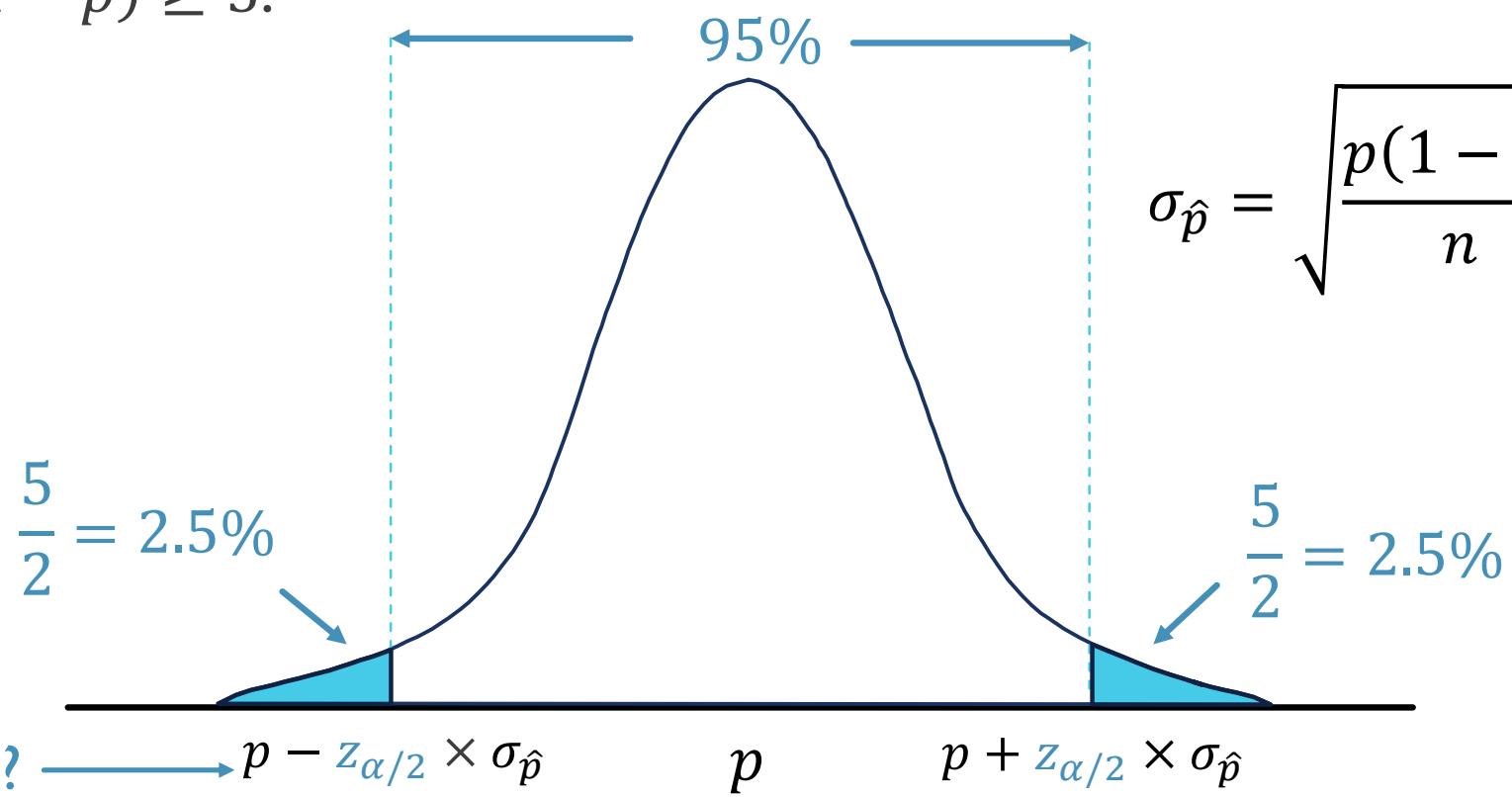
- The **sampling distribution of $\hat{p}$** is approximately the **Normal distribution** whenever $np \geq 5$ and $n(1-p) \geq 5$.



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

SAMPLING DISTRIBUTION OF $\hat{p}$

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



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

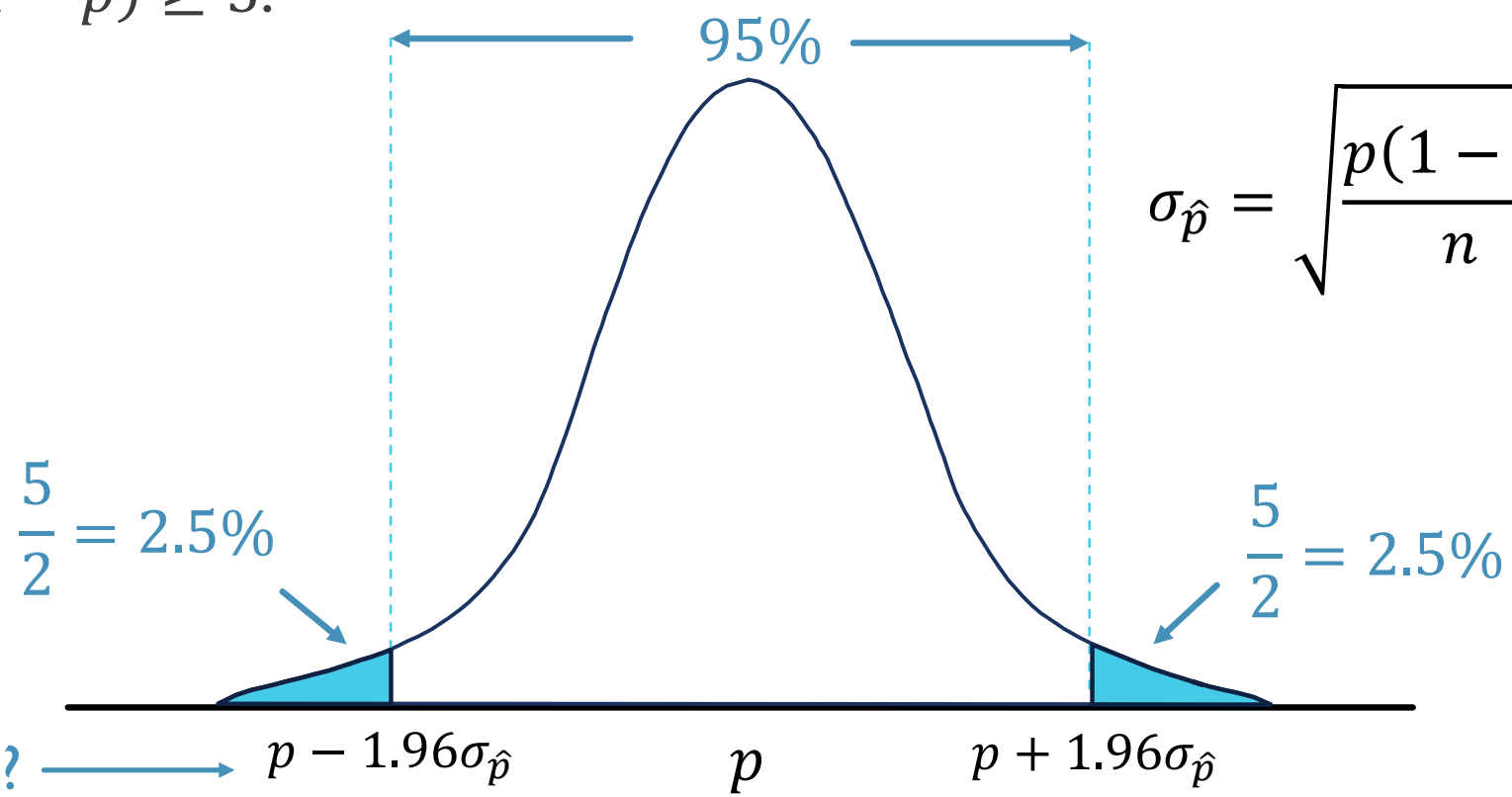
HOW TO CALCULATE $z_{\alpha/2}$

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 $\hat{p}$

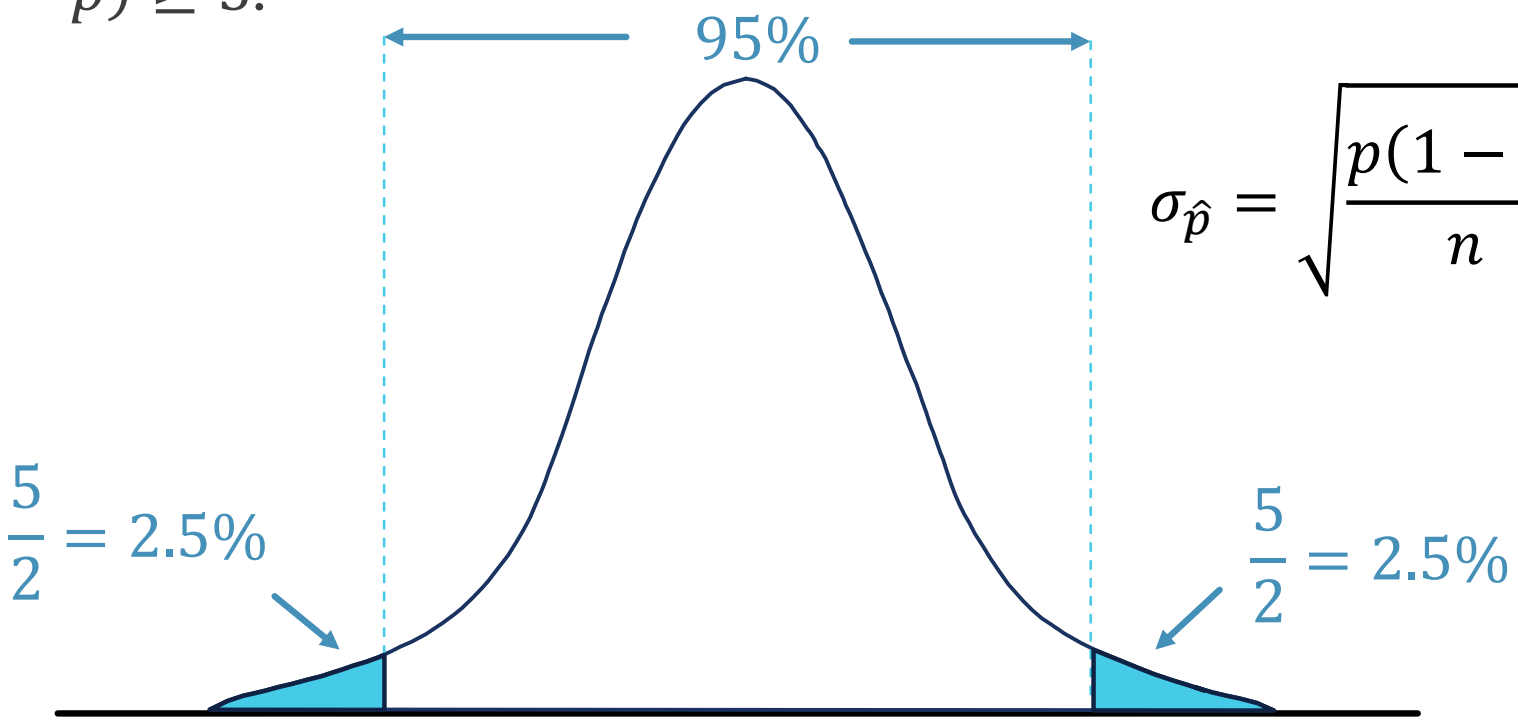
- The **sampling distribution of $\hat{p}$** is approximately the **Normal distribution** whenever $np \geq 5$ and $n(1-p) \geq 5$.



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

SAMPLING DISTRIBUTION OF $\hat{p}$

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



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

CONFIDENCE INTERVAL FOR $\hat{p}$

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

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

CONFIDENCE INTERVAL FOR $\hat{p}$

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

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

Do not know p!

CONFIDENCE INTERVAL FOR $\hat{p}$

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

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

- When you estimate a standard deviation of a statistic (in this case $\sigma_{\hat{p}}$) it is now called a **standard error**.

CONFIDENCE INTERVAL FOR $\hat{p}$

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

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

Standard error of $\hat{p}$!

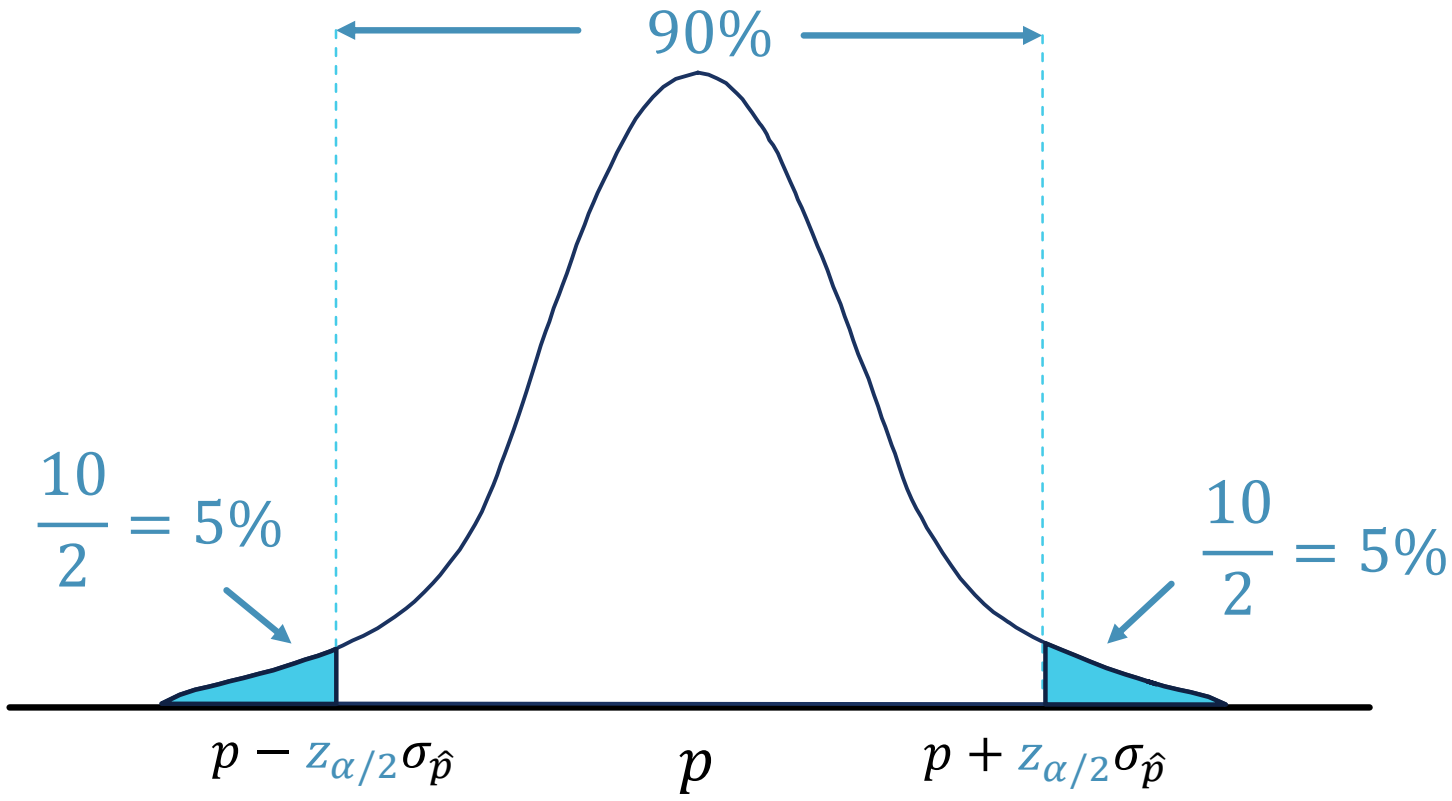
- When you estimate a standard deviation of a statistic (in this case $\sigma_{\hat{p}}$) it is now called a **standard error**.

CONFIDENCE INTERVAL BIKE DATA EXAMPLE

- You think that people are more likely to rent a bike on a clear or cloudy day compared to misty / rain / snow. Your data is a sample of 731 days with 63% clear or cloudy. Build a 90% confidence interval for the true proportion of clear or cloudy days where your company operates.

CONFIDENCE INTERVAL BIKE DATA EXAMPLE

- You think that people are more likely to rent a bike on a clear or cloudy day compared to misty / rain / snow. Your data is a sample of 731 days with 63% clear or cloudy. Build a 90% confidence interval for the true proportion of clear or cloudy days where your company operates.



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

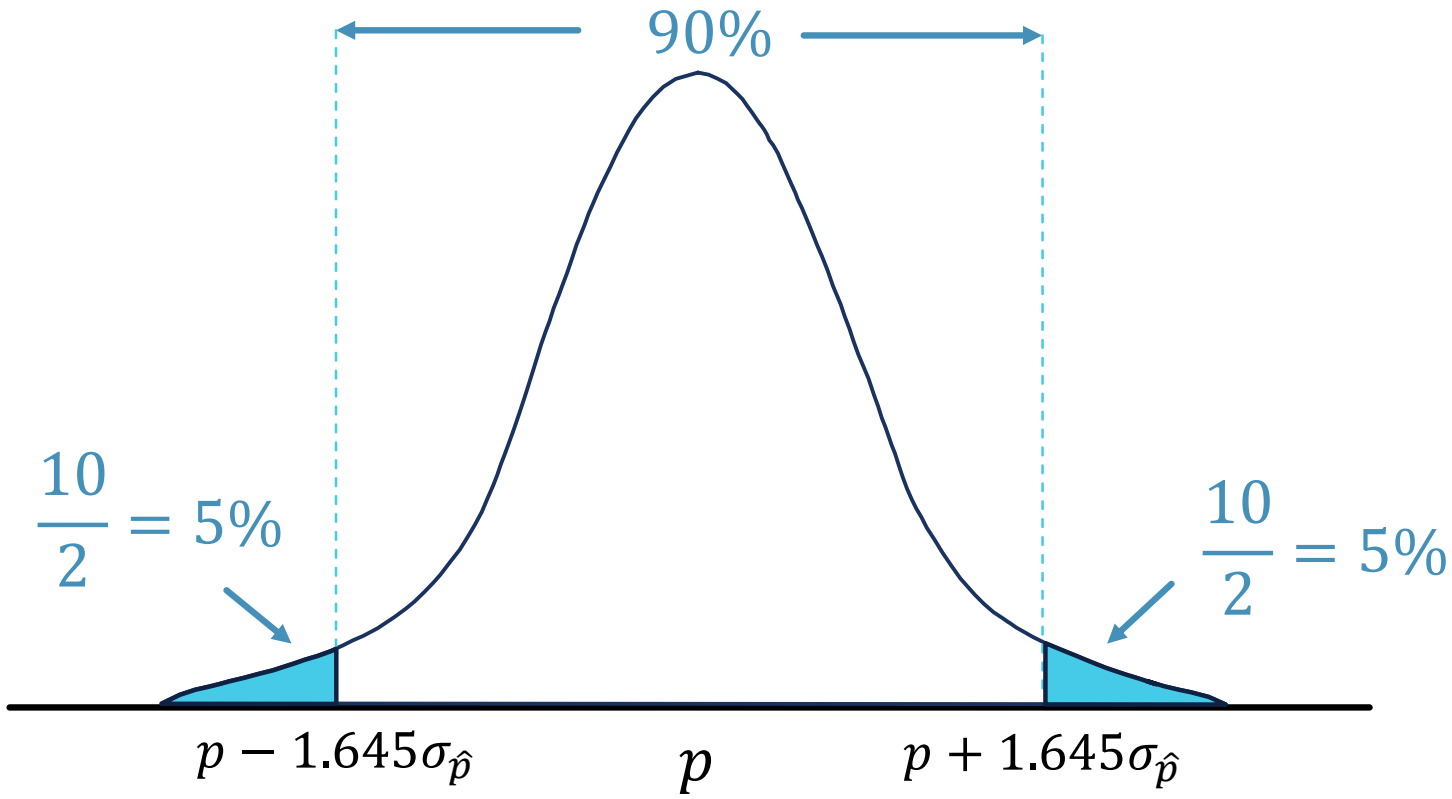
CONFIDENCE INTERVAL BIKE DATA EXAMPLE

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
.	.	.	.	.	.	.	.	.	.	.

$P(z \leq ?) = 0.05$

CONFIDENCE INTERVAL BIKE DATA EXAMPLE

- You think that people are more likely to rent a bike on a clear or cloudy day compared to misty / rain / snow. Your data is a sample of 731 days with 63% clear or cloudy. Build a 90% confidence interval for the true proportion of clear or cloudy days where your company operates.



What is $z_{\alpha/2}$? 1.645

CONFIDENCE INTERVAL BIKE DATA EXAMPLE

- You think that people are more likely to rent a bike on a clear or cloudy day compared to misty / rain / snow. Your data is a sample of 731 days with 63% clear or cloudy. Build a 90% confidence interval for the true proportion of clear or cloudy days where your company operates.

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

$$0.63 \pm 1.645\sqrt{\frac{0.63(1-0.63)}{731}}$$

$$0.63 \pm 1.645 \times 0.018$$

$$0.63 \pm 0.03$$

CONFIDENCE INTERVAL BIKE DATA EXAMPLE

- You think that people are more likely to rent a bike on a clear or cloudy day compared to misty / rain / snow. Your data is a sample of 731 days with 63% clear or cloudy. Build a 90% confidence interval for the true proportion of clear or cloudy days where your company operates.

$$0.63 \pm 0.03$$

OR

$$(0.60, 0.66)$$

SUMMARY

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

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

- When you estimate a standard deviation of a statistic it is now called a standard error.

INTERVAL ESTIMATION OF $\bar{x}$

INTERVAL ESTIMATION WITH DATA

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.

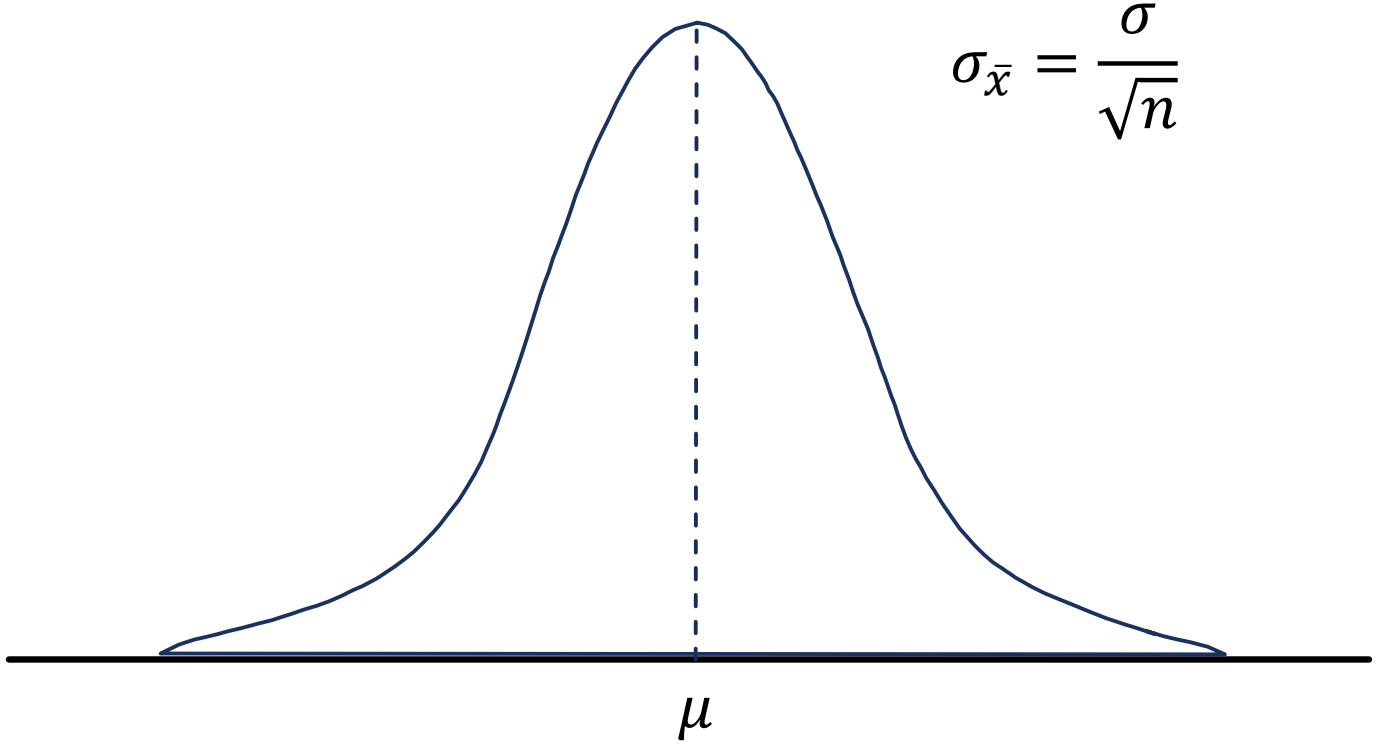
SAMPLING DISTRIBUTION OF $\bar{x}$

- 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 $\bar{x}$

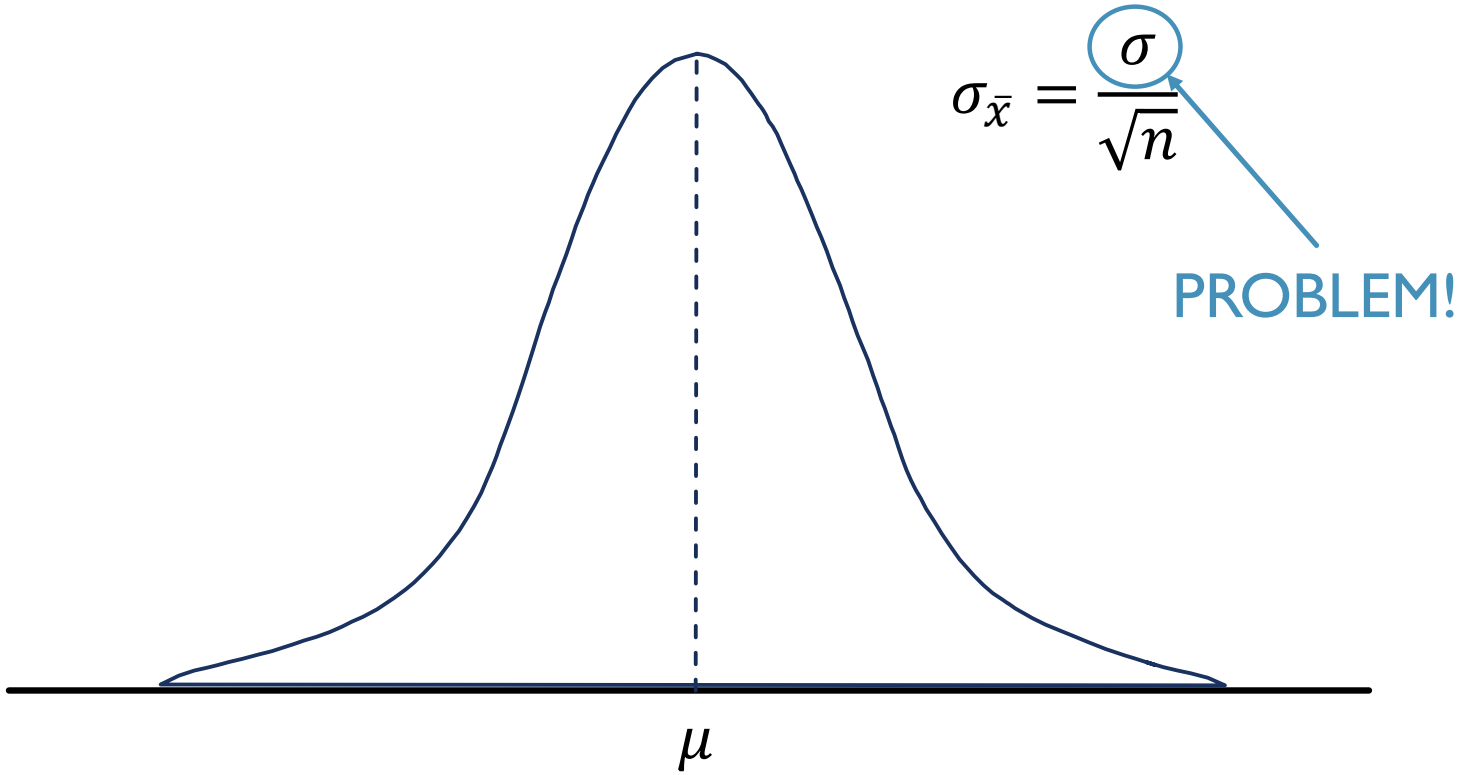
- The **sampling distribution of $\bar{x}$** is approximately the **Normal distribution** whenever $n \geq 50$.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$



SAMPLING DISTRIBUTION OF $\bar{x}$

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

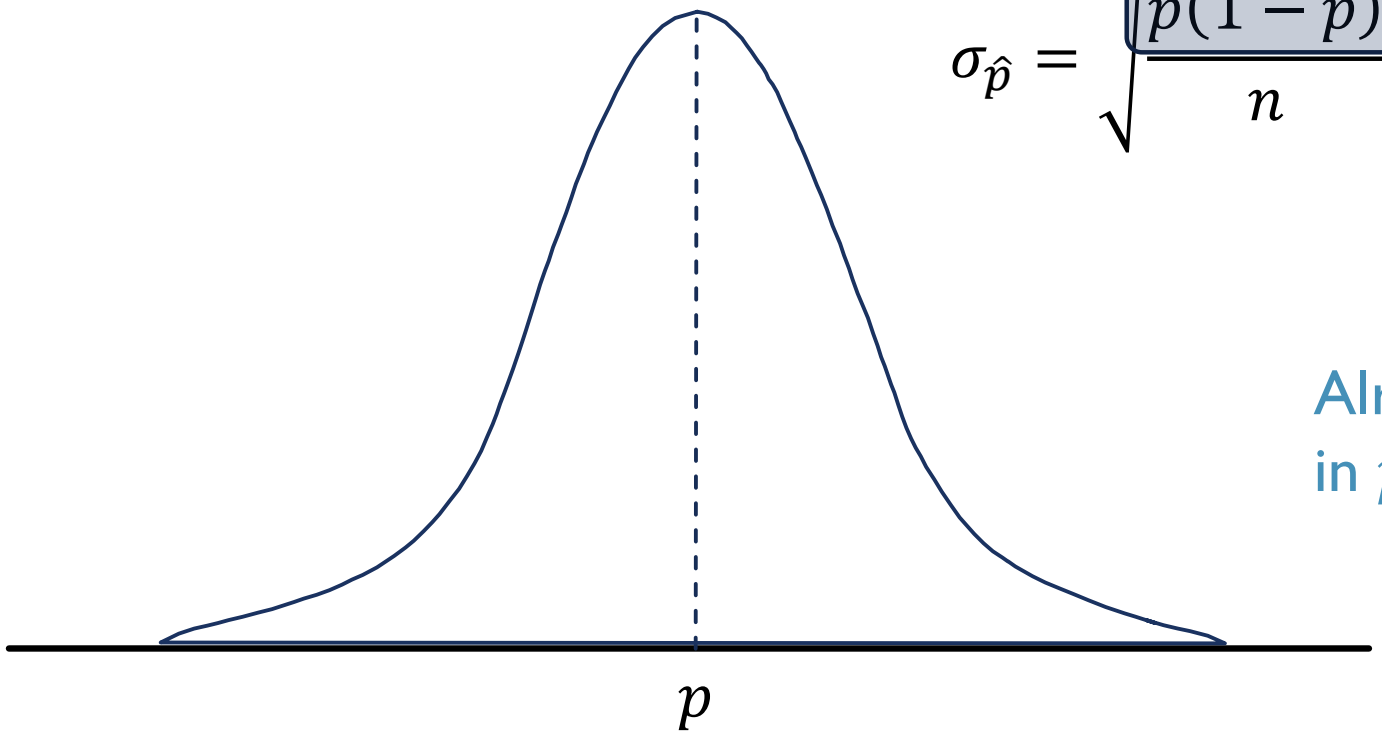


SAMPLING DISTRIBUTION OF $\hat{p}$

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

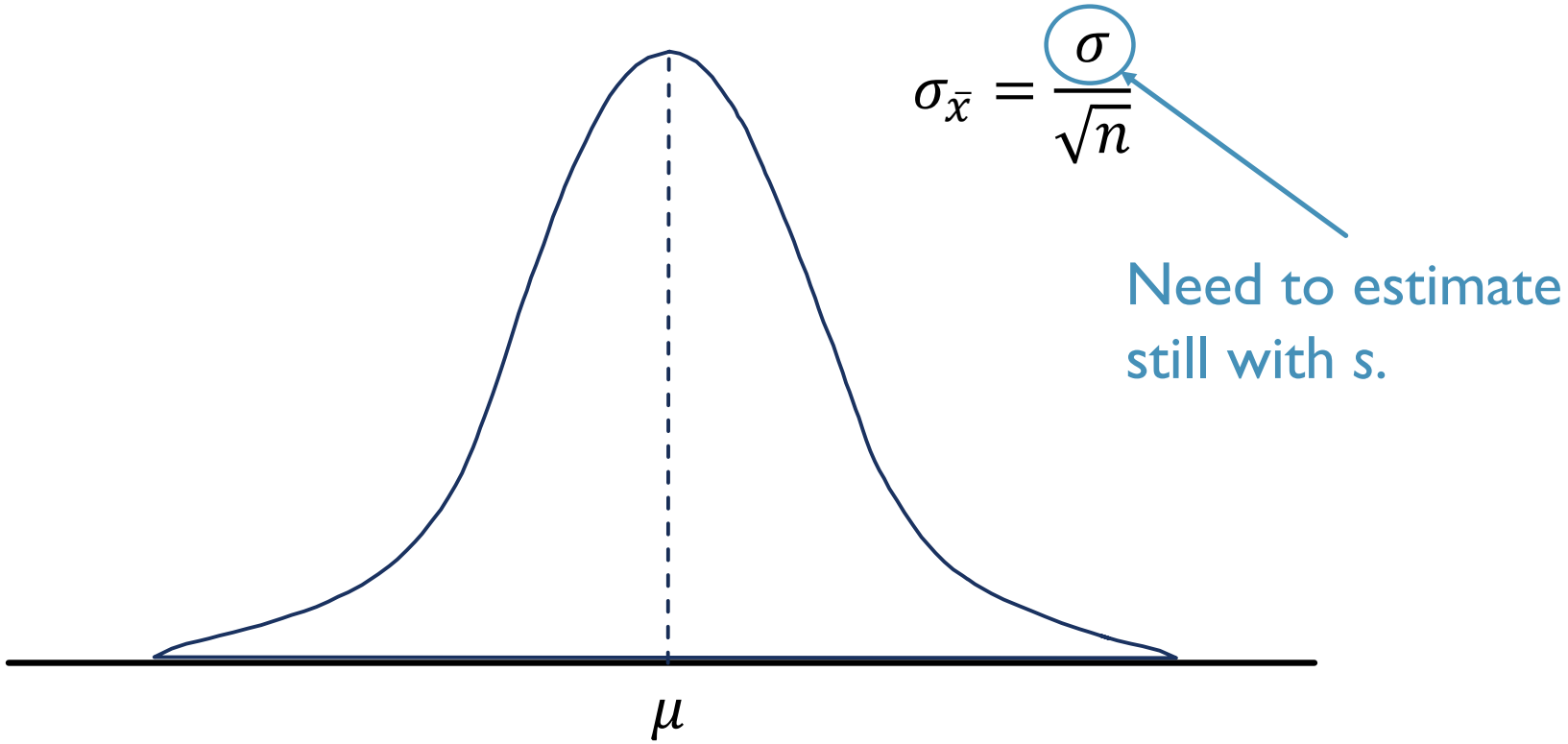
$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Already have estimate in $\hat{p}$



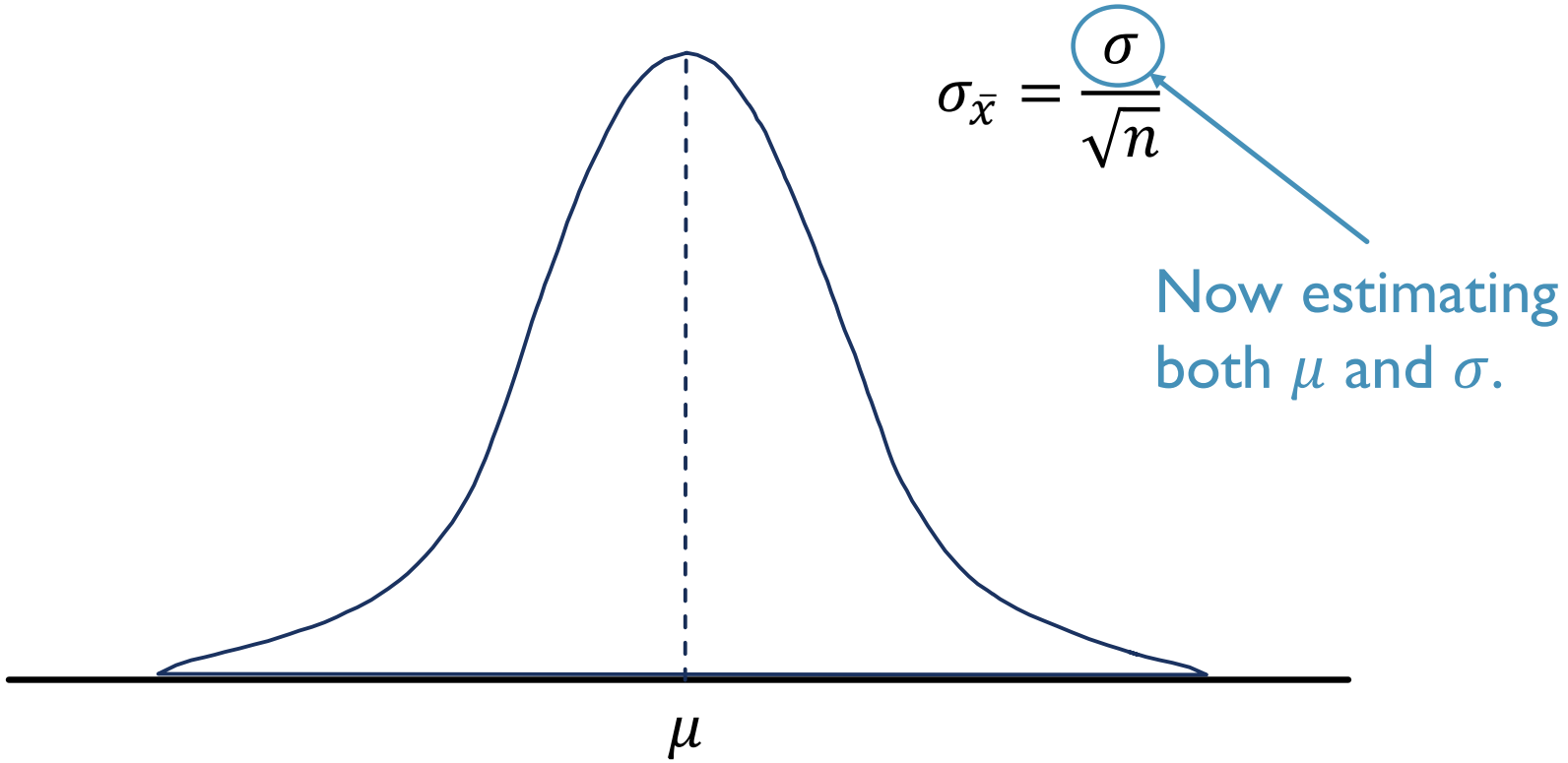
SAMPLING DISTRIBUTION OF $\bar{x}$

- 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 σ

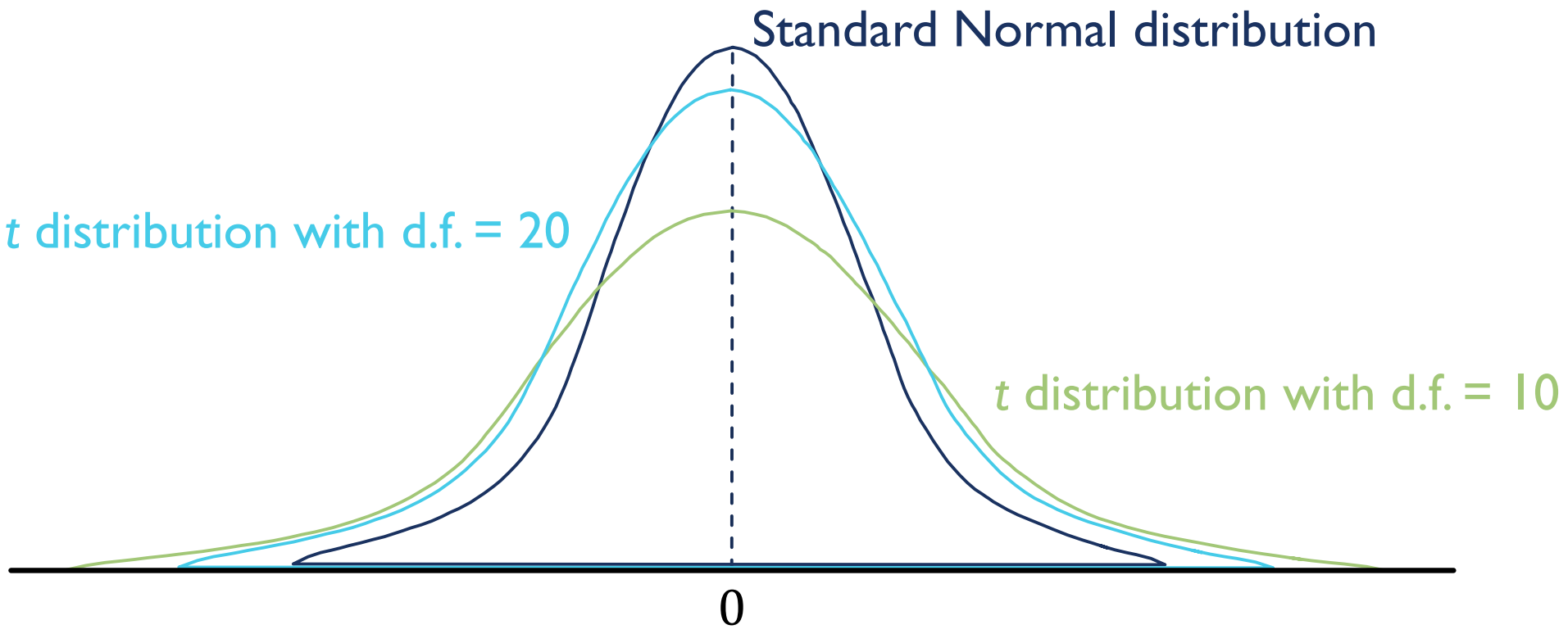
- Since we do not know the population standard deviation and need to estimate it with the sample standard deviation, we have added extra error into our calculations.
- Estimating two statistics has more error than just estimating one.
- Normal distribution is no longer a good approximation for the sampling 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: $\mathrm{d.f.} = n - 1$
- 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.

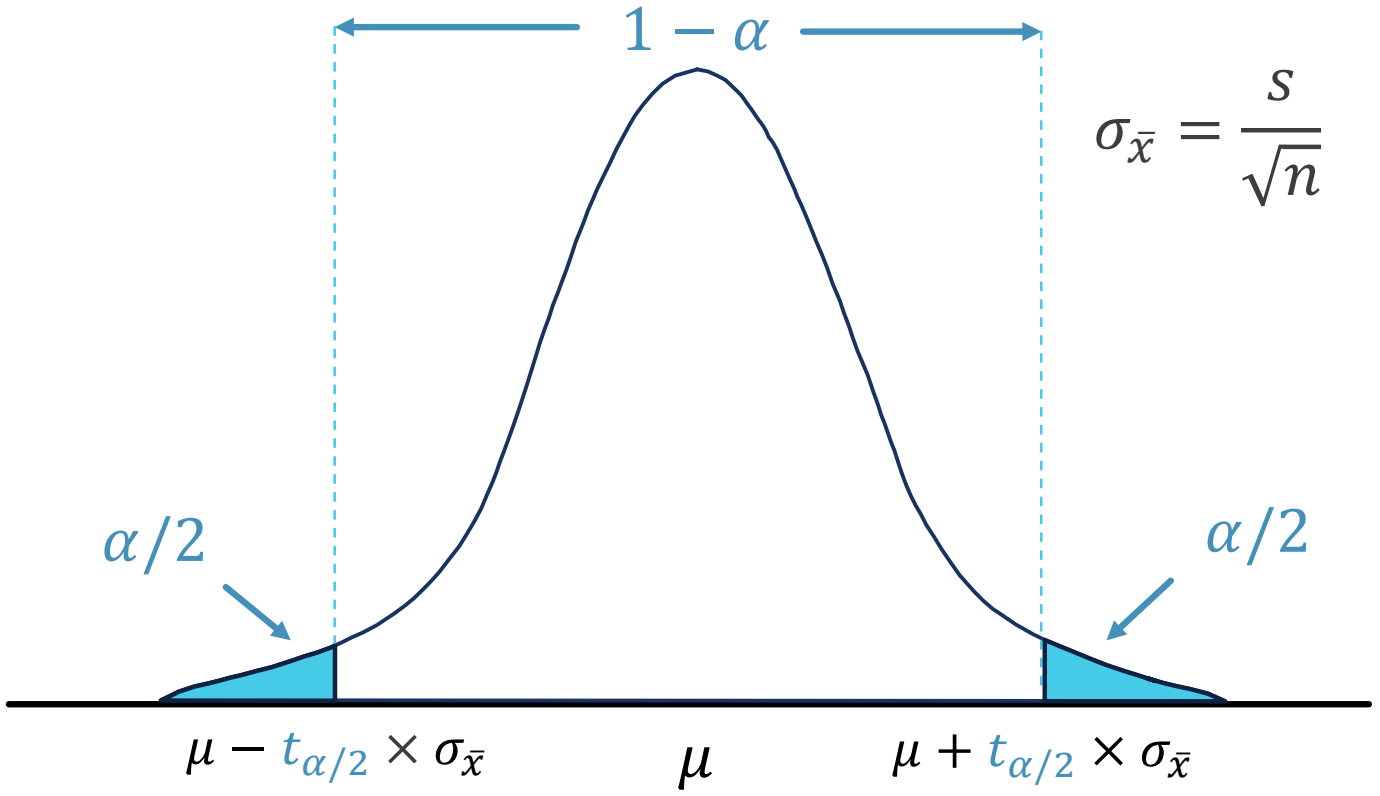
STANDARD NORMAL VS. *t* DISTRIBUTION

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



SAMPLING DISTRIBUTION OF $\bar{x}$

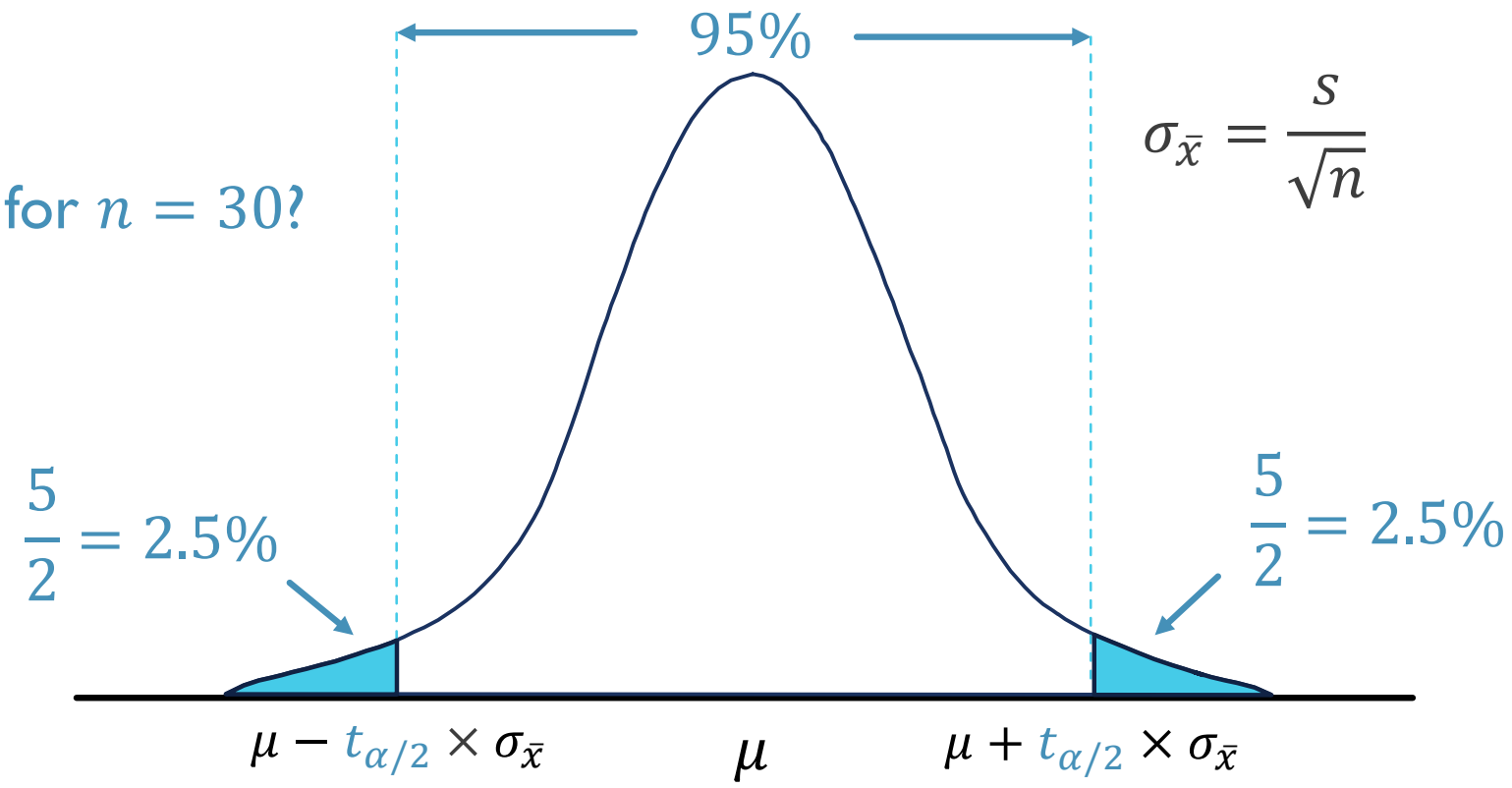
- Need to use the *t* distribution instead for the confidence intervals of $\bar{x}$.



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 $n = 30$?



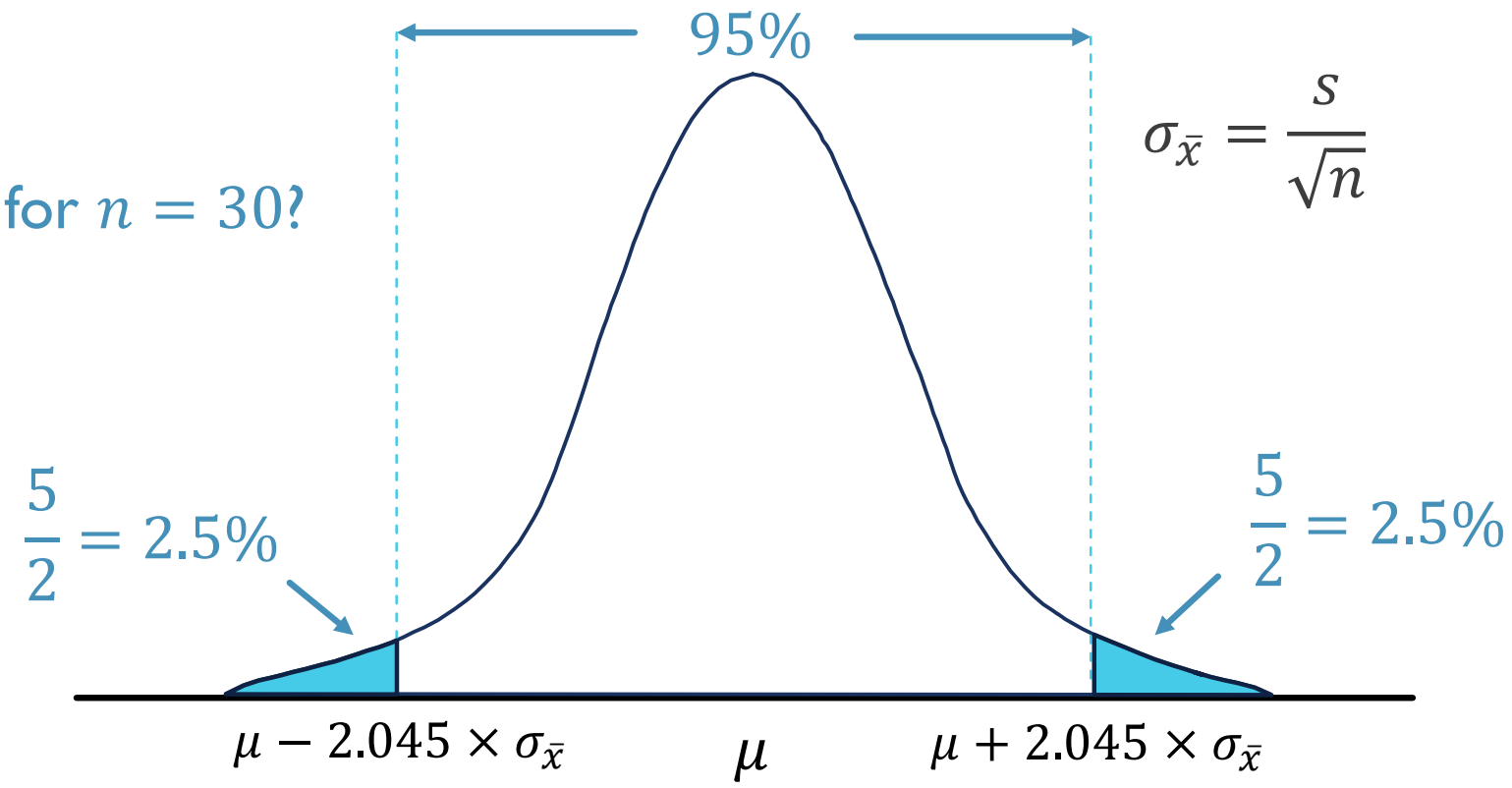
HOW TO CALCULATE $t_{\alpha/2}$

d.f.	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
.	.	.	.	.	.	.	.	.	.	.
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
.	.	.	.	.	.	.	.	.	.	.

SAMPLING DISTRIBUTION OF $\bar{x}$

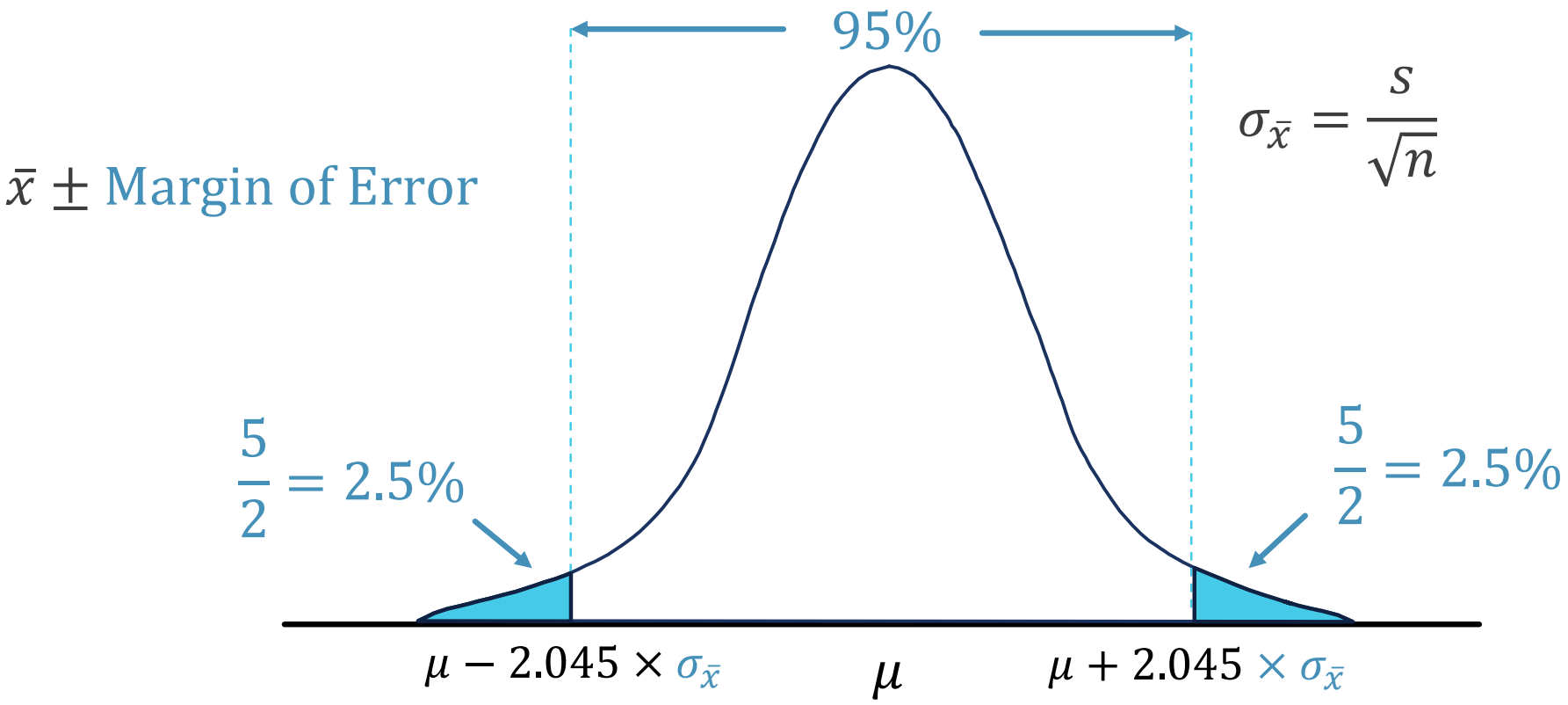
- Need to use the *t* distribution instead for the confidence intervals of $\bar{x}$.

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 α) is the following:

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

CONFIDENCE INTERVAL FOR $\bar{x}$

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

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

Standard error of $\bar{x}$!

- When you estimate a standard deviation of a statistic (in this case $\sigma_{\bar{x}}$) it is now called a **standard error**.

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**.

CONFIDENCE INTERVAL BIKE DATA EXAMPLE

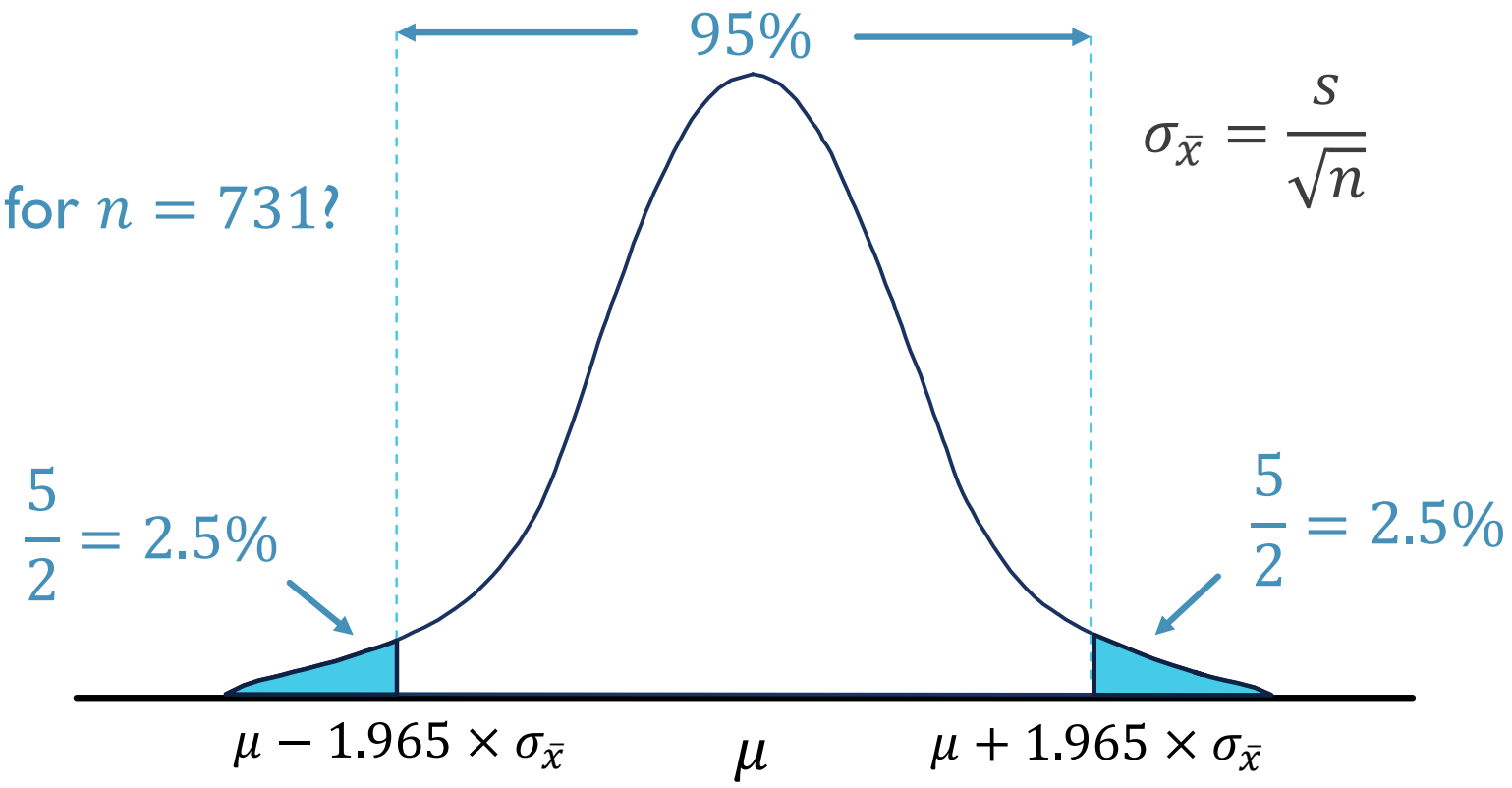
- The average daily number of total users is 4,504 with a standard deviation of 1,937 in our sample of 731 days. Build a 95% confidence interval for the average daily number of total users.

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 $n = 731$?

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$$



CONFIDENCE INTERVAL BIKE DATA EXAMPLE

- The average daily number of total users is 4,504 with a standard deviation of 1,937 in our sample of 731 days. Build a 95% confidence interval for the average daily number of total users.

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

$$4{,}504 \pm 1.965 \times \frac{1937}{\sqrt{731}}$$

$$4{,}504 \pm 140.8 \quad \text{OR} \quad (4{,}363.2, 4{,}644.8)$$

SUMMARY

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

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

- When you estimate a standard deviation of a statistic it is now called a standard error.

SAMPLE SIZE CALCULATION

INTERVAL ESTIMATION WITH DATA

REVERSING THE PROBLEM

- What if we wanted to know what sample size I 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:

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

SAMPLE SIZE NEEDED FOR $\hat{p}$

- Solve for sample size:

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

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

SAMPLE SIZE NEEDED FOR $\hat{p}$

- Solve for sample size:

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

Use p^* as best guess ahead of sampling

SAMPLE SIZE NEEDED FOR $\hat{p}$

- Sample size calculation:

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

- 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 $p^* = 0.5$.

CONFIDENCE INTERVAL BIKE DATA EXAMPLE

- You think that people are more likely to rent a bike on a clear or cloudy day compared to misty / rain / snow. Your data is a sample of 731 days with 63% clear or cloudy. Build a 90% confidence interval for the true proportion of clear or cloudy days where your company operates.

0.63 ± 0.03

OR

$(0.60, 0.66)$

What if that margin of error is too big for what the company wants?

CONFIDENCE INTERVAL BIKE DATA EXAMPLE

- You think that people are more likely to rent a bike on a clear or cloudy day compared to misty / rain / snow. You want to know the proportion of clear or cloudy days within 2% error. What sample size would we need for that?

CONFIDENCE INTERVAL BIKE DATA EXAMPLE

- You think that people are more likely to rent a bike on a clear or cloudy day compared to misty / rain / snow. You want to know the proportion of clear or cloudy days within 2% error for a 90% confidence interval. What sample size would we need for that?

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

$$n = \frac{(1.645^2)0.5(1-0.5)}{0.02^2}$$

$$n = 1{,}691.266$$

$$n \approx 1{,}692$$

SAMPLE SIZE NEEDED FOR $\bar{x}$

- Take the margin of error:

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

- Solve for sample size:

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

SAMPLE SIZE NEEDED FOR $\bar{x}$

- Solve for sample size:

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

Don't know s^2 ahead of sampling!

SAMPLE SIZE NEEDED FOR $\bar{x}$

- Solve for sample size:

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

Use s^{*2} as 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:

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

Typically use Normal distribution approximation

SAMPLE SIZE NEEDED FOR $\bar{x}$

- Solve for sample size:

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

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

SUMMARY

- To calculate the sample size needed for a confidence interval to have a certain margin of error, we can reverse the confidence interval equation to solve for sample size.
- Due to not knowing certain information ahead of time ($\hat{p}, s^2$), we need to use estimates of these values.
  - The best estimate for $\hat{p}$ is 0.5.
  - The best estimate for s^2 is typically taken from previous studies or a pilot study.