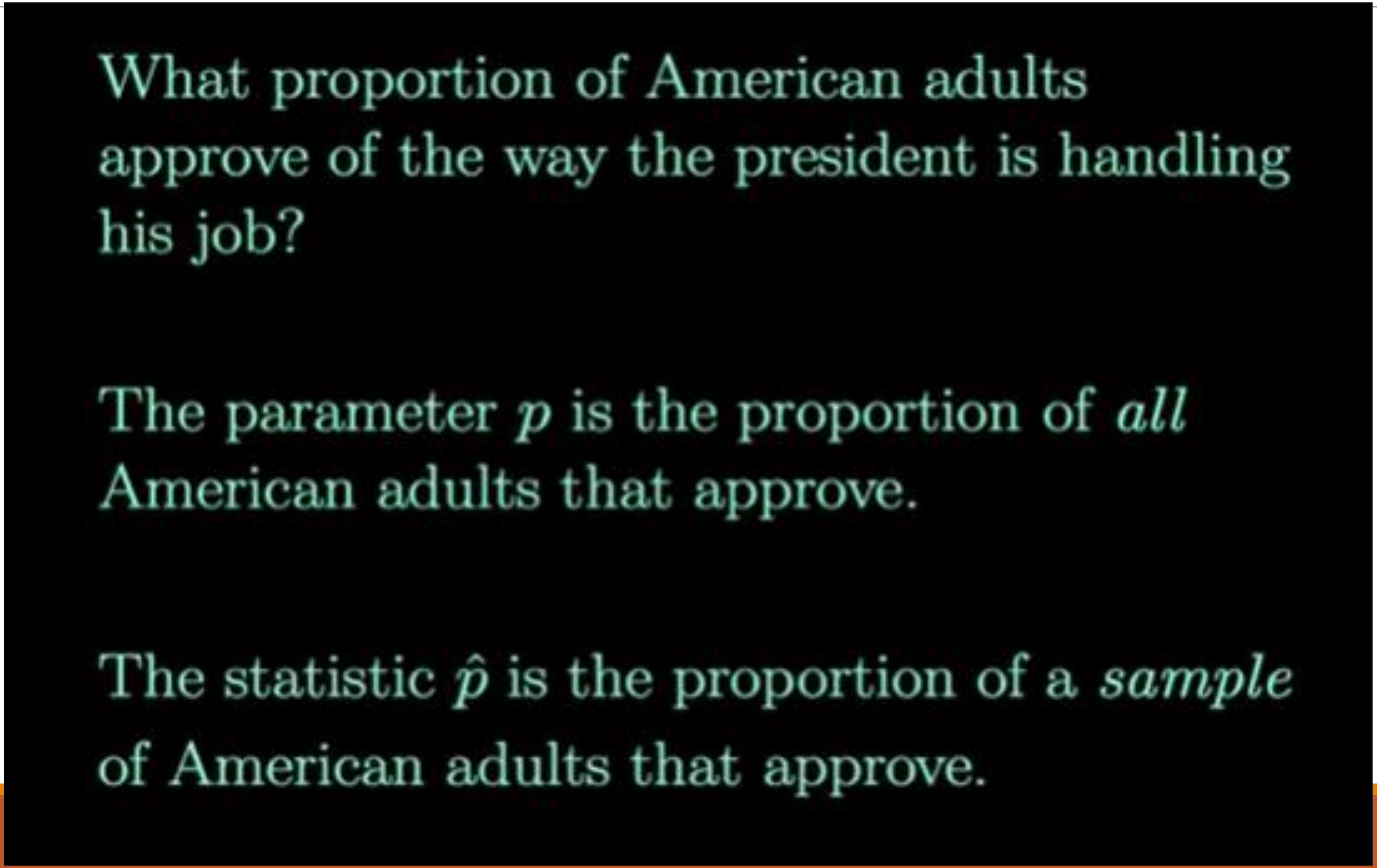
Statistic Inference for Proportions

ANISHA JOSEPH

Example

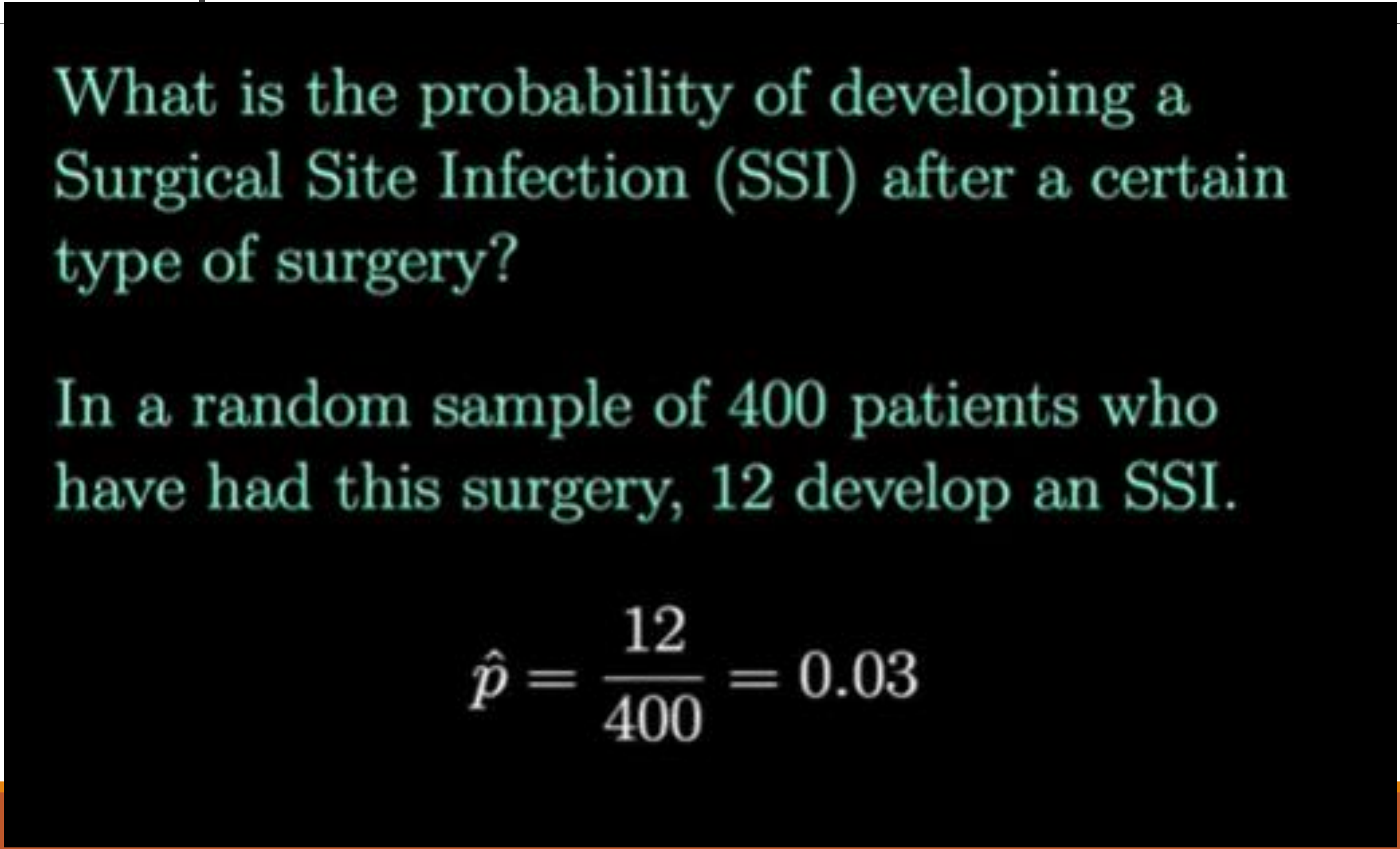


What proportion of American adults approve of the way the president is handling his job?

The parameter p is the proportion of *all* American adults that approve.

The statistic $\hat{p}$ is the proportion of a *sample* of American adults that approve.

Example

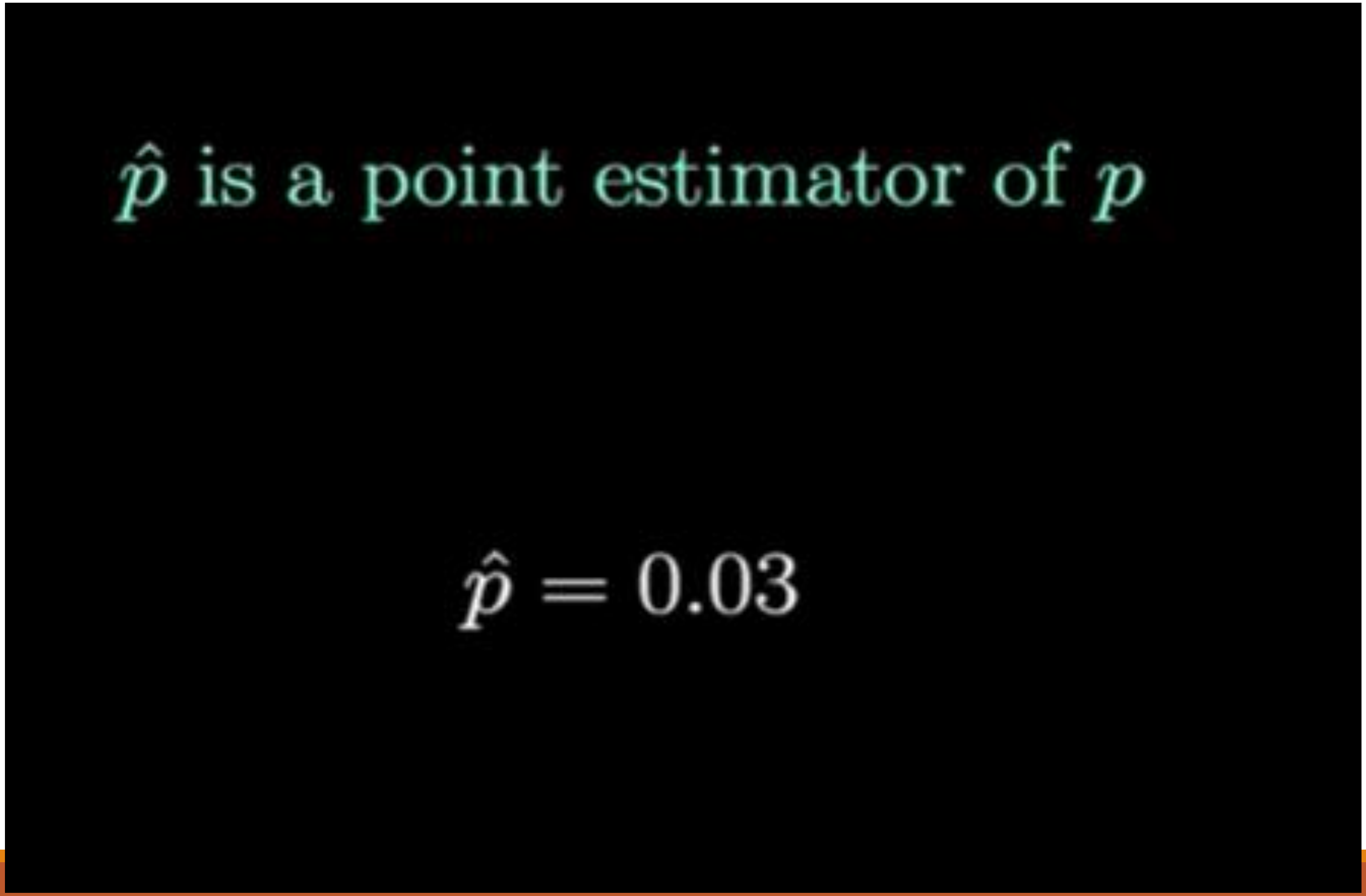


What is the probability of developing a Surgical Site Infection (SSI) after a certain type of surgery?

In a random sample of 400 patients who have had this surgery, 12 develop an SSI.

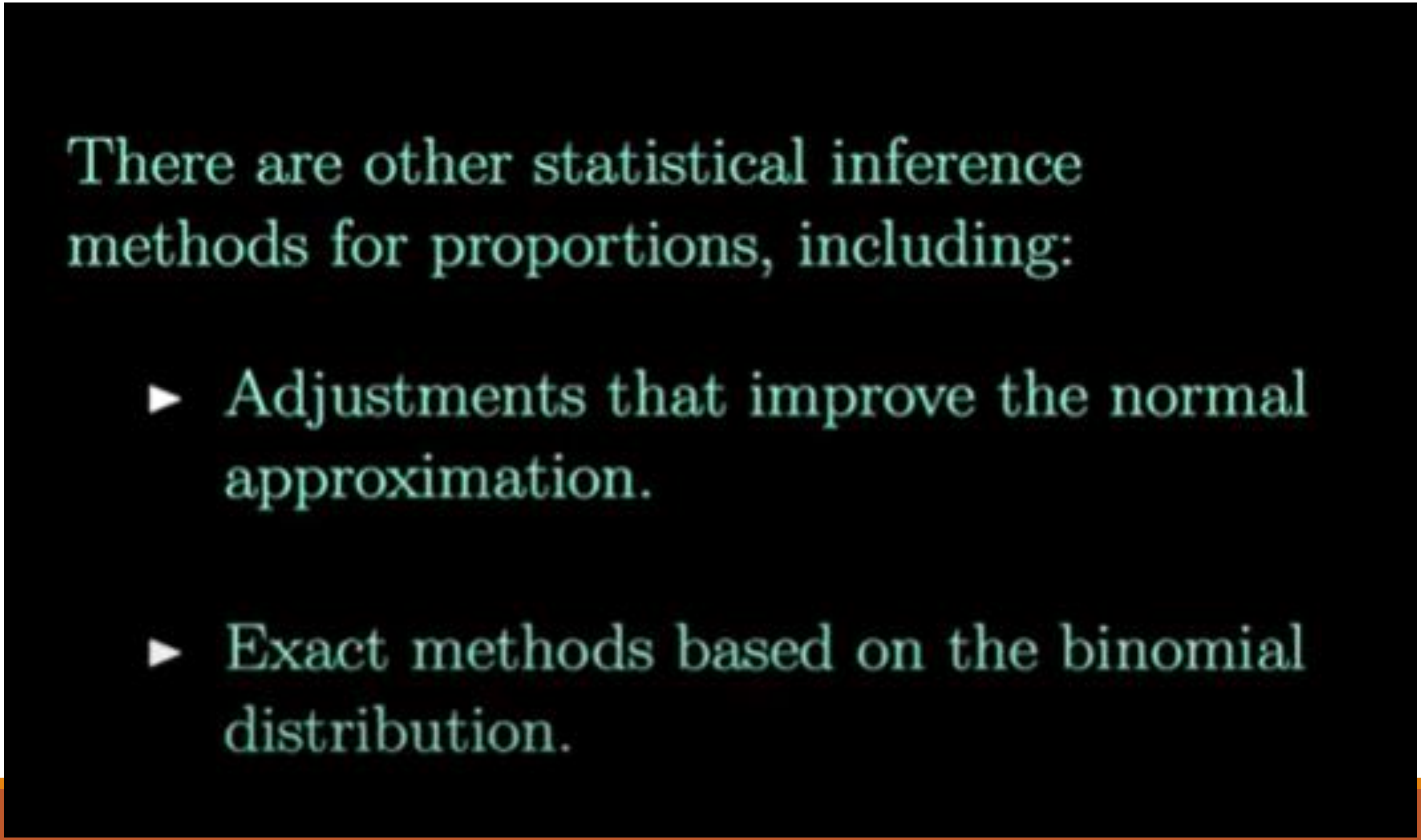
$$\hat{p} = \frac{12}{400} = 0.03$$

Example



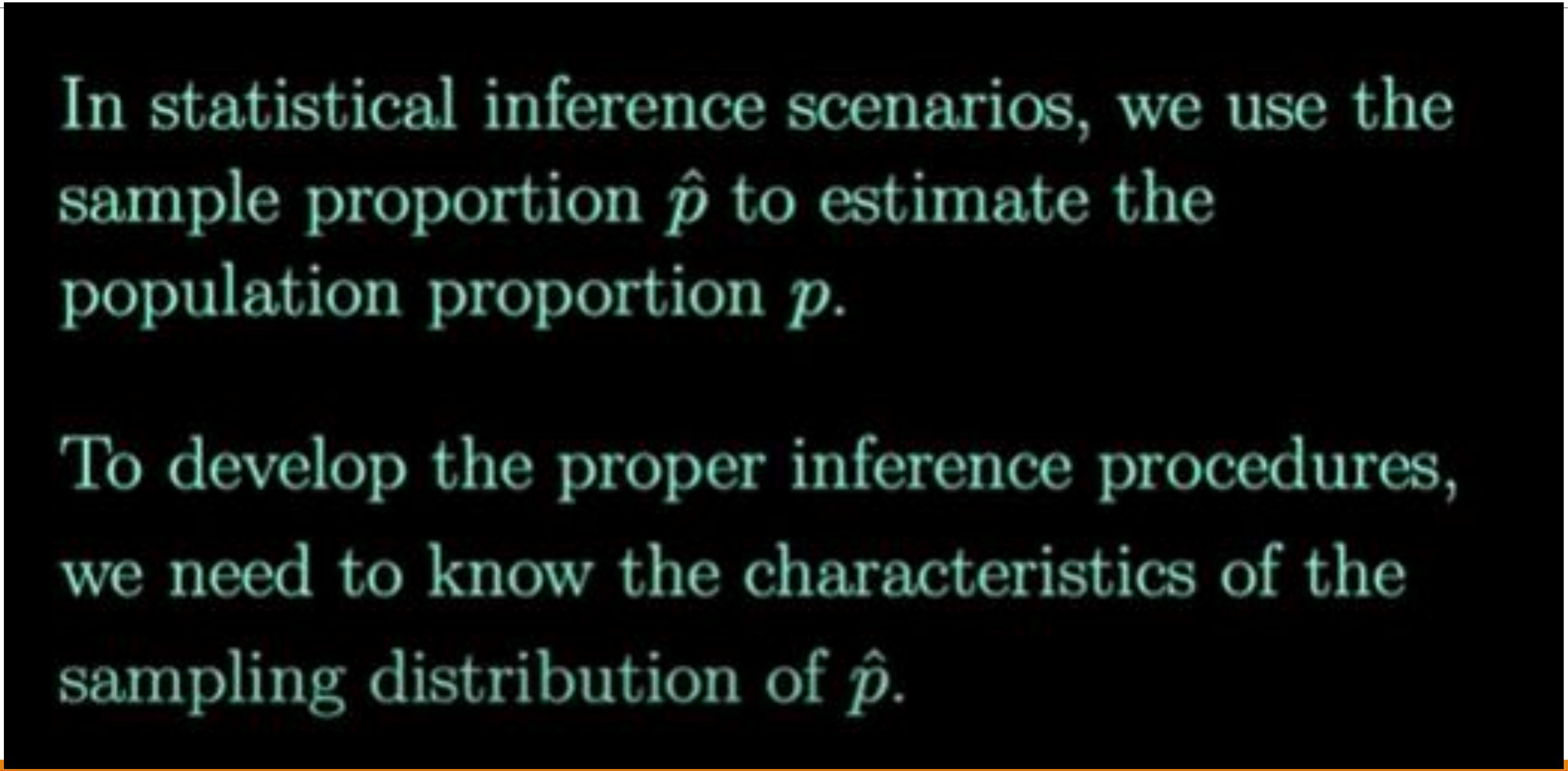
$\hat{p}$ is a point estimator of p

$$\hat{p} = 0.03$$



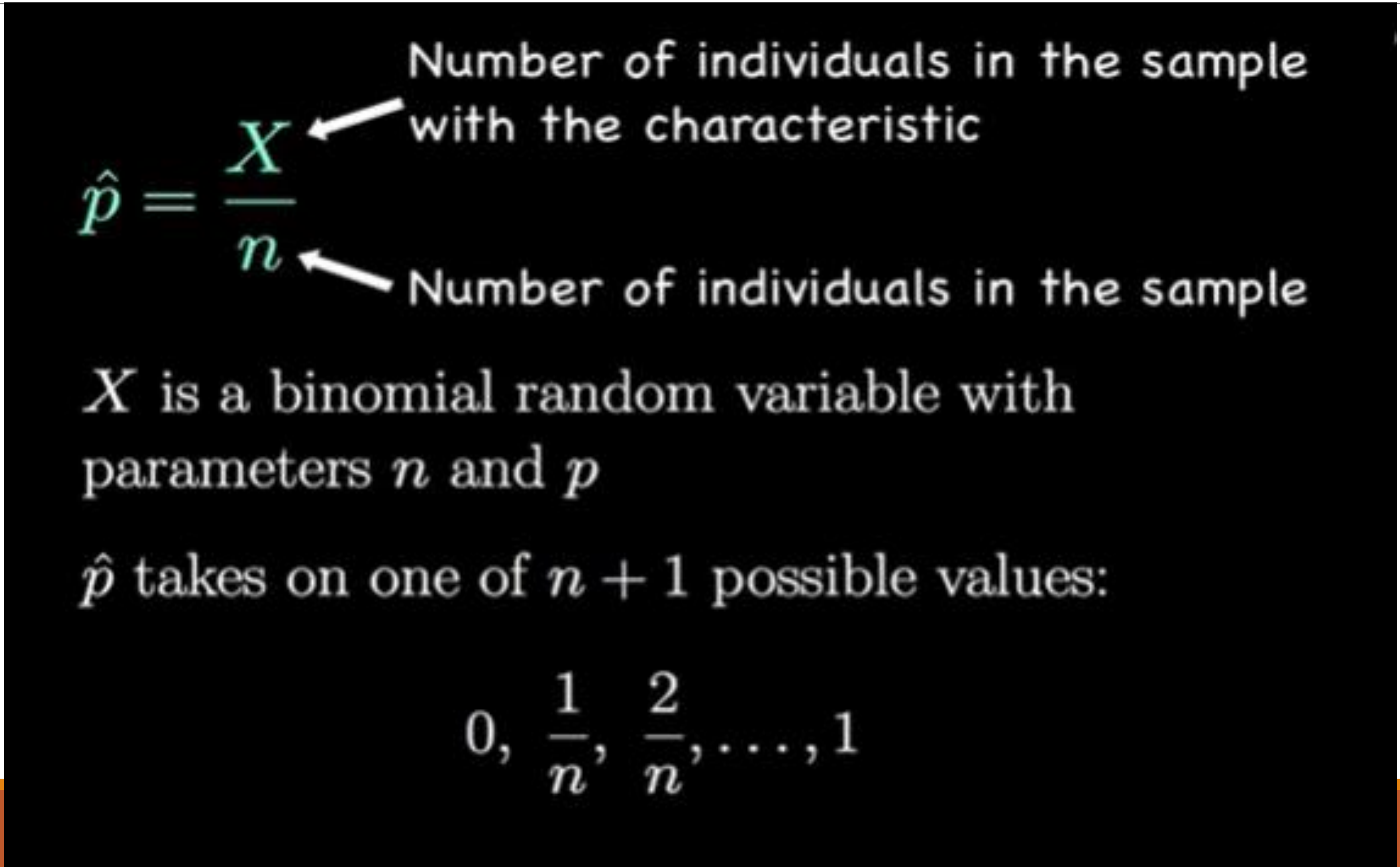
There are other statistical inference methods for proportions, including:

- Adjustments that improve the normal approximation.
- Exact methods based on the binomial distribution.



In statistical inference scenarios, we use the sample proportion $\hat{p}$ to estimate the population proportion p.

To develop the proper inference procedures, we need to know the characteristics of the sampling distribution of $\hat{p}$.



$$\hat{p} = \frac{X}{n}$$

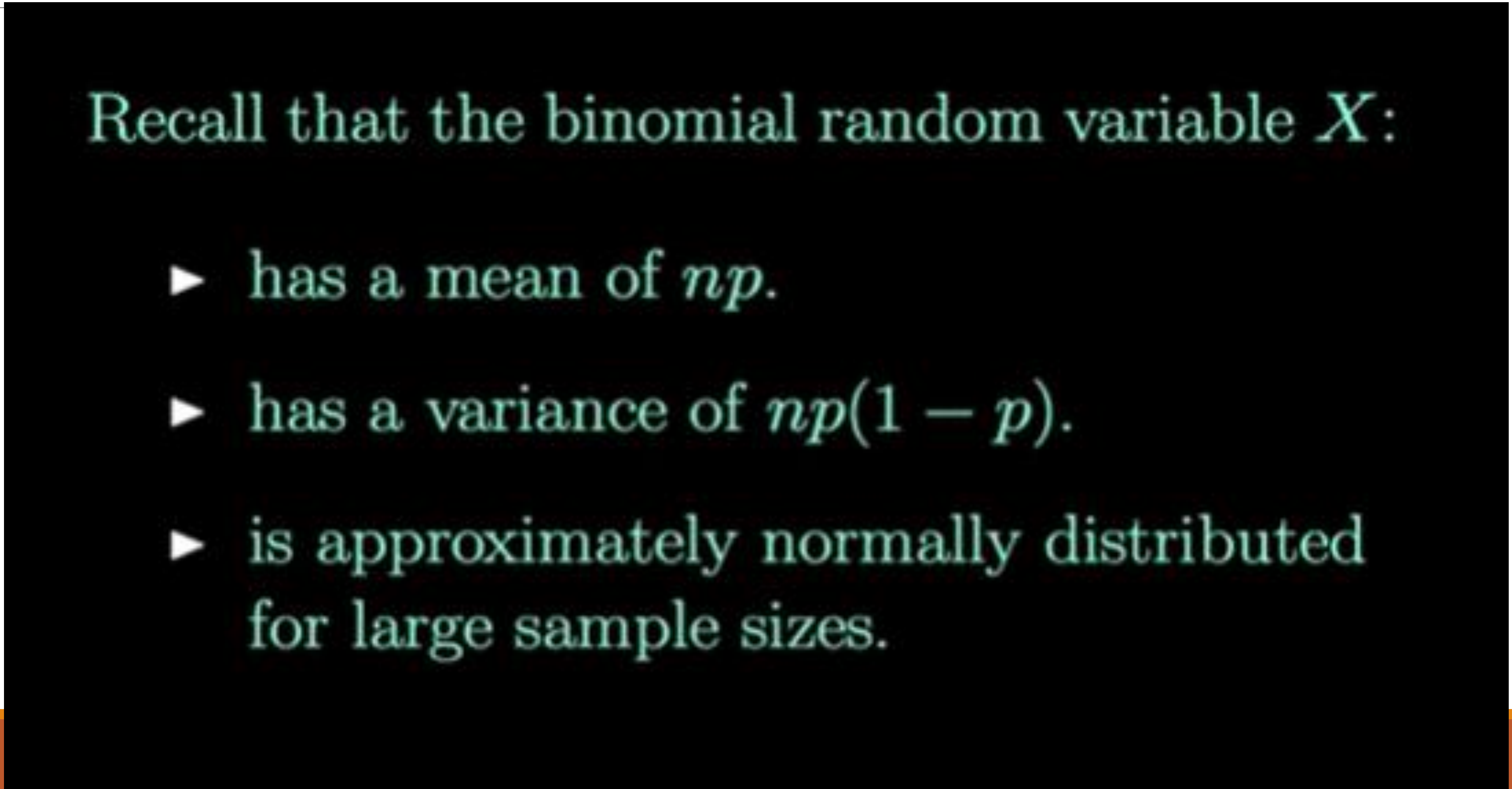
Number of individuals in the sample with the characteristic

Number of individuals in the sample

X is a binomial random variable with parameters n and p

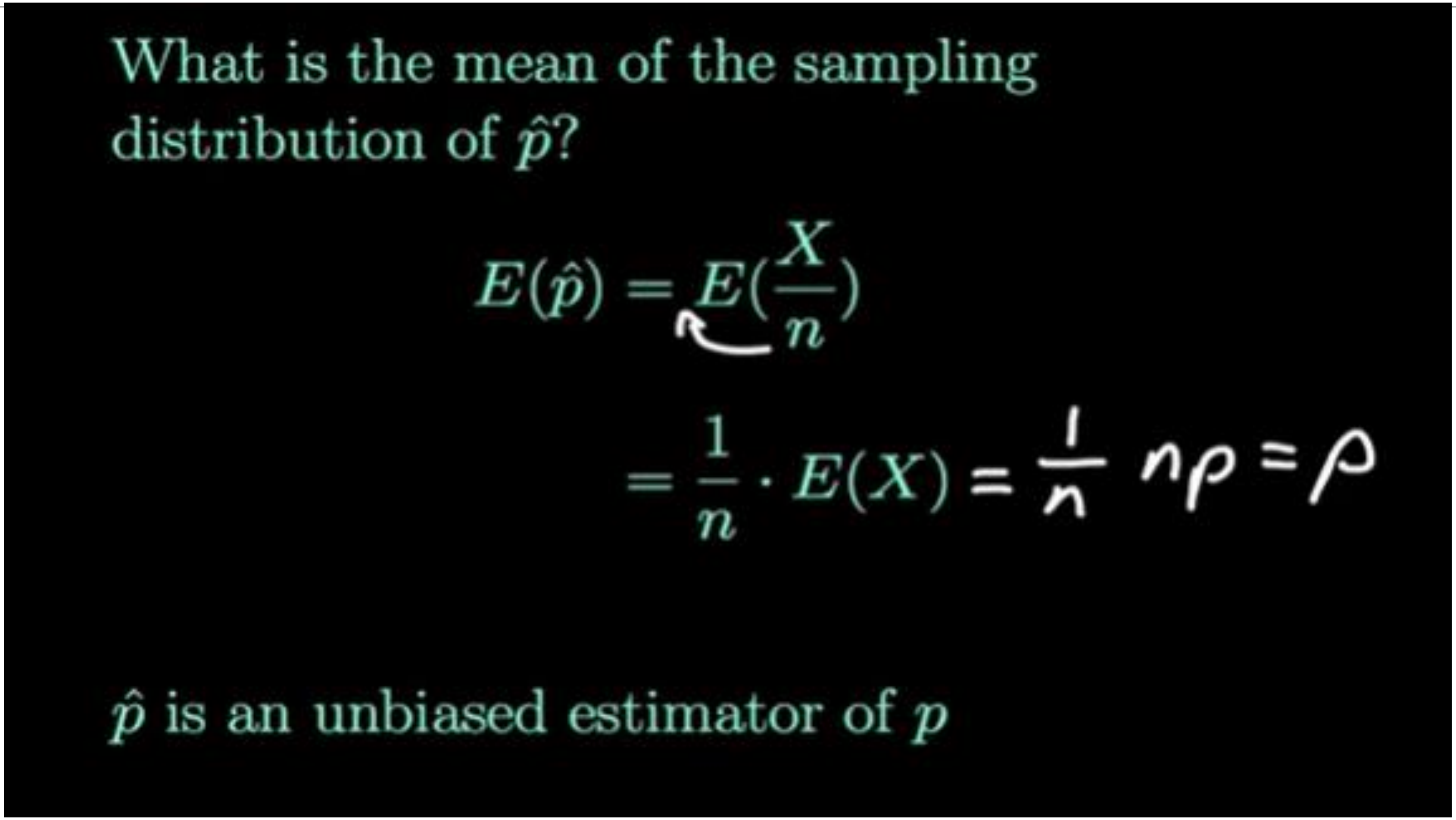
$\hat{p}$ takes on one of $n+1$ possible values:

$$0,\ \frac{1}{n},\ \frac{2}{n}, \ldots, 1$$



Recall that the binomial random variable X:

- has a mean of np.
- has a variance of $np(1-p)$.
- is approximately normally distributed for large sample sizes.

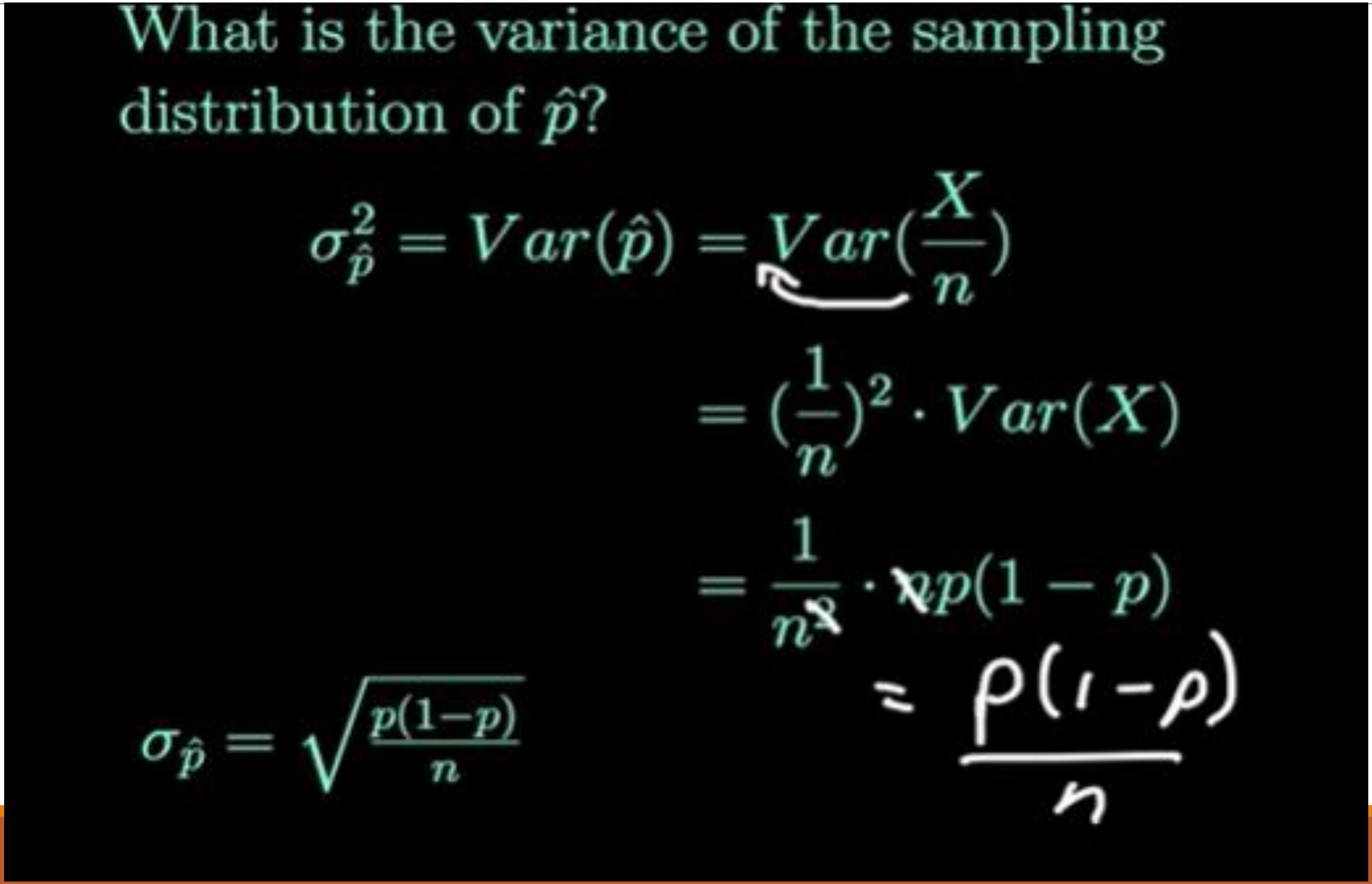


What is the mean of the sampling distribution of $\hat{p}$?

$$E(\hat{p}) = E(\frac{X}{n})$$

$$= \frac{1}{n} \cdot E(X) = \frac{1}{n} np = p$$

$\hat{p}$ is an unbiased estimator of p



What is the variance of the sampling distribution of $\hat{p}$?

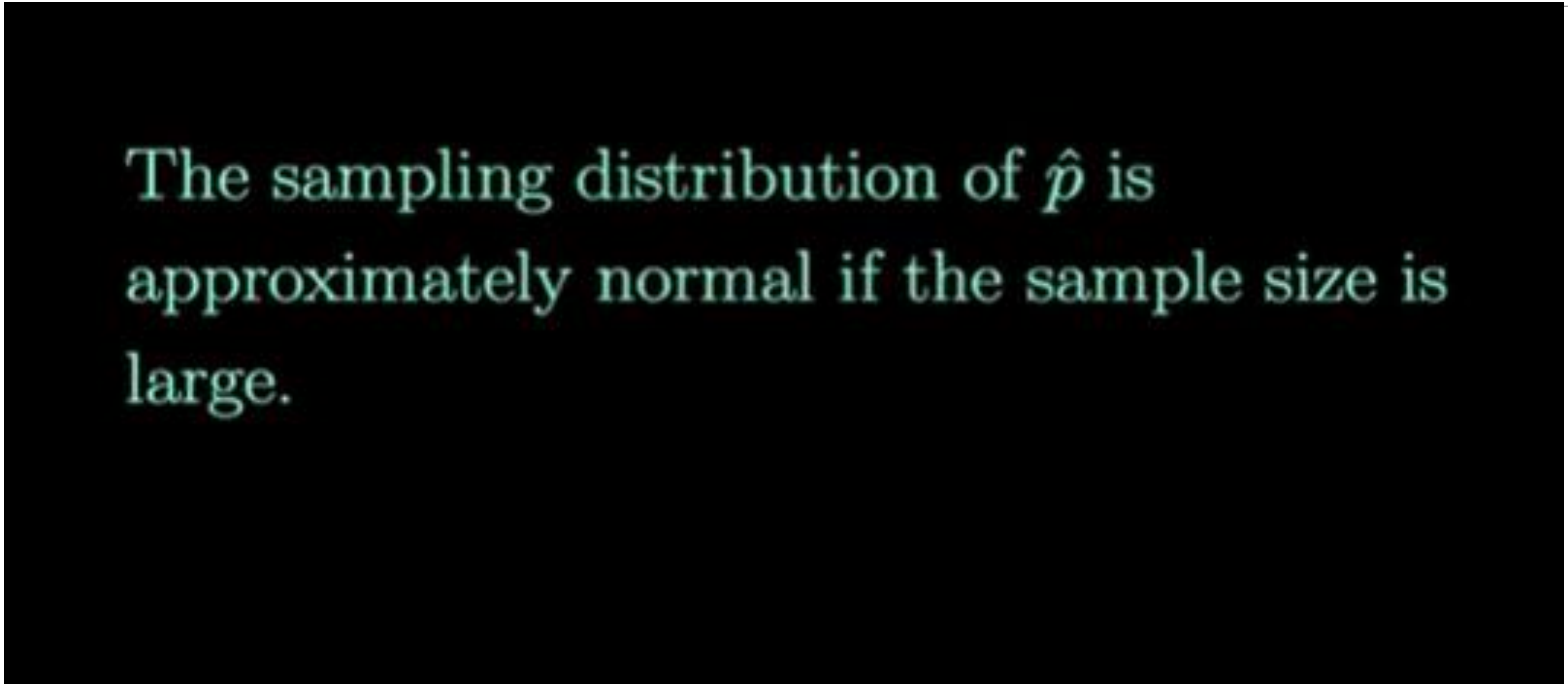
$$\sigma^2_{\hat{p}} = Var(\hat{p}) = Var(\frac{X}{n})$$

$$= (\frac{1}{n})^2 \cdot Var(X)$$

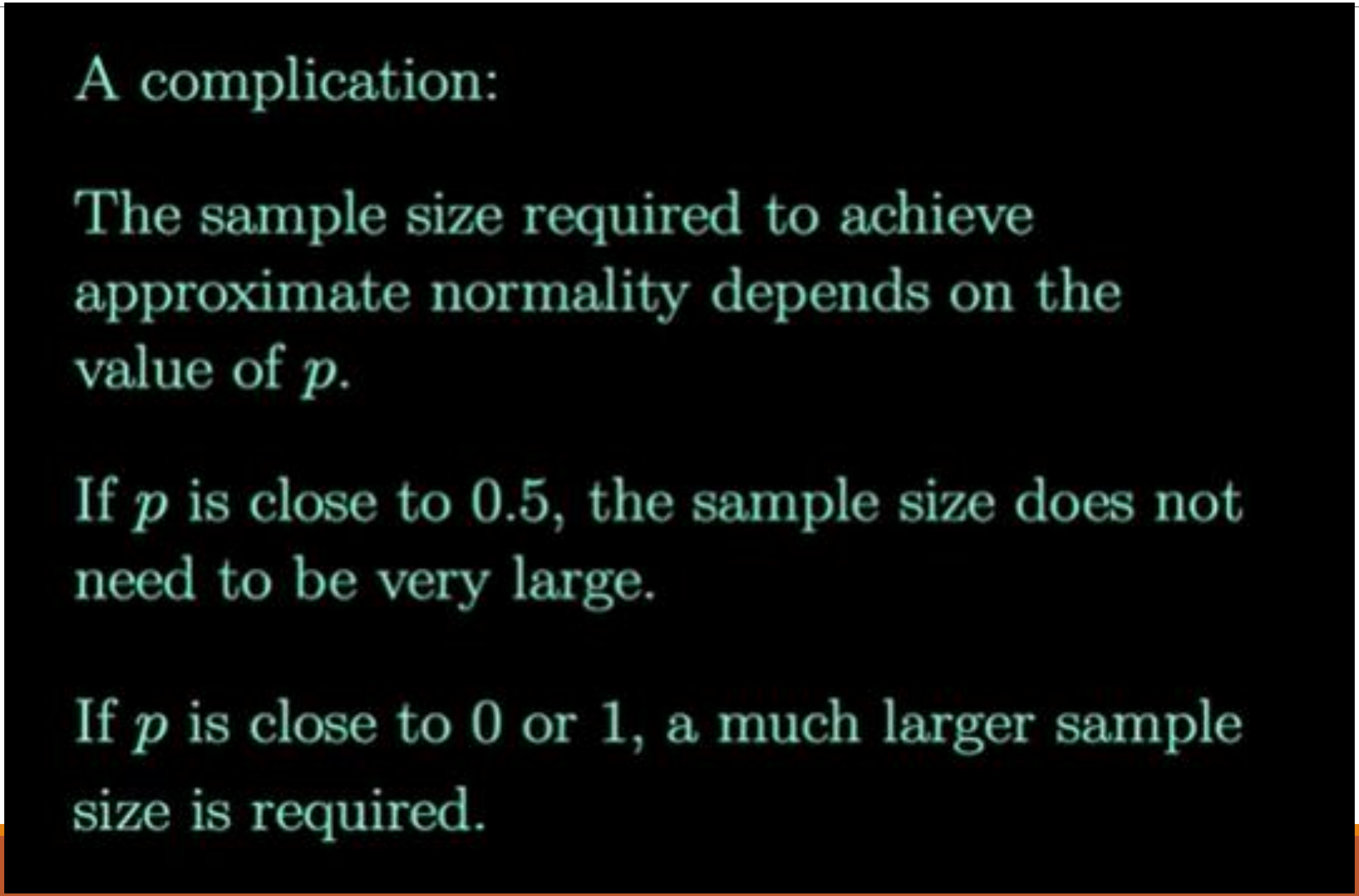
$$= \frac{1}{n^{\cancel{2}}} \cdot \cancel{n}p(1-p)$$

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

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



The sampling distribution of $\hat{p}$ is approximately normal if the sample size is large.

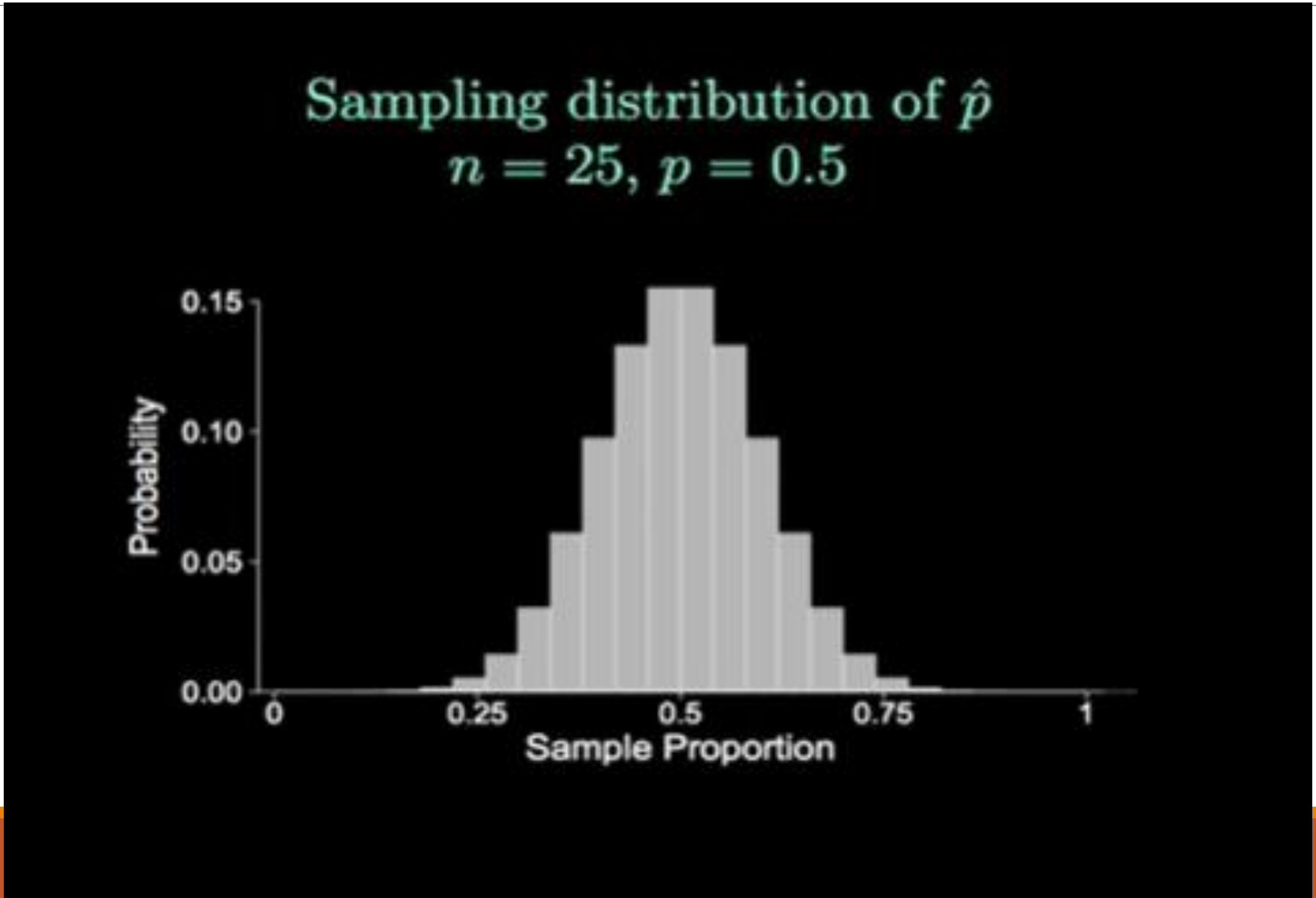


A complication:

The sample size required to achieve approximate normality depends on the value of p.

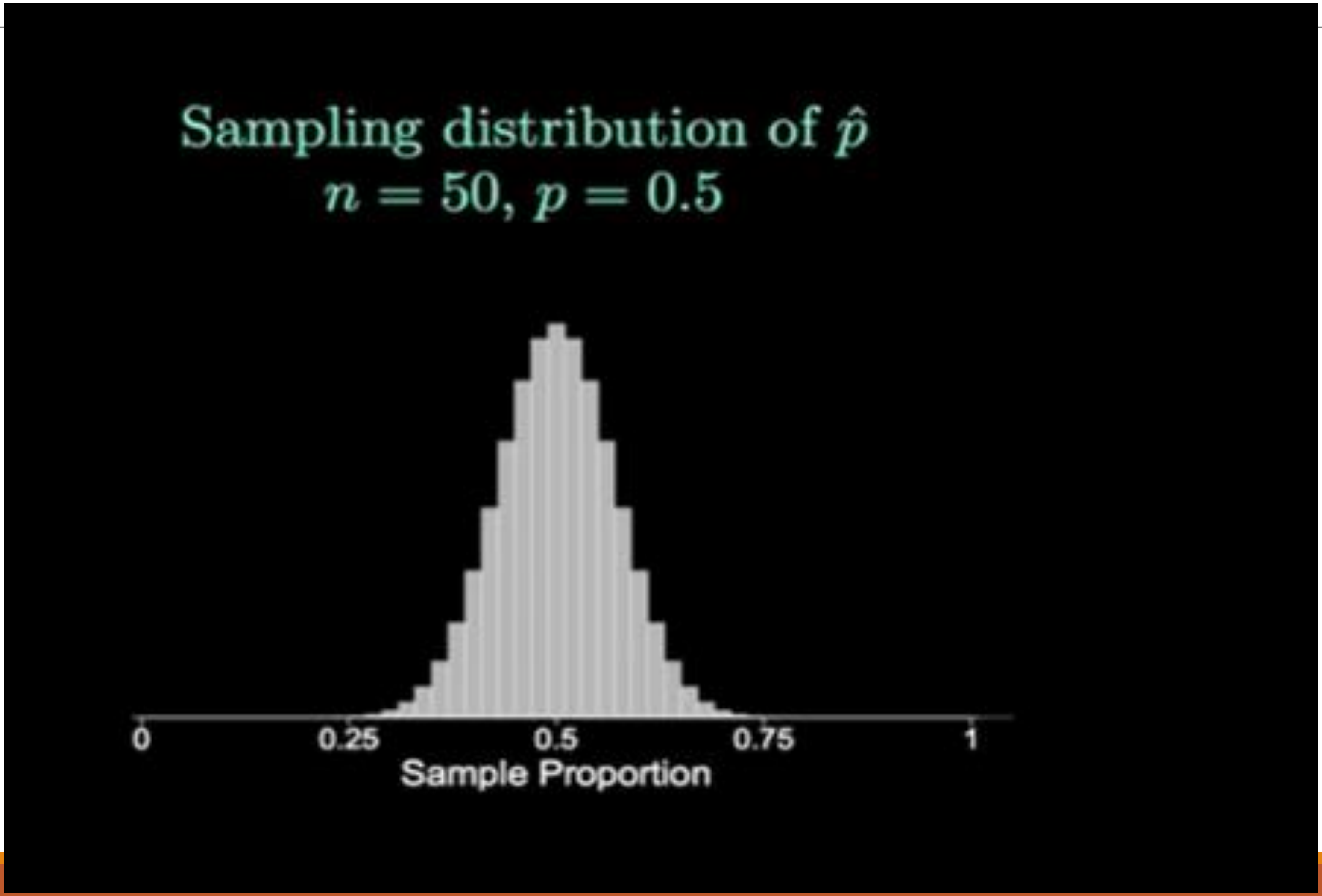
If p is close to 0.5, the sample size does not need to be very large.

If p is close to 0 or 1, a much larger sample size is required.



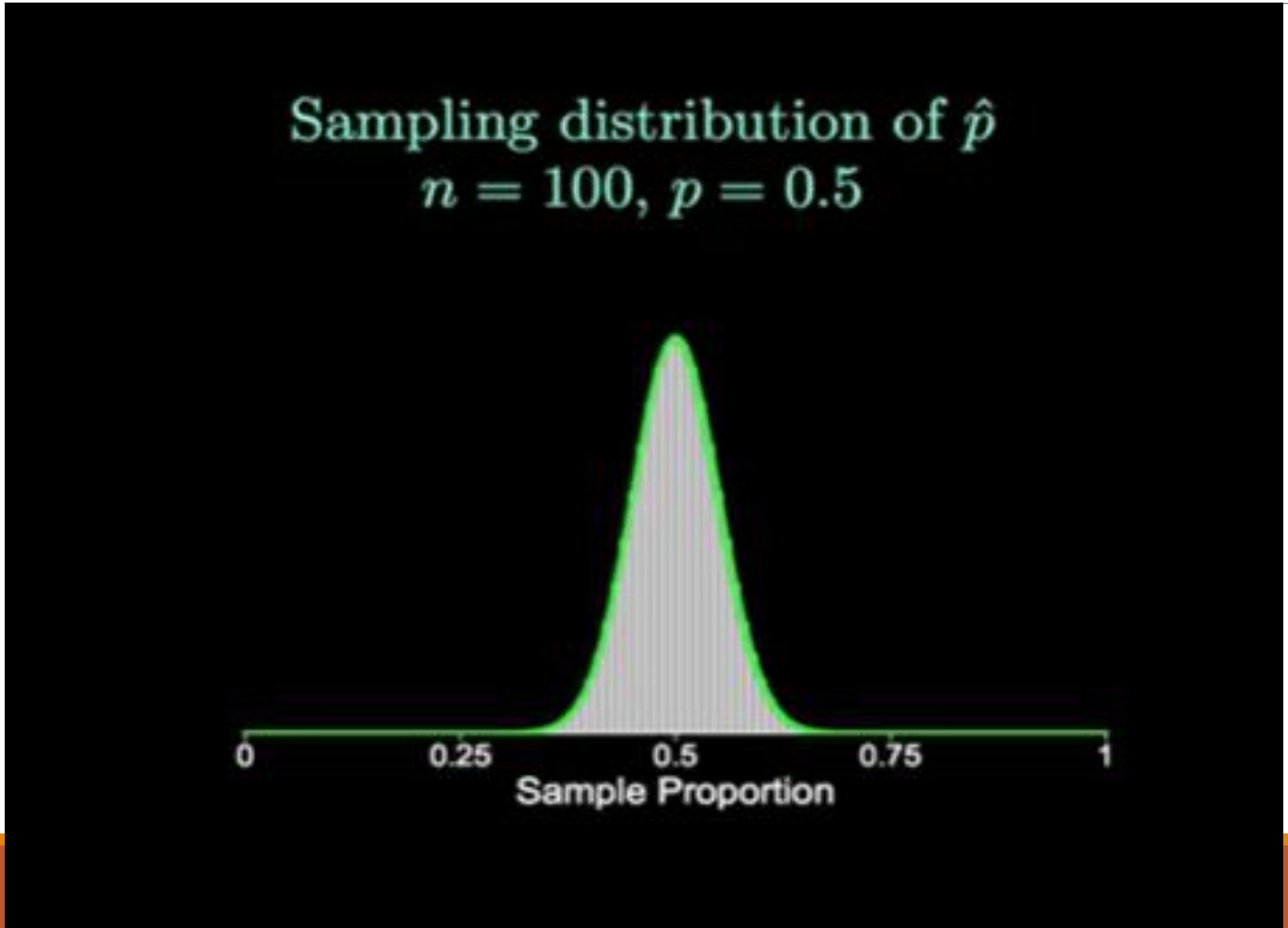
Sampling distribution of $\hat{p}$

$n = 25,\ p = 0.5$

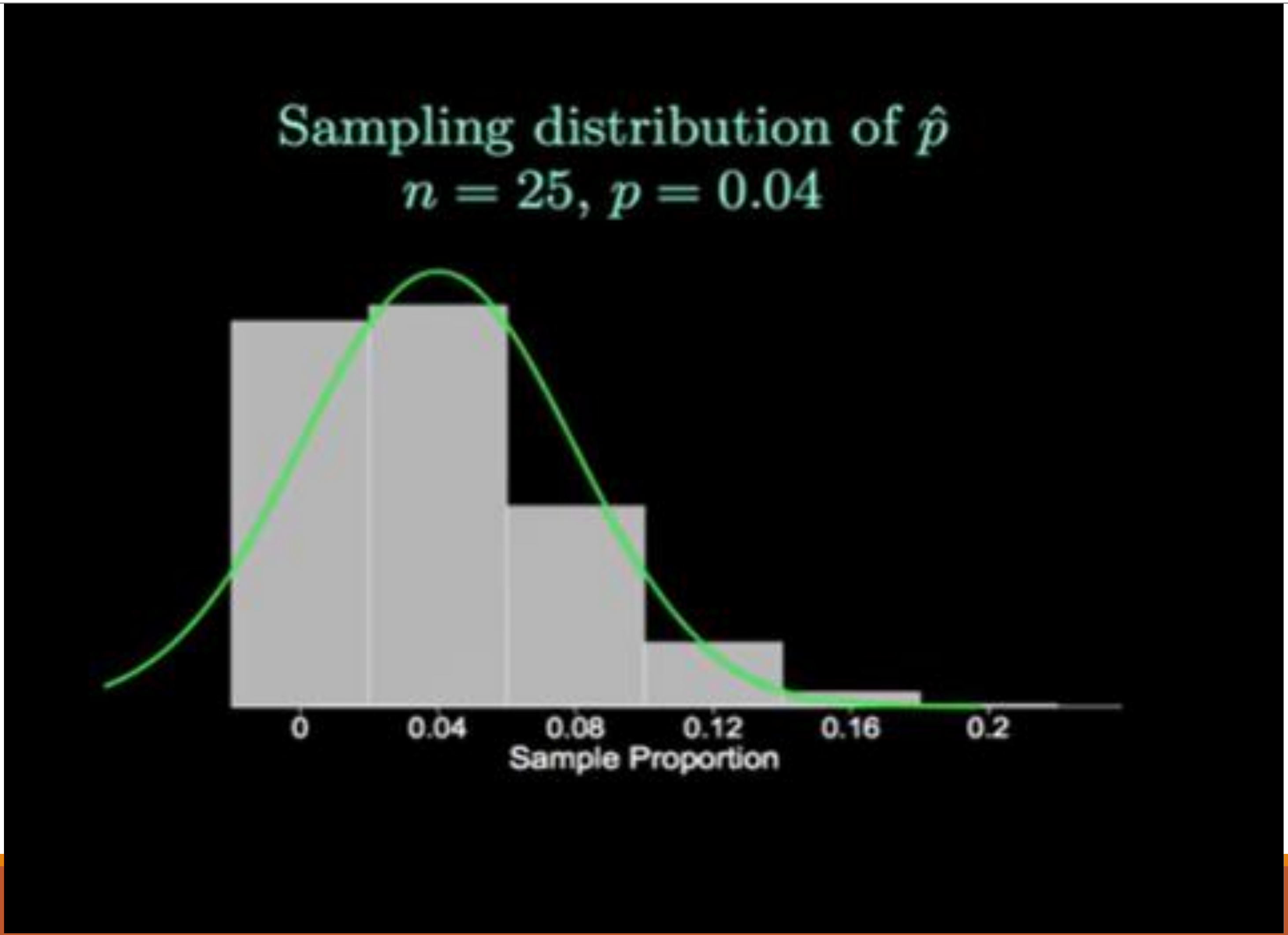


Sampling distribution of $\hat{p}$

$n = 50,\ p = 0.5$

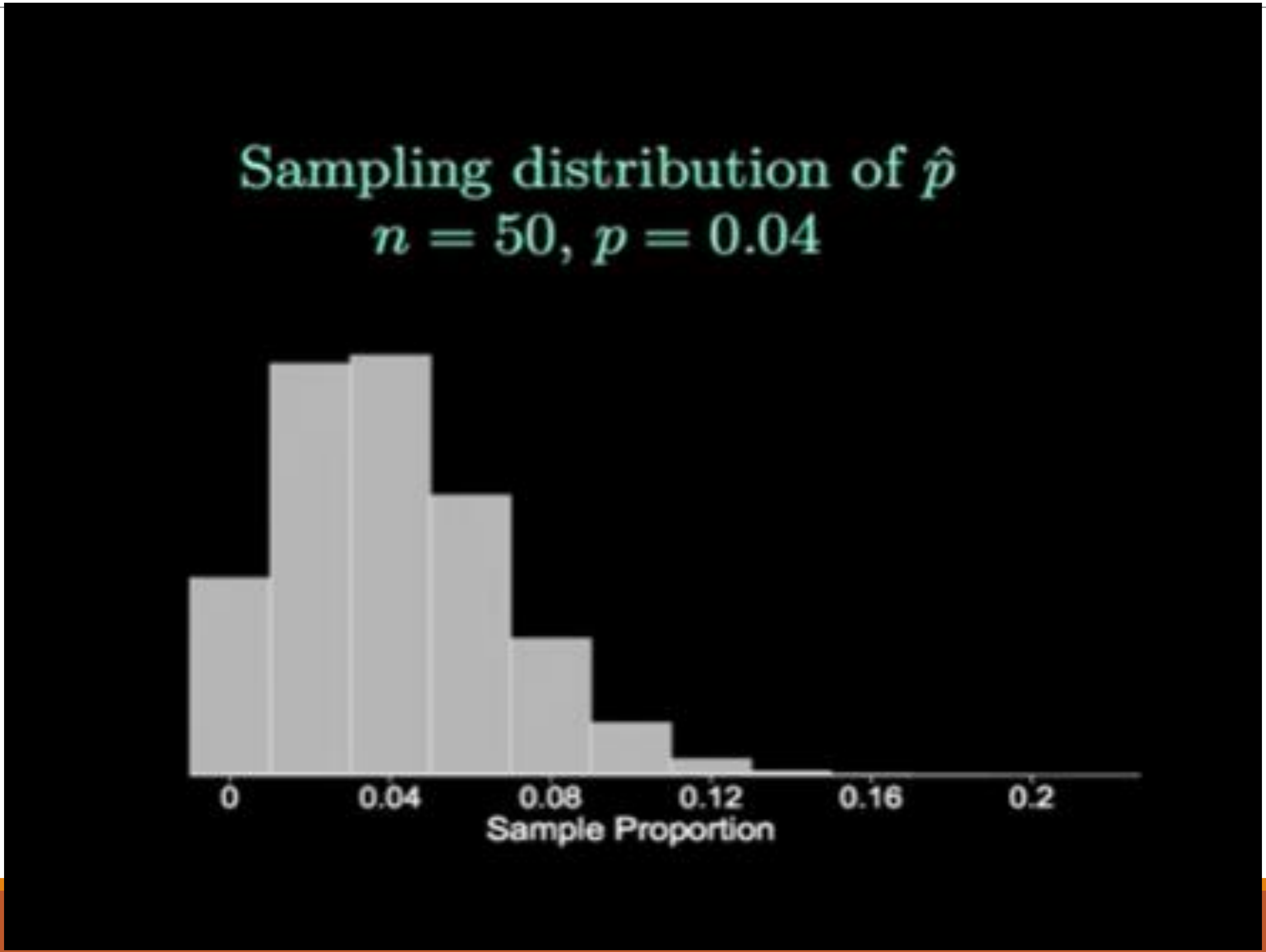


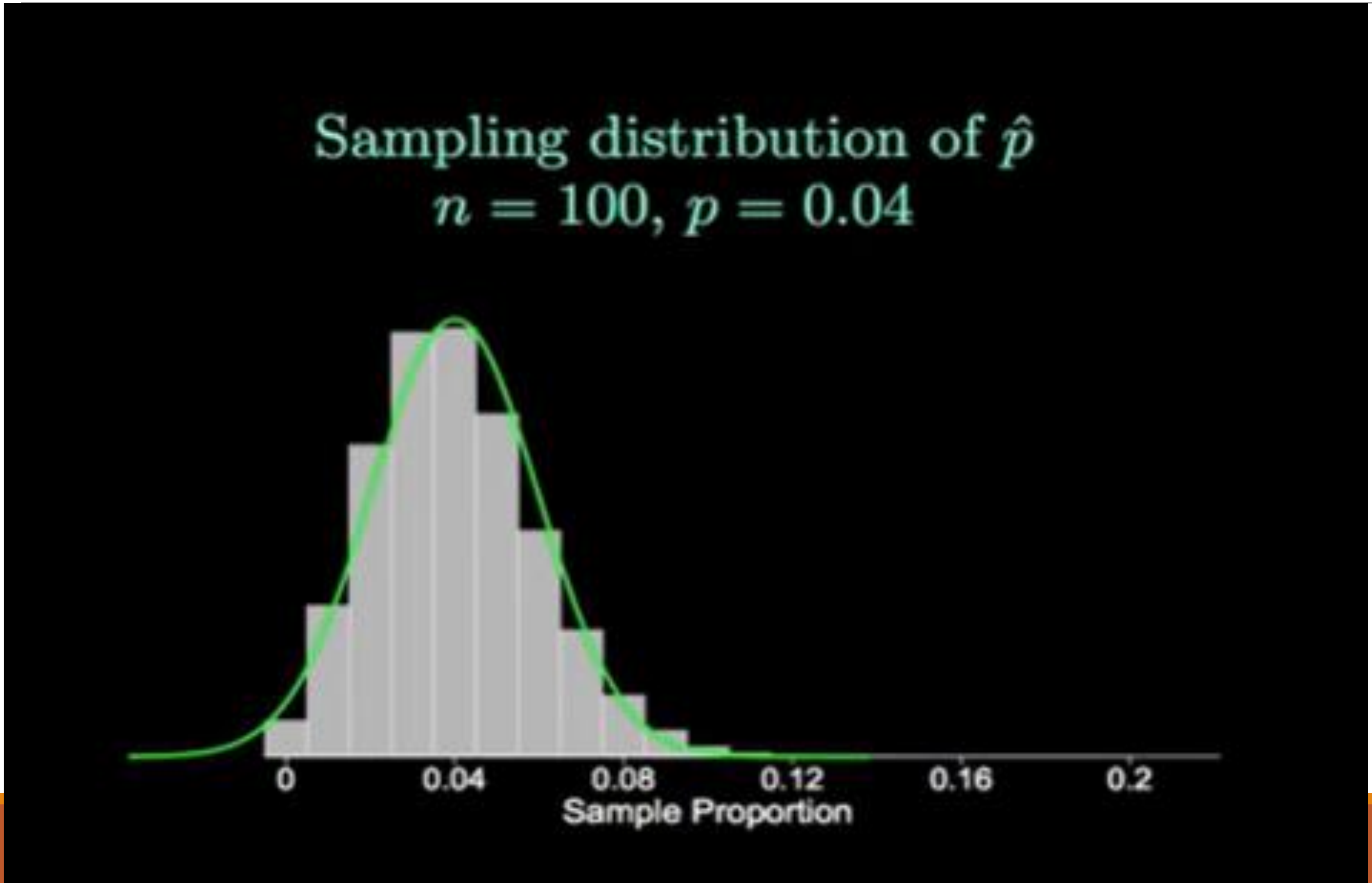
Sampling distribution of $\hat{p}$
$n = 100$, $p = 0.5$

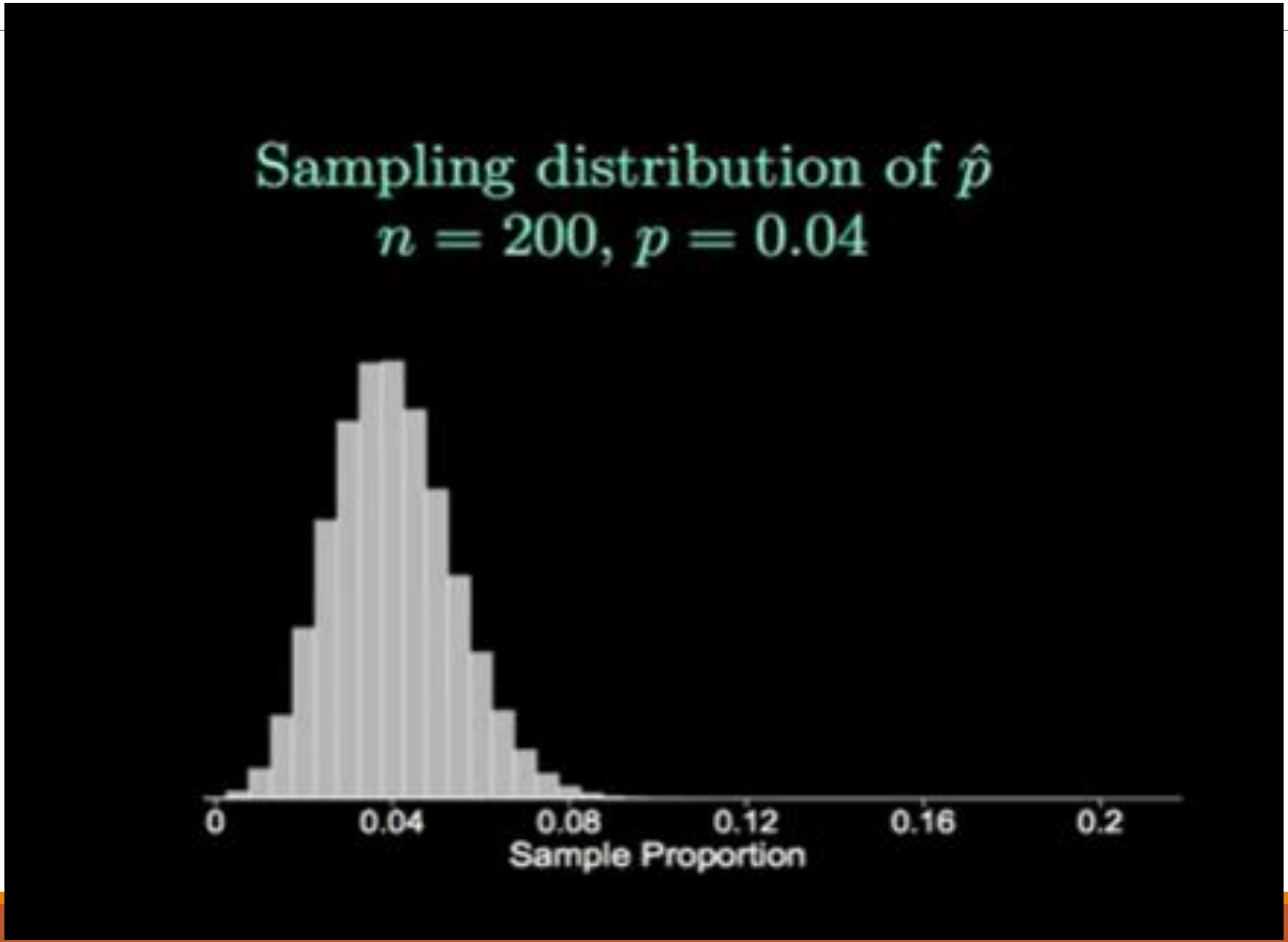


Sampling distribution of $\hat{p}$

$n = 25,\ p = 0.04$

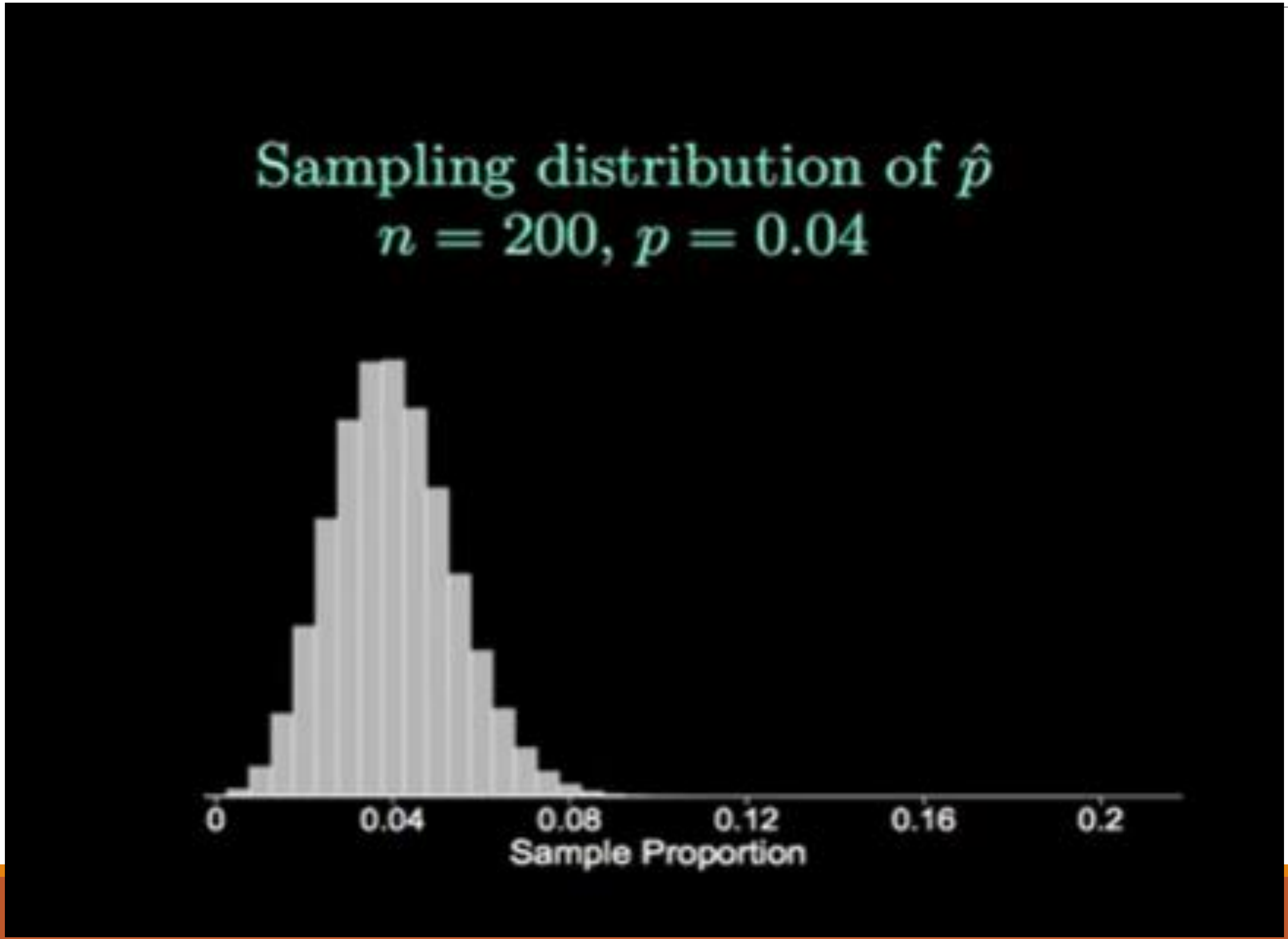

Sampling distribution of $\hat{p}$
$n = 50,\ p = 0.04$
0
0.04
0.08
0.12
0.16
0.2
Sample Proportion

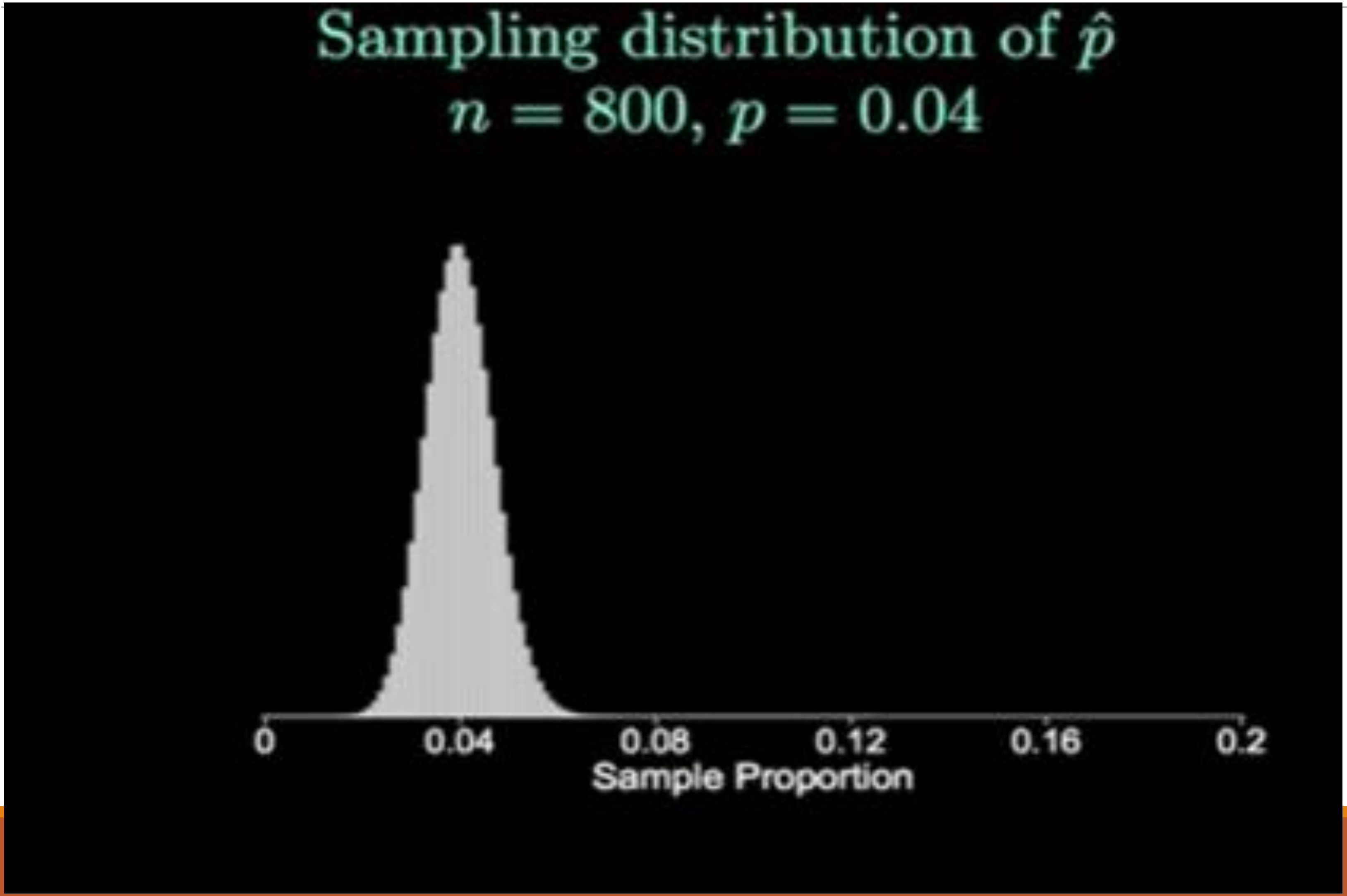

Sampling distribution of $\hat{p}$
$n = 100$, $p = 0.04$
0
0.04
0.08
0.12
0.16
0.2
Sample Proportion



Sampling distribution of $\hat{p}$

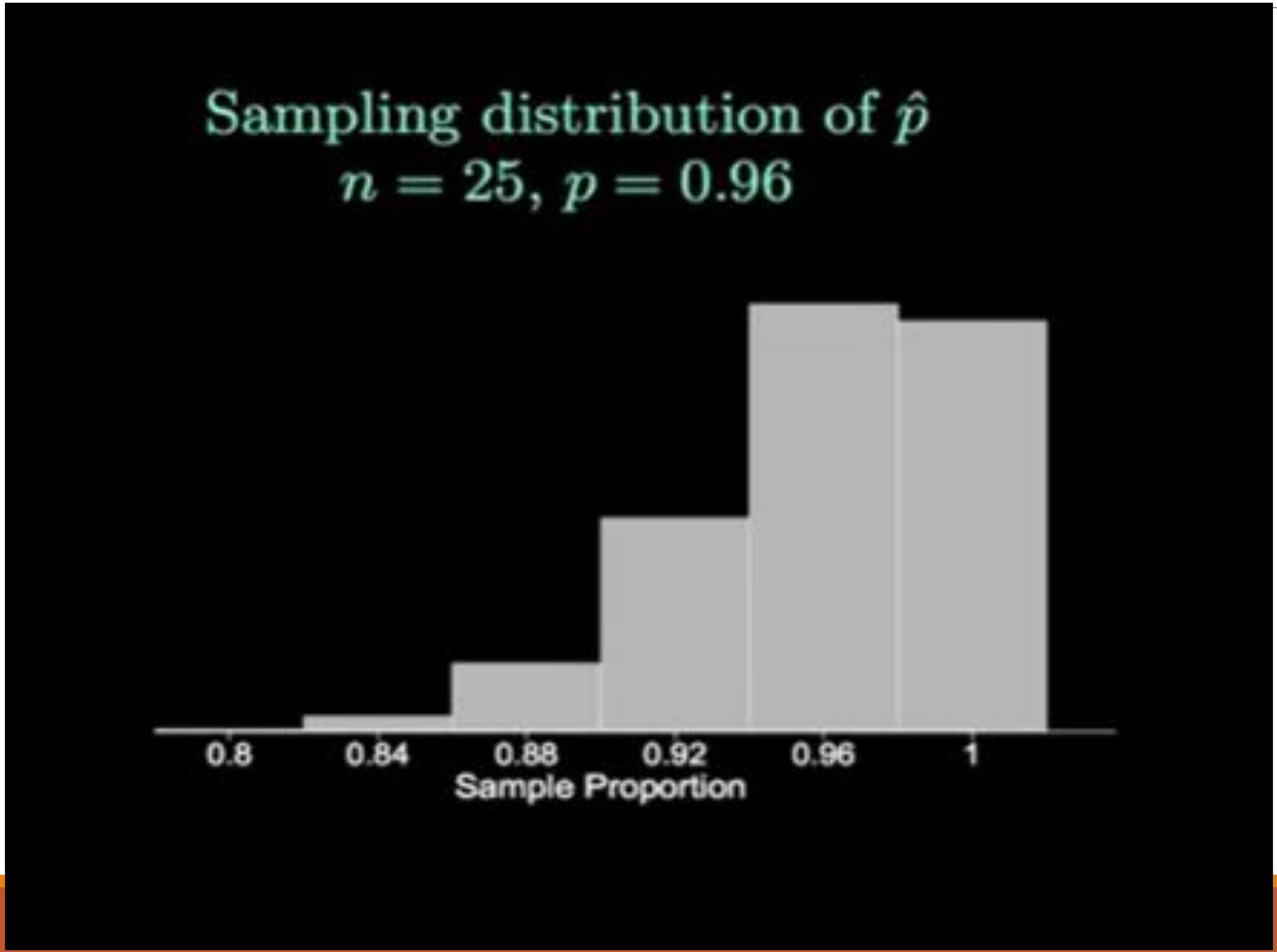
$n = 200$, $p = 0.04$

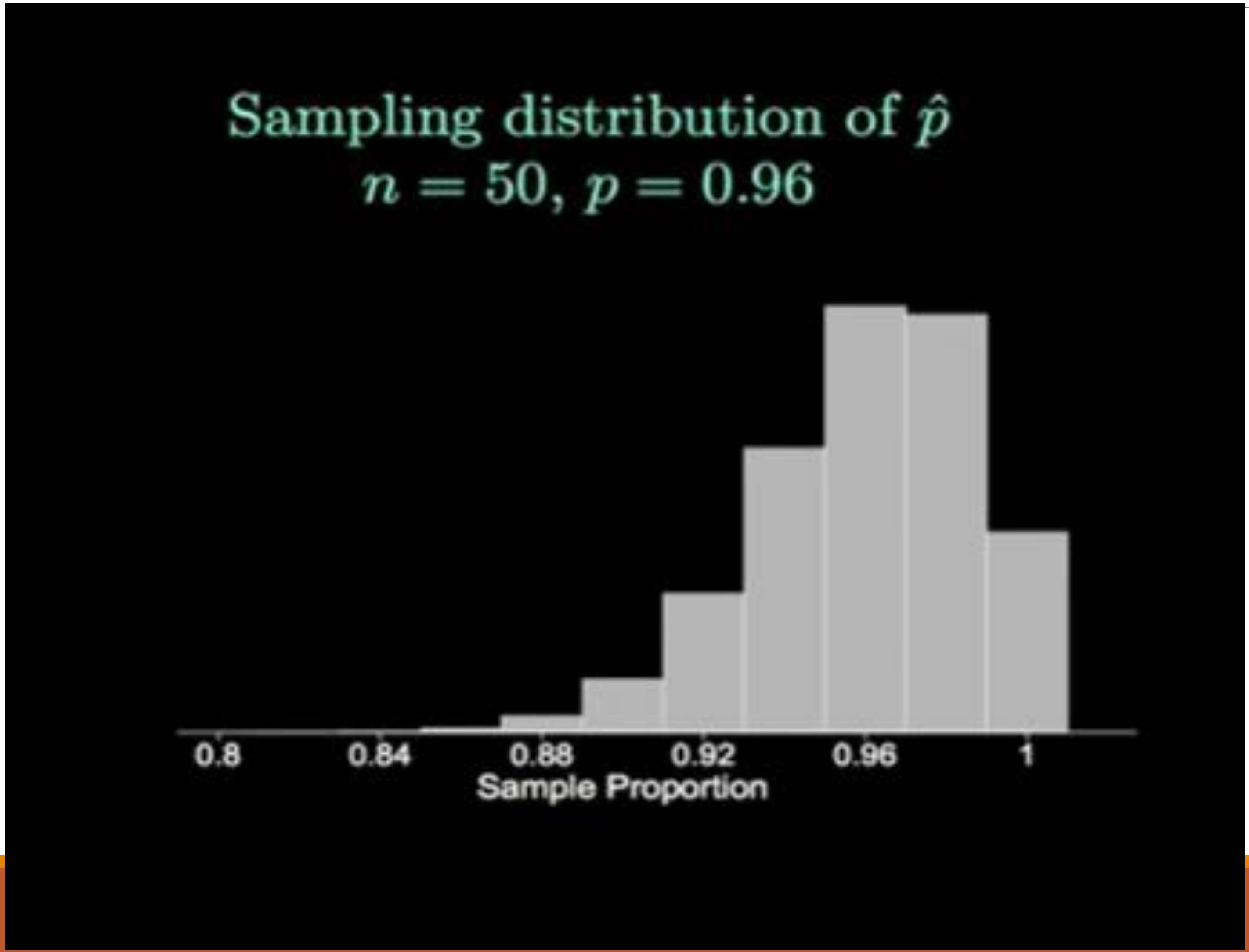

Sampling distribution of $\hat{p}$
$n = 200,\ p = 0.04$
0
0.04
0.08
0.12
0.16
0.2
Sample Proportion



Sampling distribution of $\hat{p}$

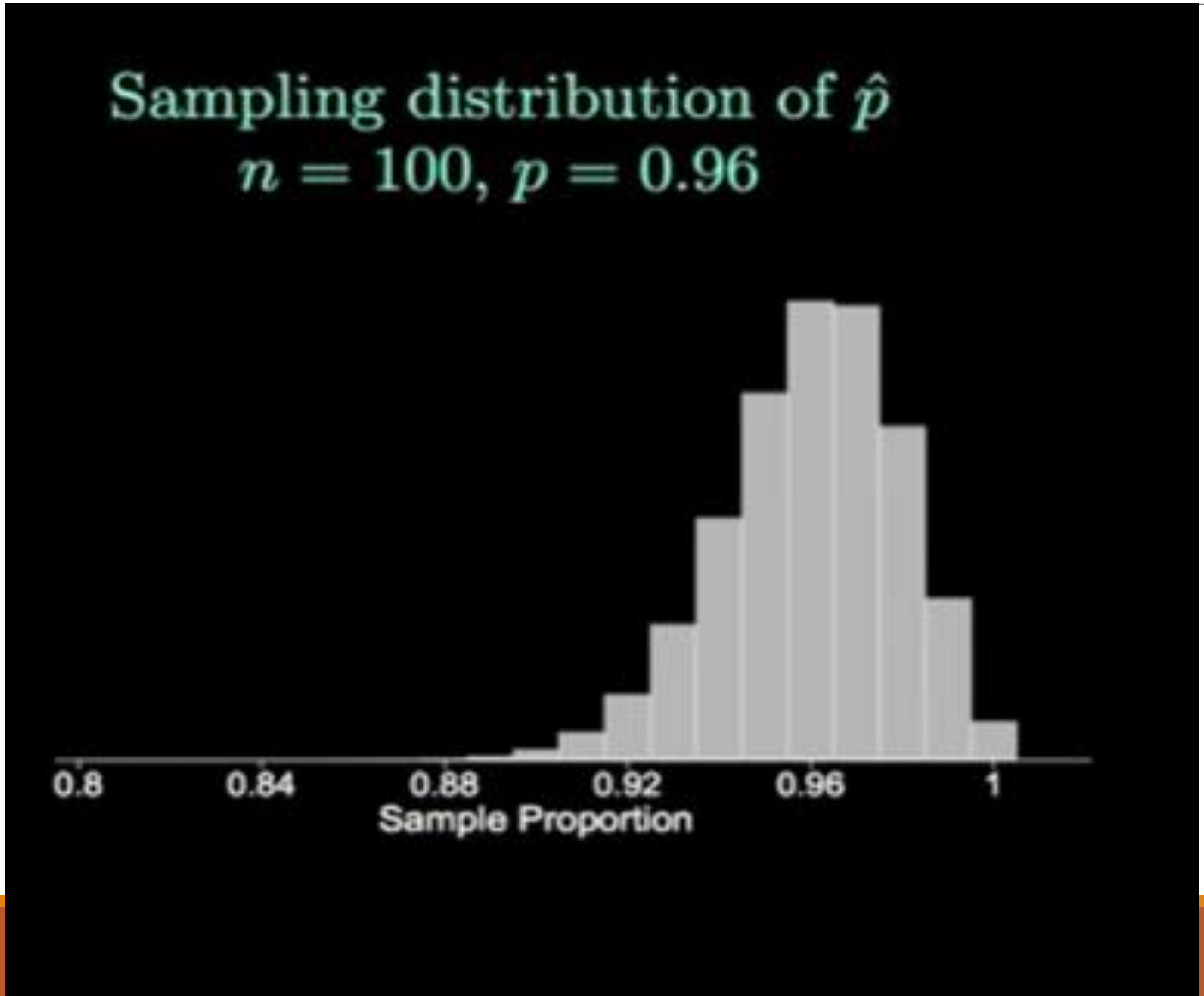
$n = 800, p = 0.04$

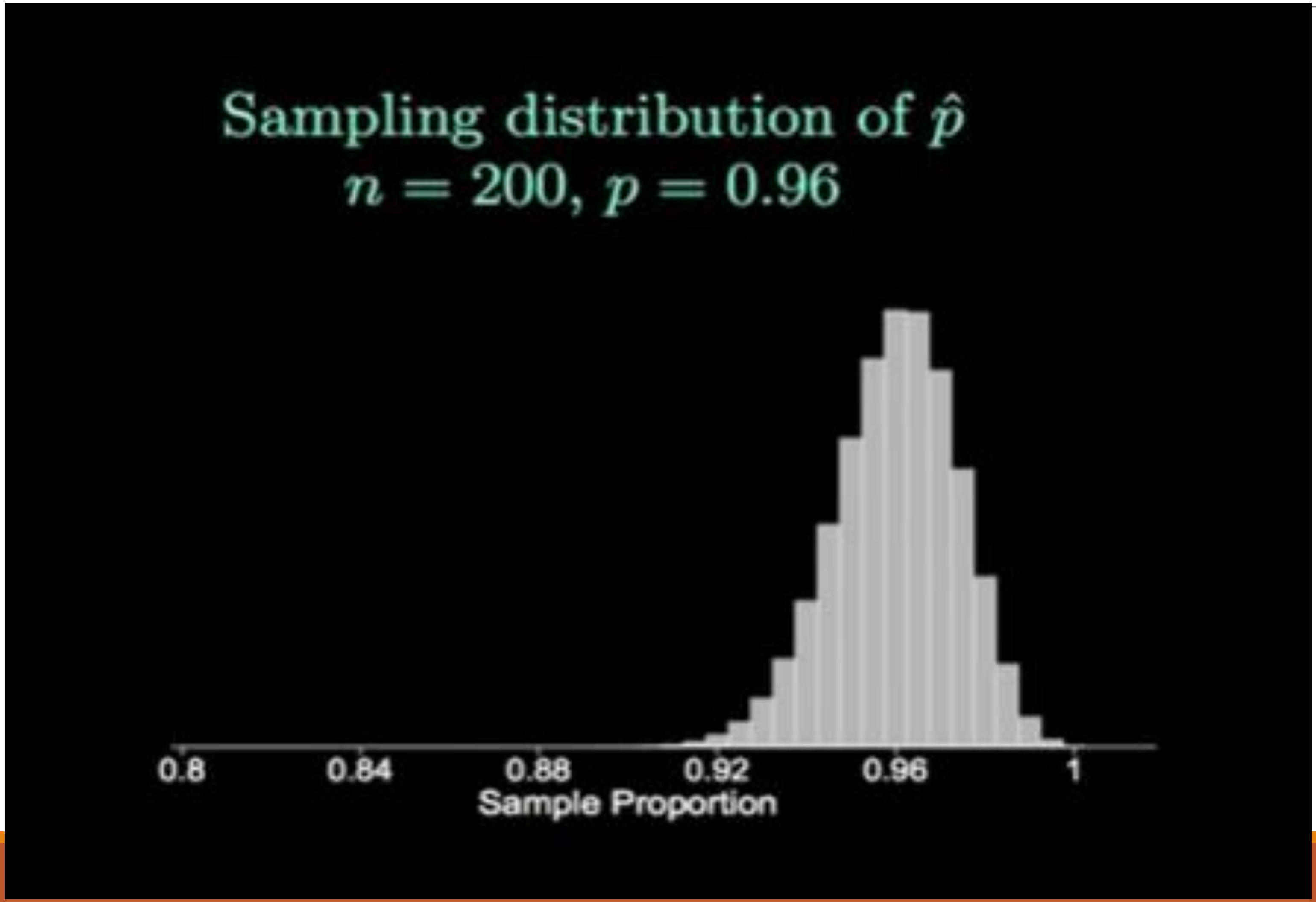

Sampling distribution of $\hat{p}$
$n = 25,\ p = 0.96$
0.8
0.84
0.88
0.92
0.96
1
Sample Proportion



Sampling distribution of $\hat{p}$

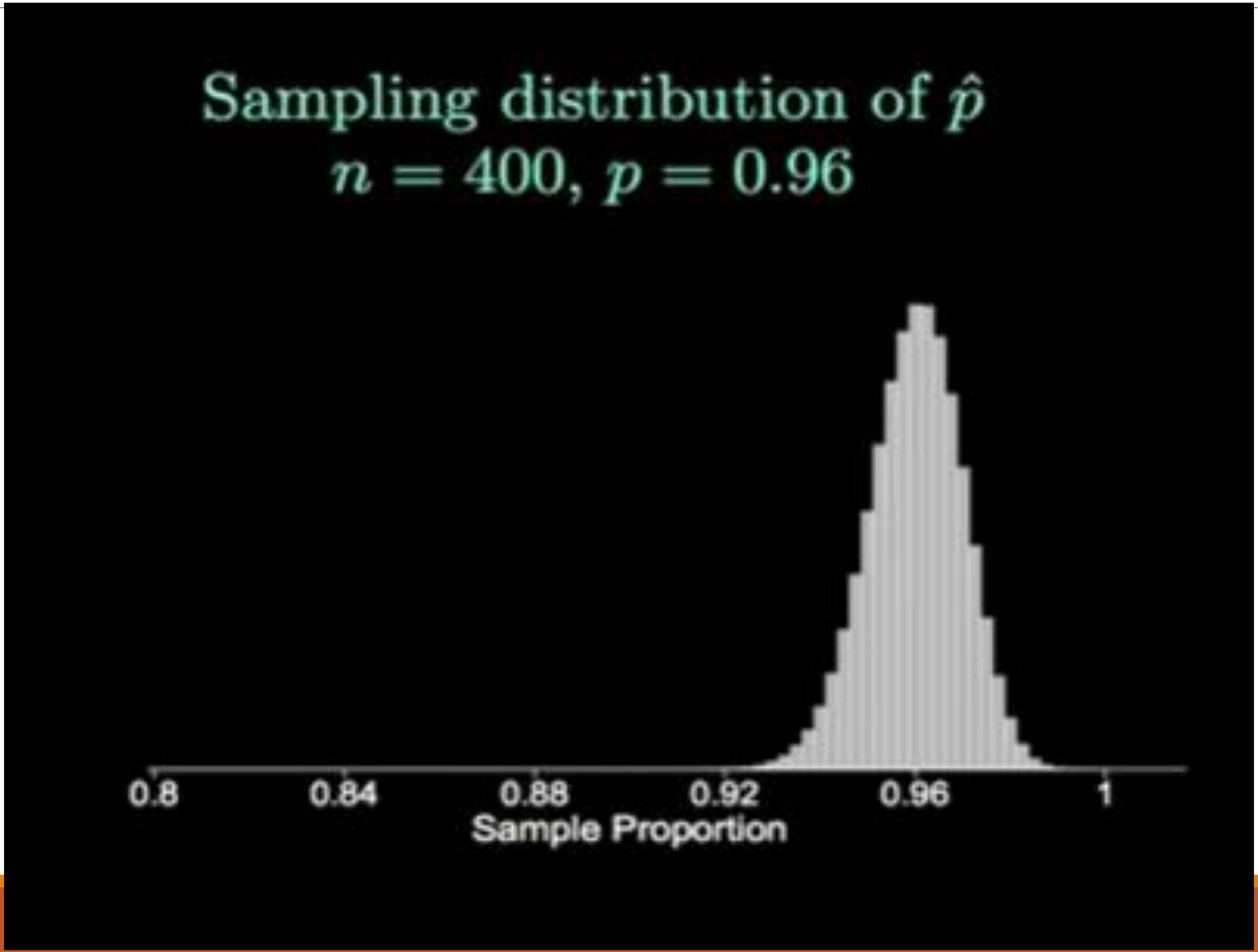
$n = 50,\ p = 0.96$

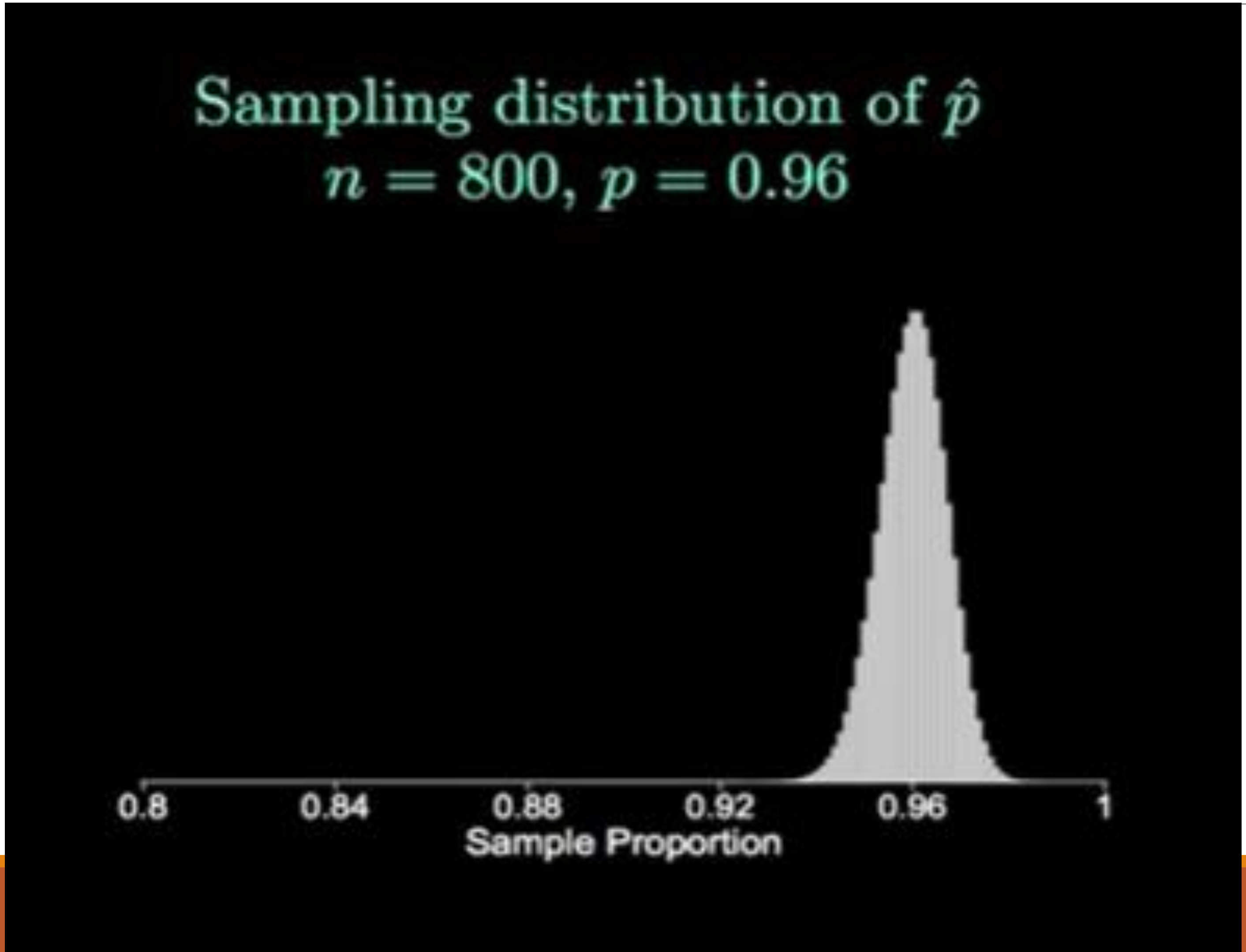

Sampling distribution of $\hat{p}$
$n = 100, p = 0.96$
0.8
0.84
0.88
0.92
0.96
1
Sample Proportion

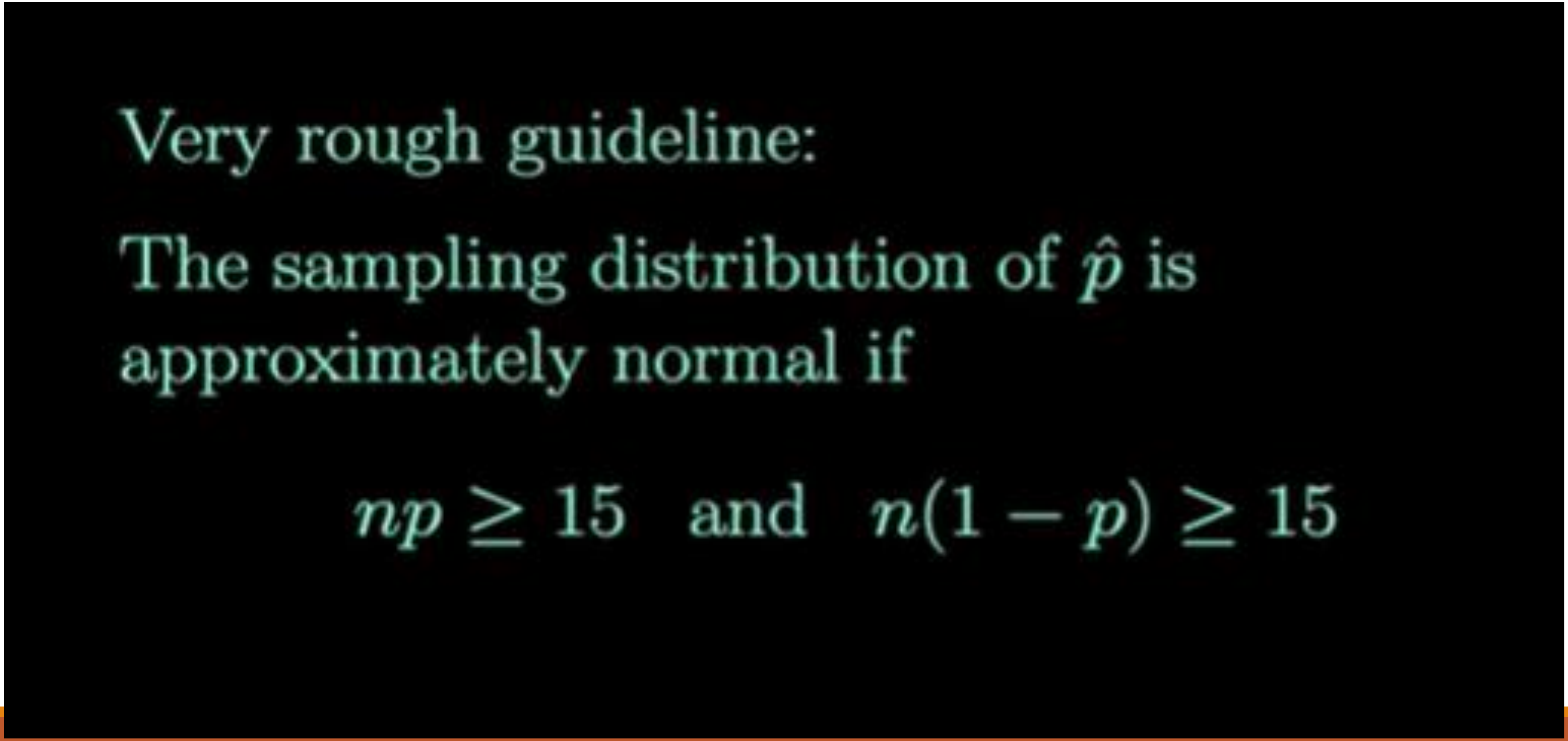


Sampling distribution of $\hat{p}$

$n = 200,\ p = 0.96$


Sampling distribution of $\hat{p}$
$n = 400,\ p = 0.96$
0.8
0.84
0.88
0.92
0.96
1
Sample Proportion

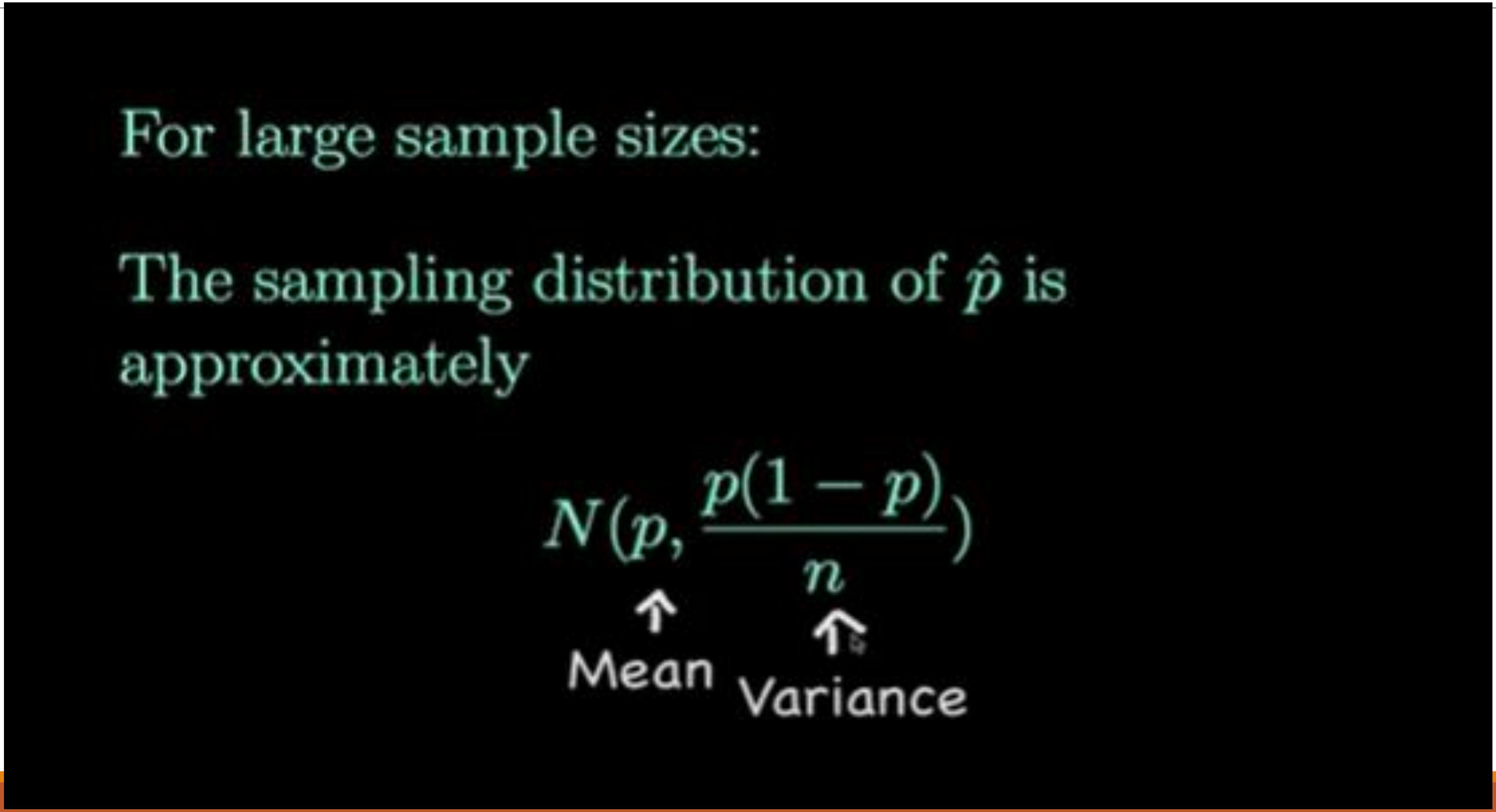

Sampling distribution of $\hat{p}$
$n = 800, p = 0.96$
0.8
0.84
0.88
0.92
0.96
1
Sample Proportion



Very rough guideline:

The sampling distribution of $\hat{p}$ is approximately normal if

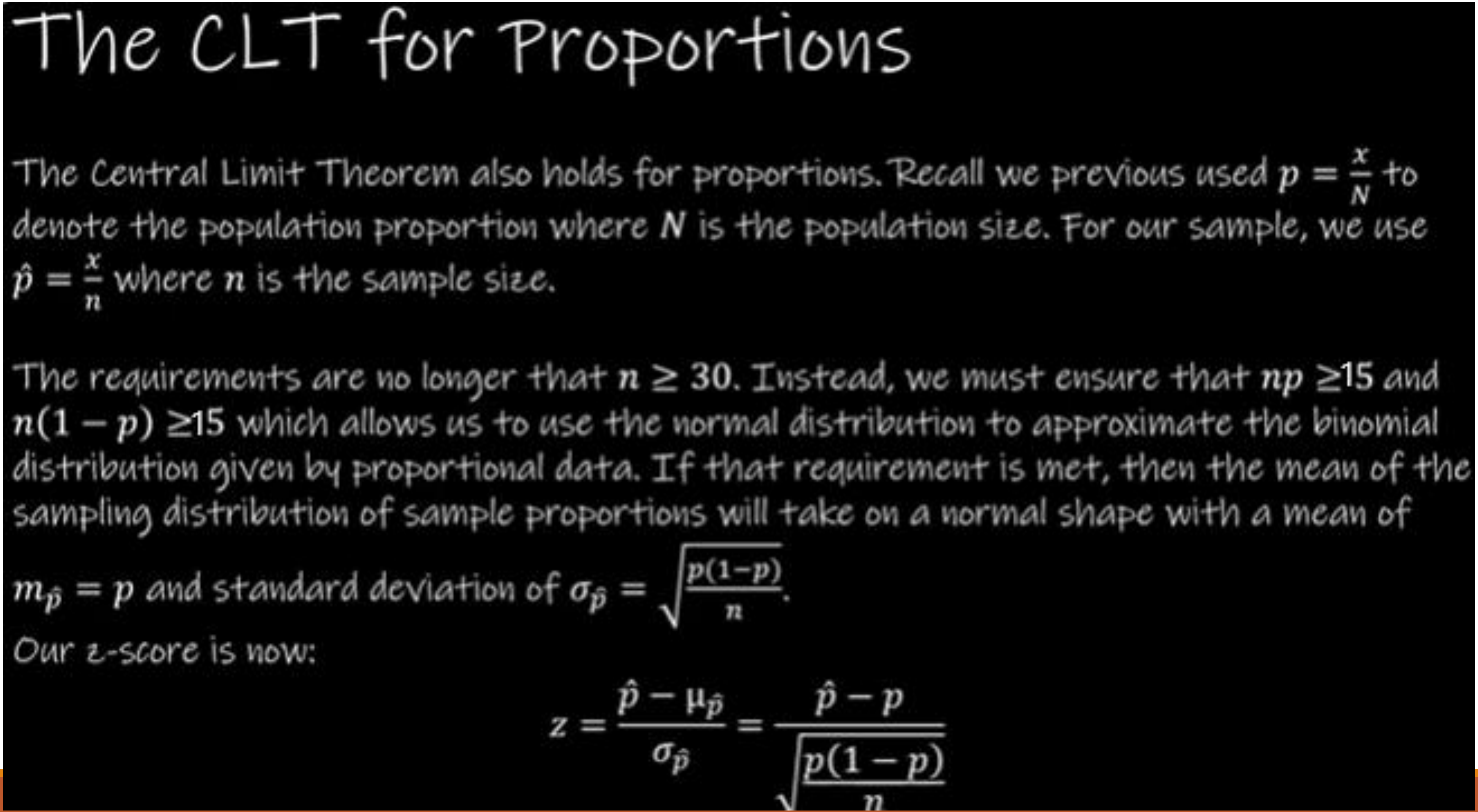
$$np \geq 15 \quad \text{and} \quad n(1-p) \geq 15$$



For large sample sizes:

The sampling distribution of $\hat{p}$ is approximately

$$N(\underset{\substack{\uparrow \\ \text{Mean}}}{p}, \underset{\substack{\uparrow \\ \text{Variance}}}{\frac{p(1-p)}{n}})$$



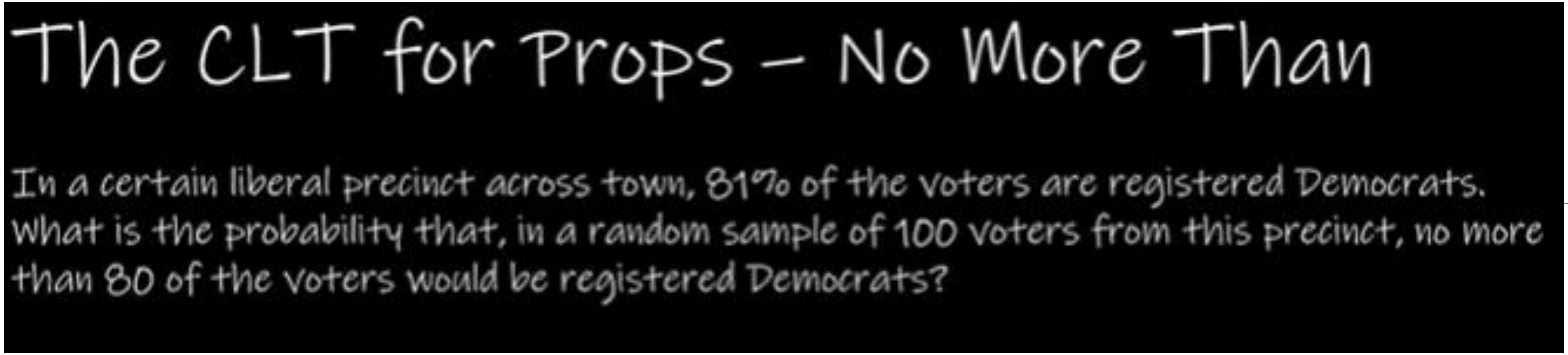
The CLT for Proportions

The Central Limit Theorem also holds for proportions. Recall we previous used $p = \frac{x}{N}$ to denote the population proportion where N is the population size. For our sample, we use $\hat{p} = \frac{x}{n}$ where n is the sample size.

The requirements are no longer that $n \geq 30$. Instead, we must ensure that $np \geq 15$ and $n(1-p) \geq 15$ which allows us to use the normal distribution to approximate the binomial distribution given by proportional data. If that requirement is met, then the mean of the sampling distribution of sample proportions will take on a normal shape with a mean of $m_{\hat{p}} = p$ and standard deviation of $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$.

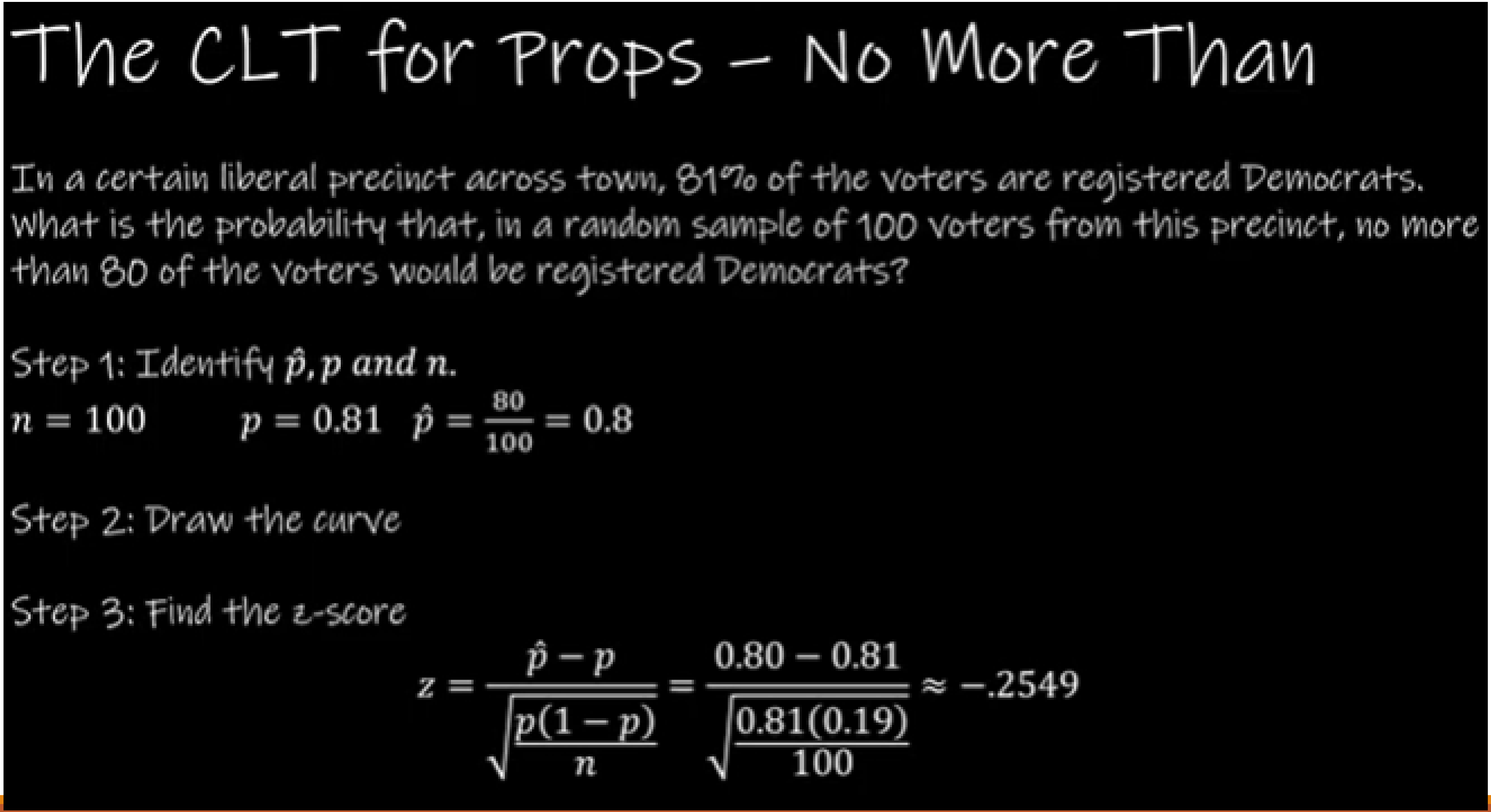
Our z-score is now:

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



The CLT for Props – No More Than

In a certain liberal precinct across town, 81% of the voters are registered Democrats. What is the probability that, in a random sample of 100 voters from this precinct, no more than 80 of the voters would be registered Democrats?



The CLT for Props – No More Than

In a certain liberal precinct across town, 81% of the voters are registered Democrats. What is the probability that, in a random sample of 100 voters from this precinct, no more than 80 of the voters would be registered Democrats?

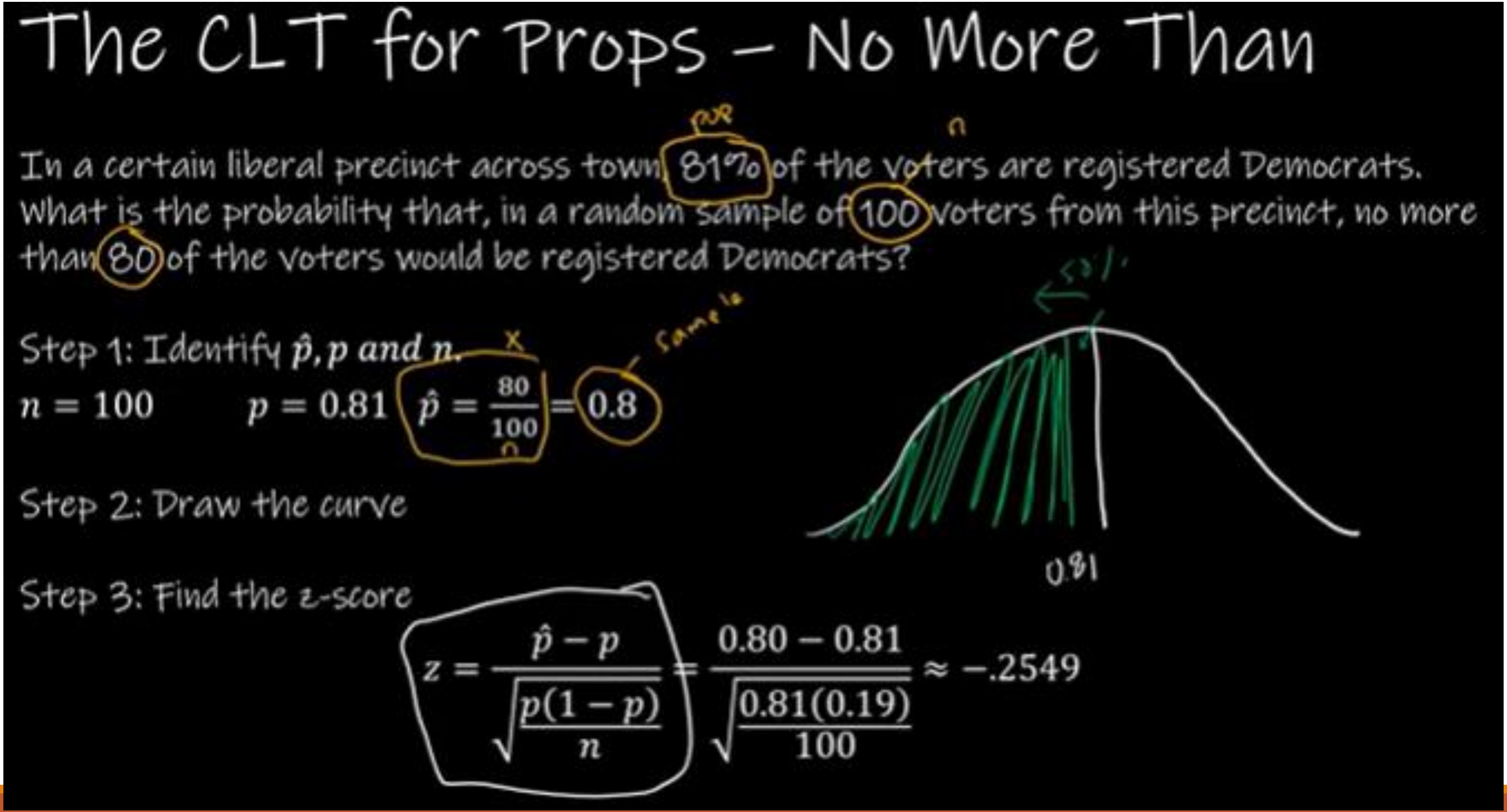
Step 1: Identify $\hat{p}, p$ *and* n.

$n = 100 \qquad p = 0.81 \quad \hat{p} = \frac{80}{100} = 0.8$

Step 2: Draw the curve

Step 3: Find the z-score

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.80 - 0.81}{\sqrt{\frac{0.81(0.19)}{100}}} \approx -.2549$$



The CLT for Props – No More Than

In a certain liberal precinct across town, 81% of the voters are registered Democrats. What is the probability that, in a random sample of 100 voters from this precinct, no more than 80 of the voters would be registered Democrats?

Step 1: Identify $\hat{p}, p$ and n.

$n = 100$ $p = 0.81$ $\boxed{\hat{p} = \frac{\overset{x}{80}}{\underset{n}{100}}} = \boxed{0.8}$

Step 2: Draw the curve

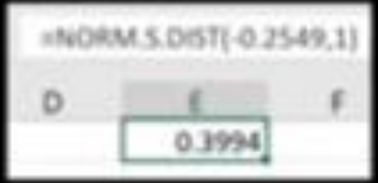
Step 3: Find the z-score

$$\boxed{z = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}} = \frac{0.80-0.81}{\sqrt{\frac{0.81(0.19)}{100}}} \approx -.2549$$

The CLT for Props – No More Than

In a certain liberal precinct across town, 81% of the voters are registered Democrats. What is the probability that, in a random sample of 100 voters from this precinct, no more than 80 of the voters would be registered Democrats?

Step 4: Calculate the probability that $\hat{p} \leq 0.8$



```
=NORM.S.DIST(-0.2549,1)
```

0.3994

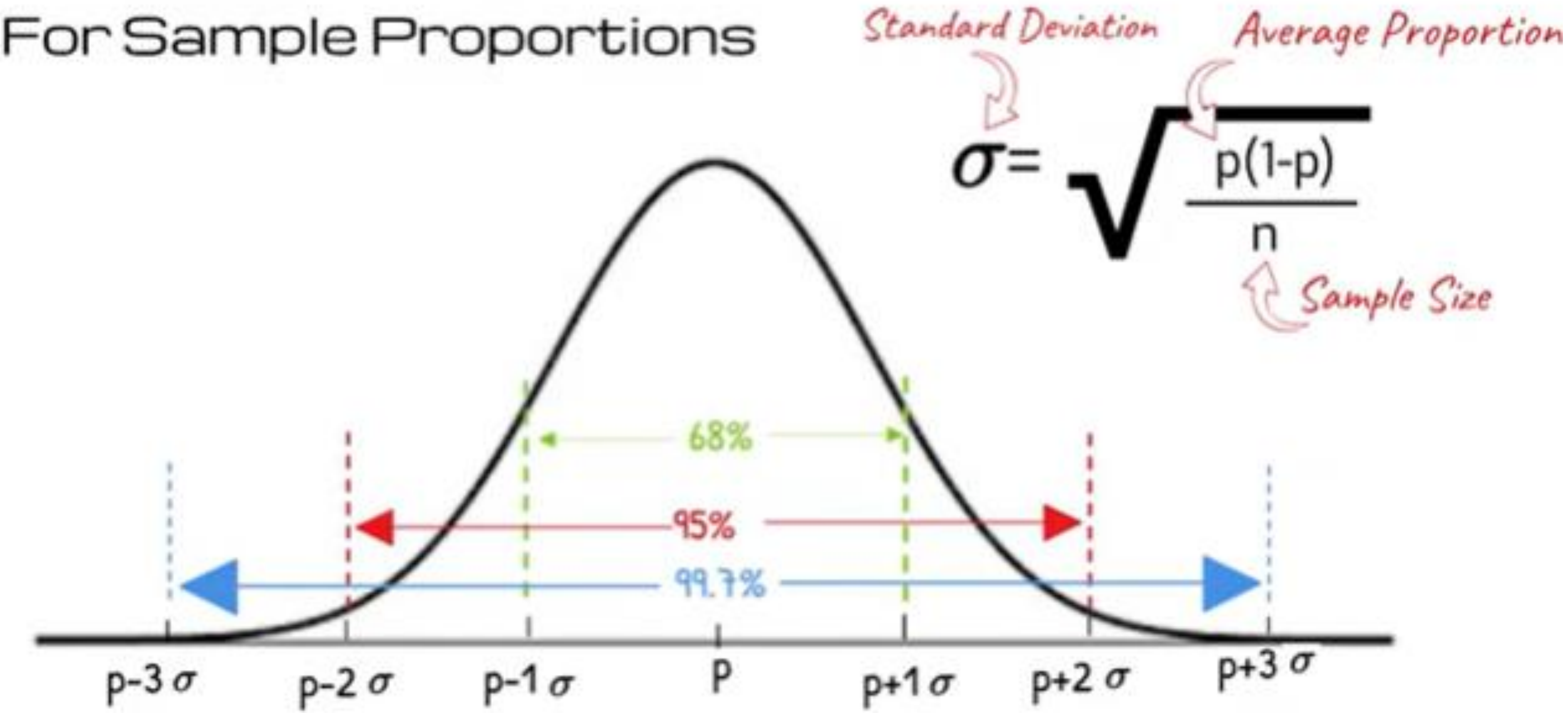
$$P(\hat{p} \leq 0.8) = P(z \leq -0.2549) \approx 0.3994$$

Step 5: State your conclusion

The probability of no more than 80 voters in a randomly selected sample of 100 voters being registered Democrats is approximately 0.3994

The Central Limit Theorem

For Sample Proportions



Standard Deviation

Average Proportion

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

Sample Size

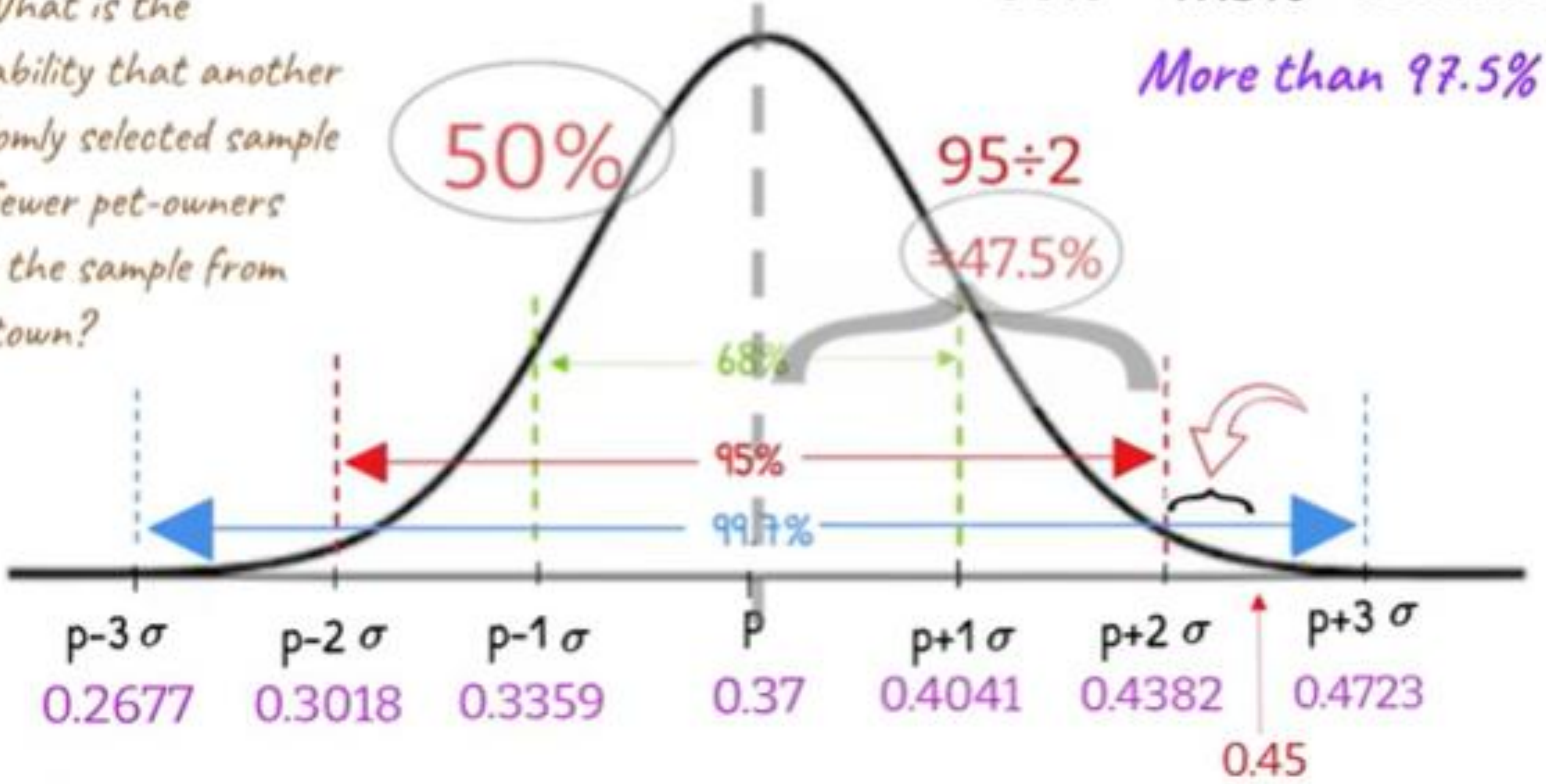
The Central Limit Theorem

For Sample Proportions

Ex. What is the probability that another randomly selected sample has fewer pet-owners than the sample from this town?

50% + 47.5% + a bit more!

More than 97.5%



The Central Limit Theorem For Sample Proportions

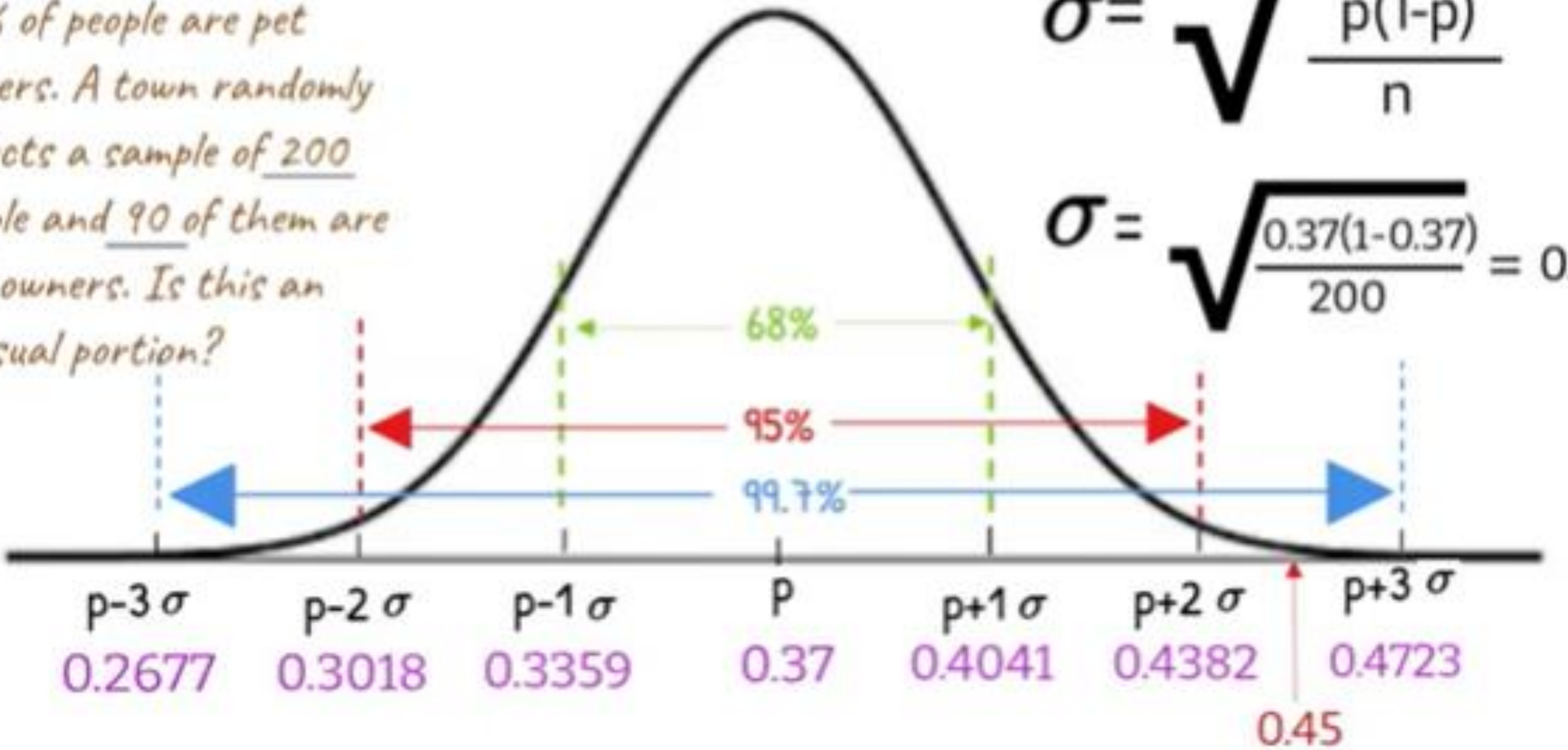
In the town 90/200 have pets!

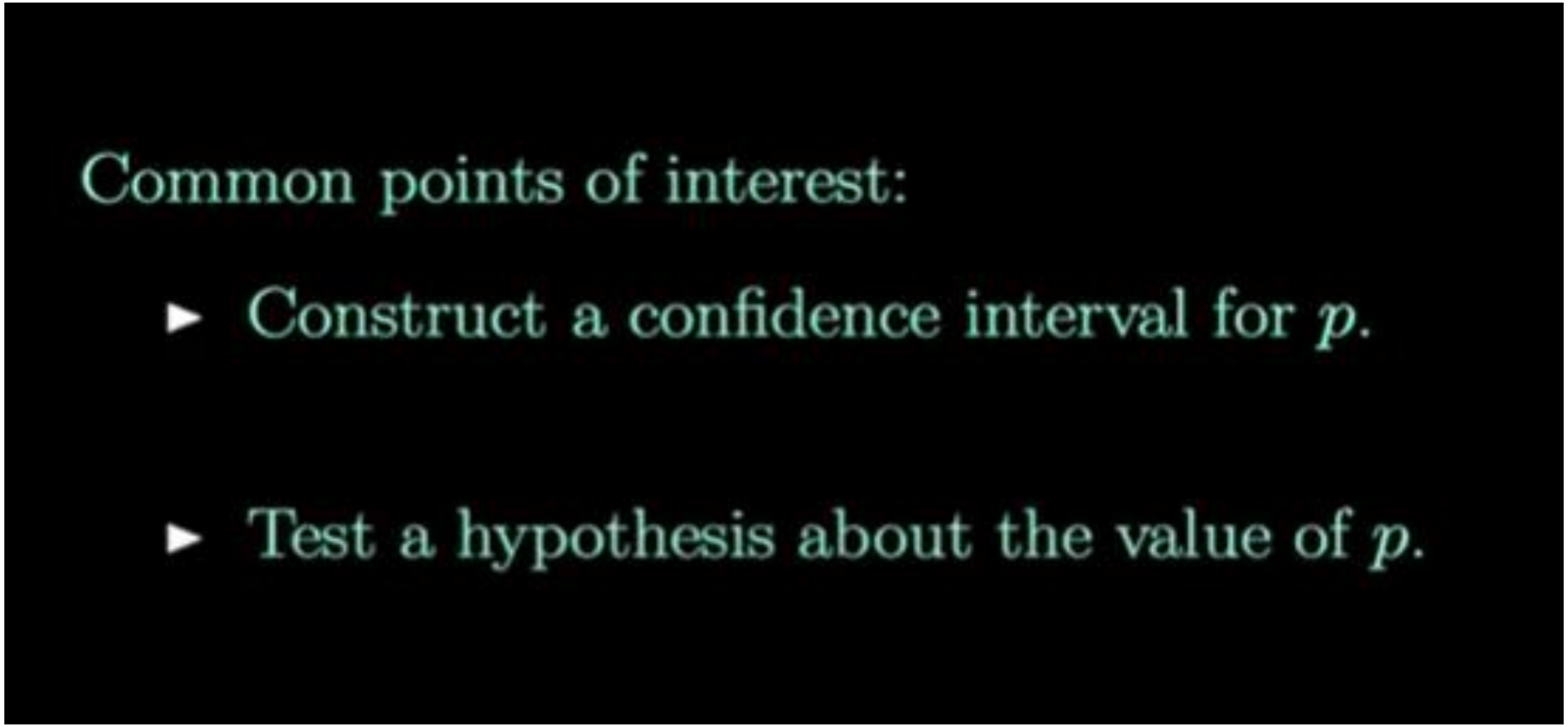
90÷200=0.45

Ex. The WPA states that 37% of people are pet owners. A town randomly collects a sample of 200 people and 90 of them are pet-owners. Is this an unusual portion?

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

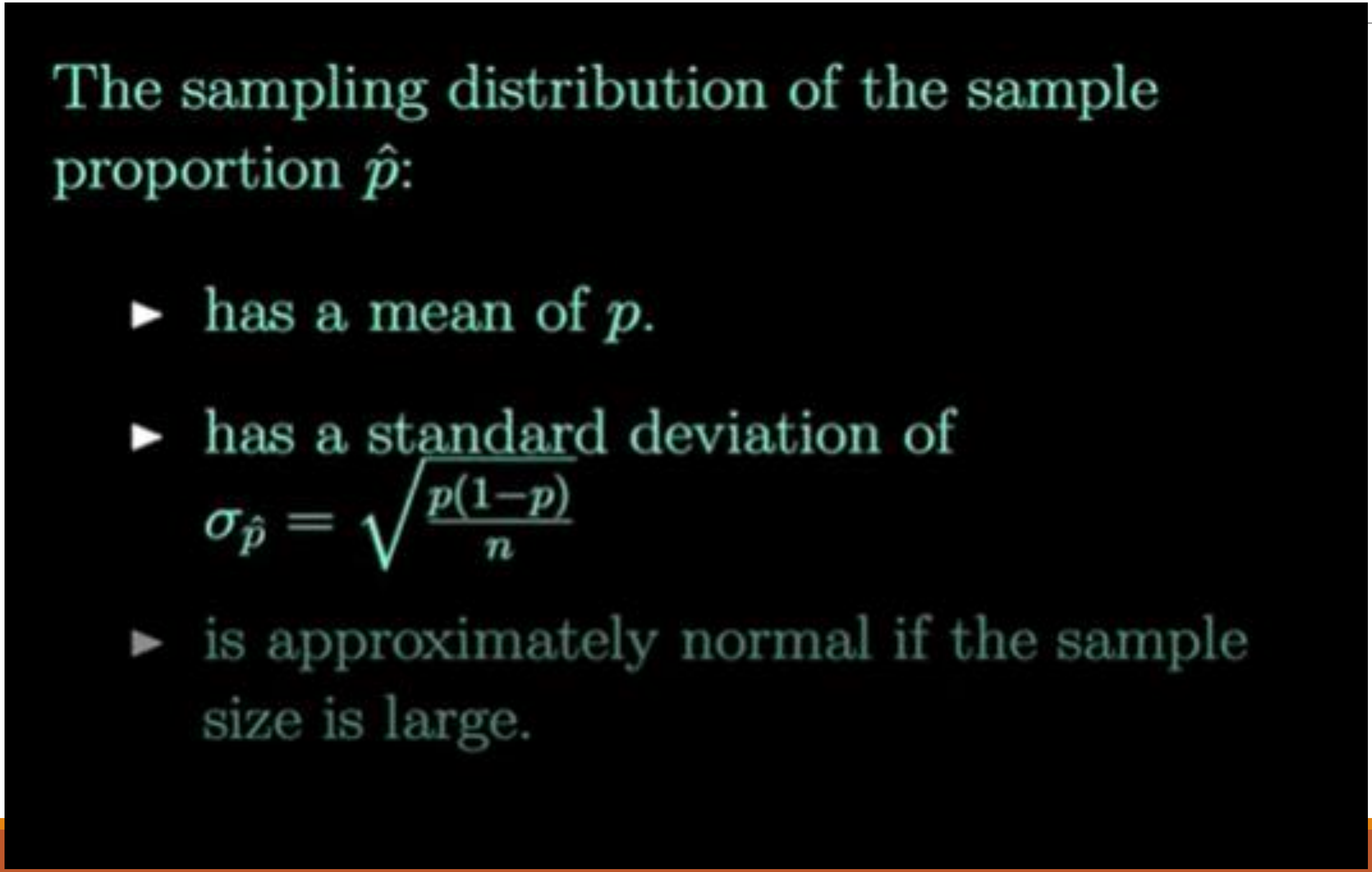
$$\sigma = \sqrt{\frac{0.37(1-0.37)}{200}} = 0.0341$$





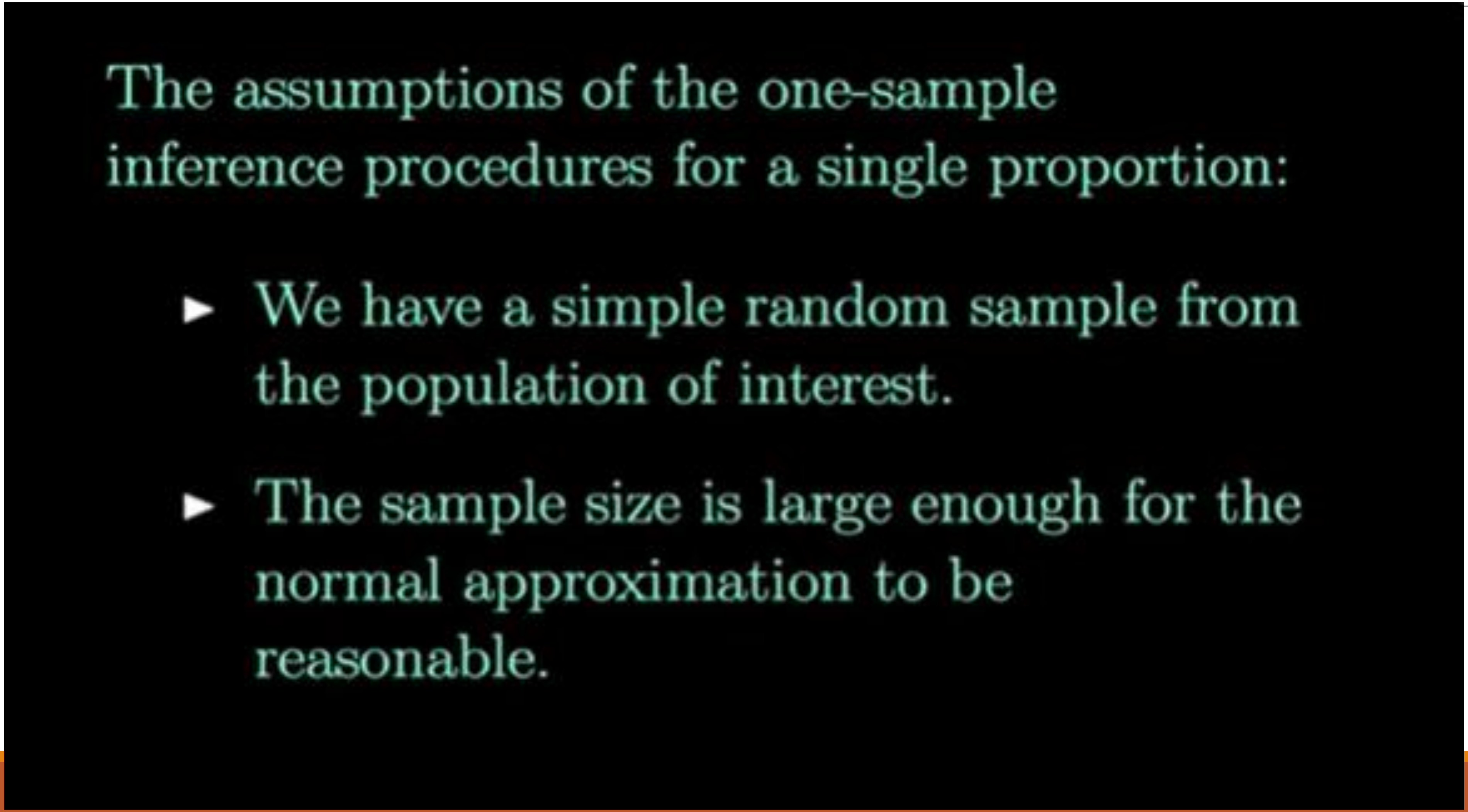
Common points of interest:

- Construct a confidence interval for p.
- Test a hypothesis about the value of p.



The sampling distribution of the sample proportion $\hat{p}$:

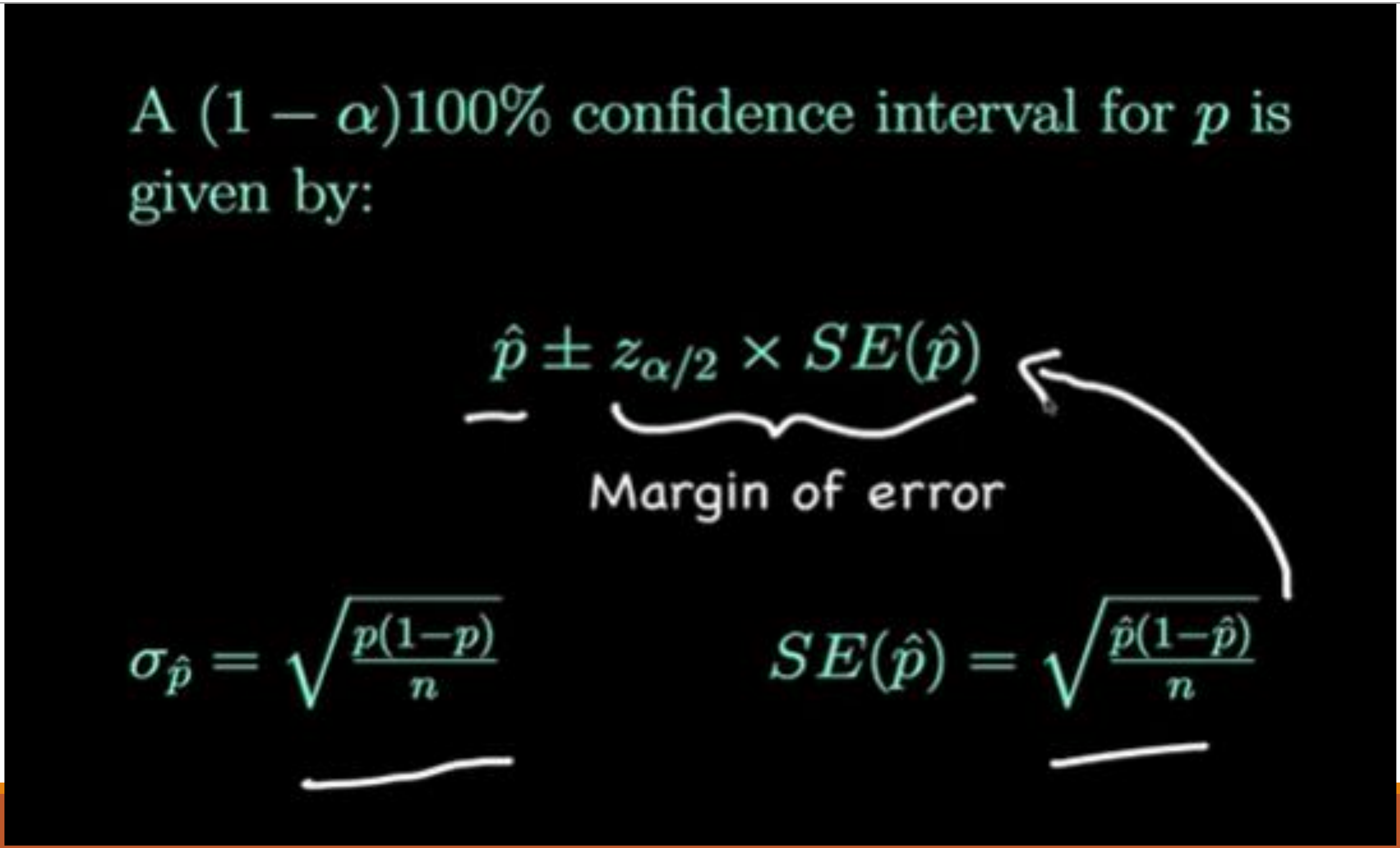
- has a mean of p.
- has a standard deviation of $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$
- is approximately normal if the sample size is large.



The assumptions of the one-sample inference procedures for a single proportion:

- We have a simple random sample from the population of interest.
- The sample size is large enough for the normal approximation to be reasonable.

CI

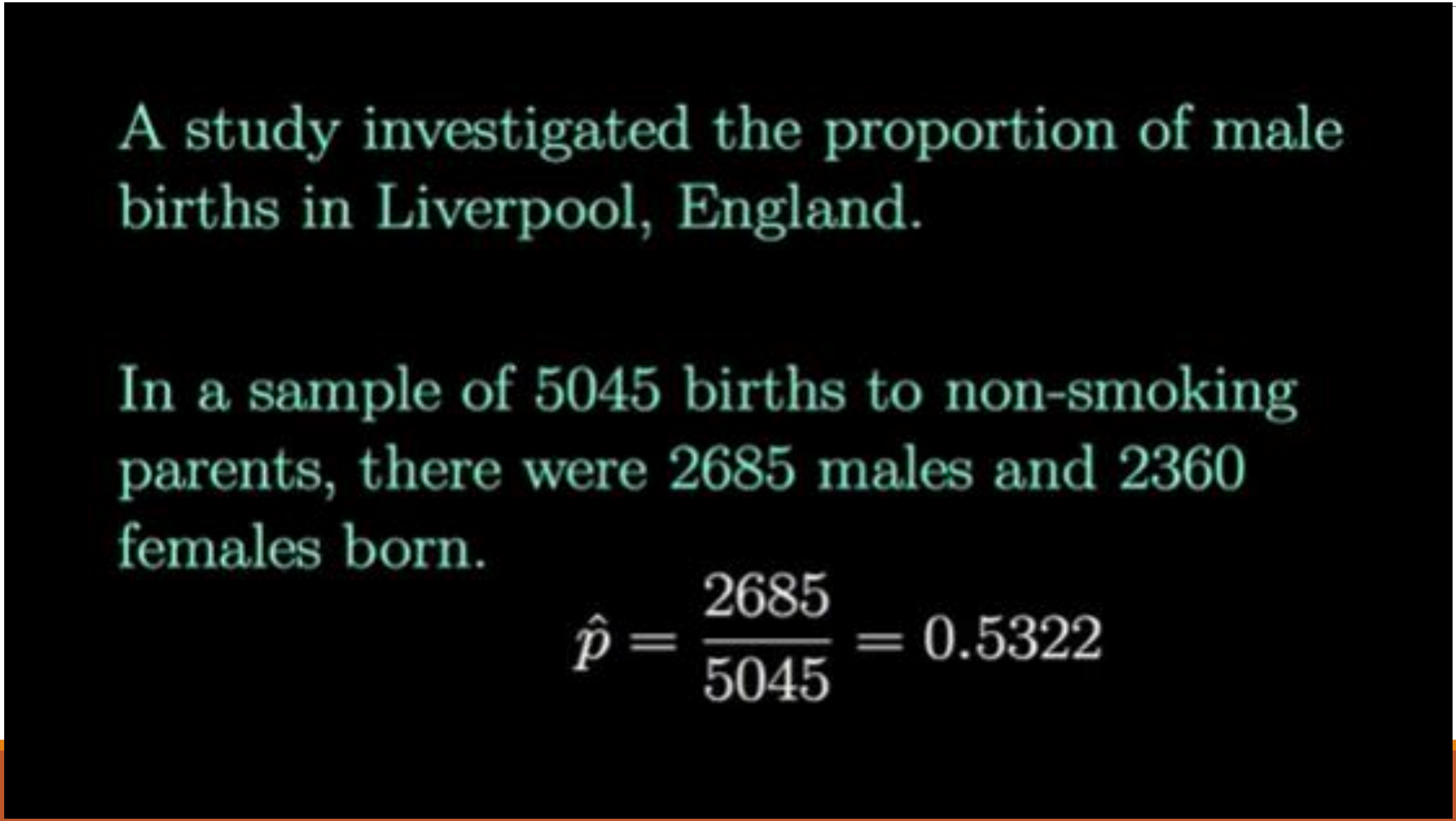


A $(1-\alpha)100\%$ confidence interval for p is given by:

$$\underline{\hat{p}} \pm \underbrace{z_{\alpha/2} \times SE(\hat{p})}_{\text{Margin of error}}$$

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

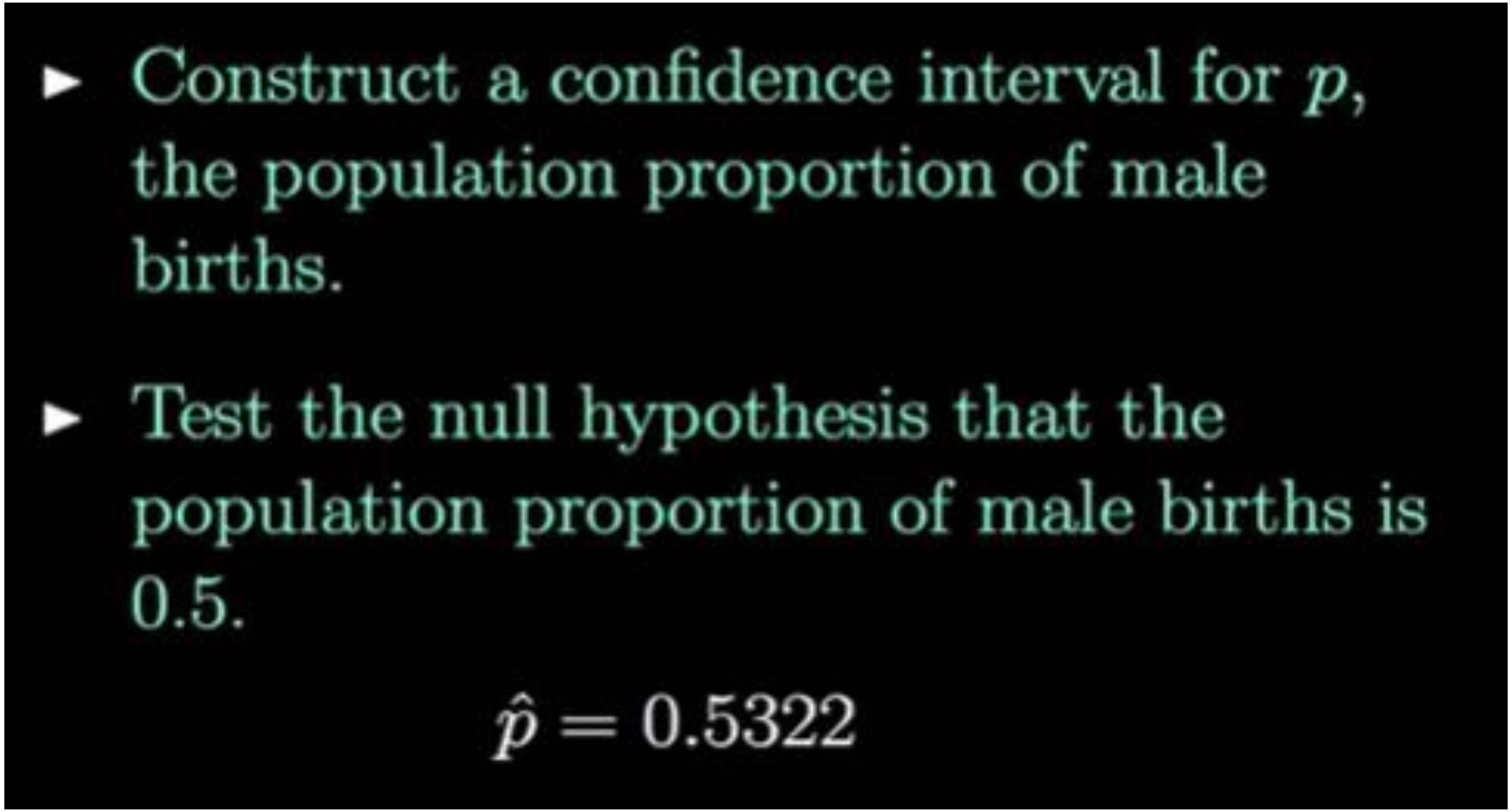
Example for CI



A study investigated the proportion of male births in Liverpool, England.

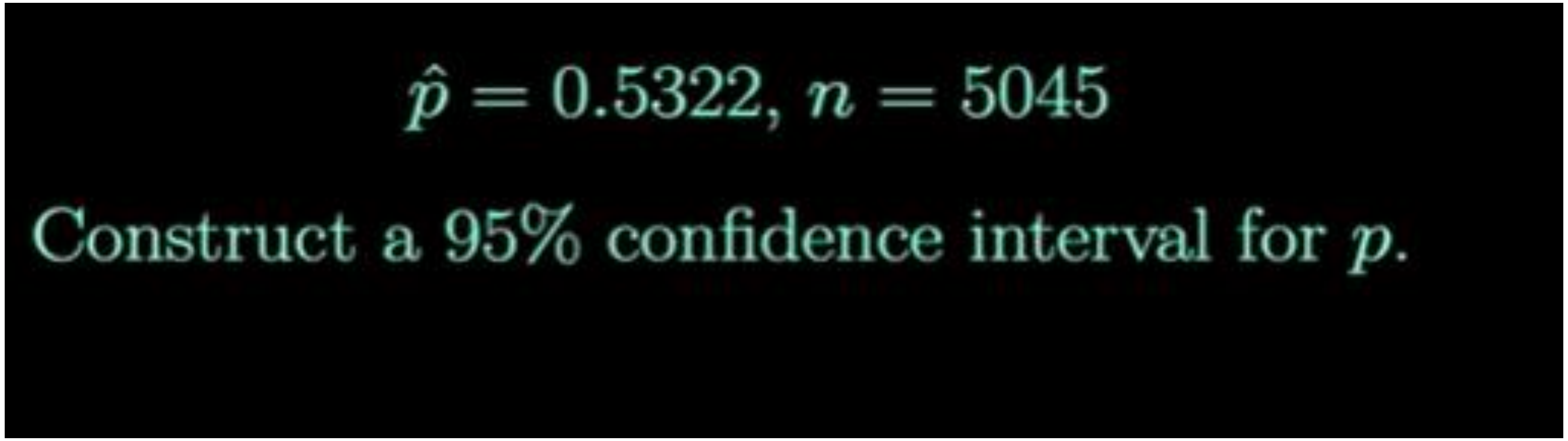
In a sample of 5045 births to non-smoking parents, there were 2685 males and 2360 females born.

$$\hat{p} = \frac{2685}{5045} = 0.5322$$



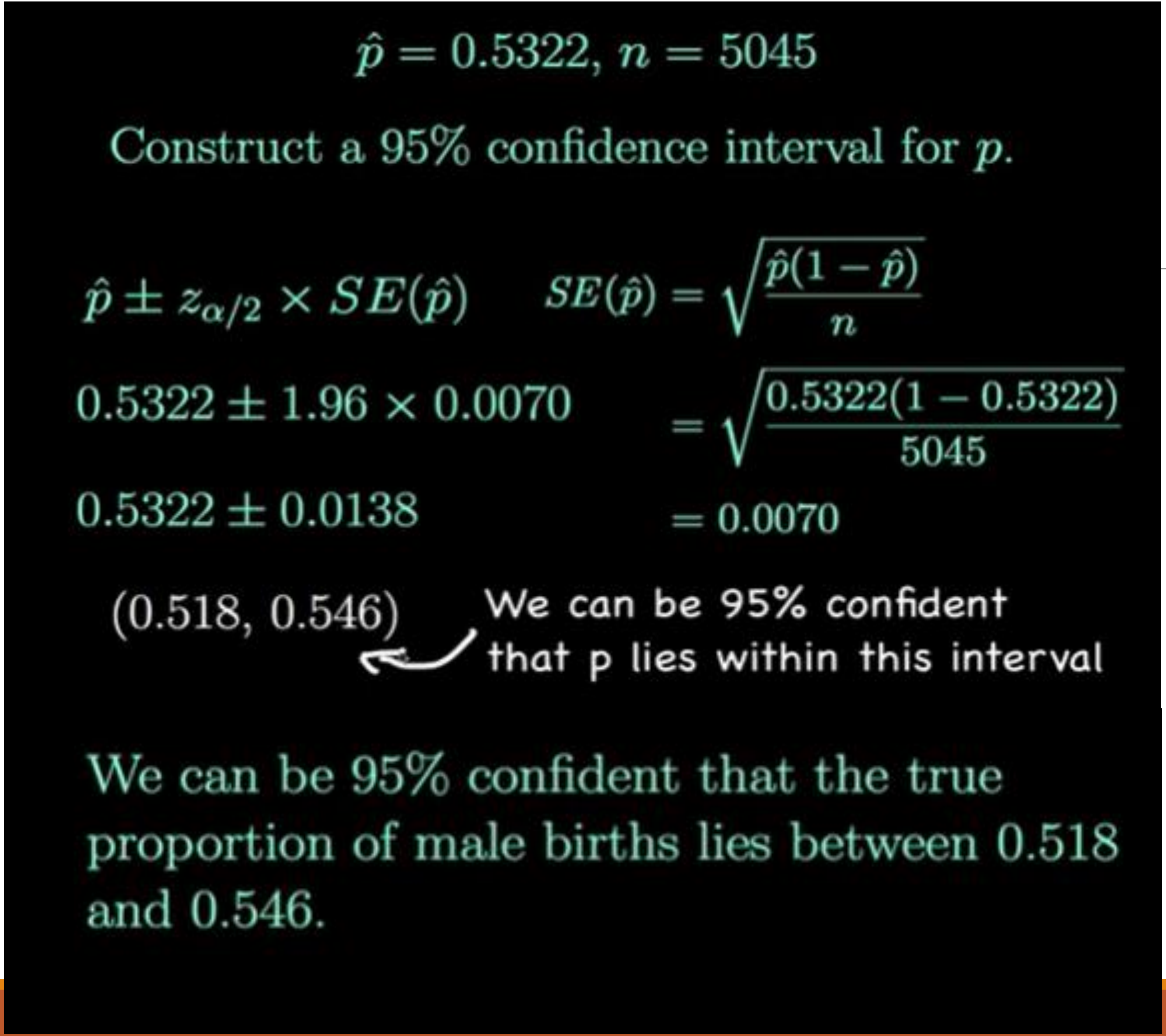
- Construct a confidence interval for p, the population proportion of male births.
- Test the null hypothesis that the population proportion of male births is 0.5.

$$\hat{p} = 0.5322$$



$$\hat{p} = 0.5322,\ n = 5045$$

Construct a 95% confidence interval for p.



$$\hat{p} = 0.5322,\ n = 5045$$

Construct a 95% confidence interval for p.

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

$$0.5322 \pm 1.96 \times 0.0070 \qquad = \sqrt{\frac{0.5322(1-0.5322)}{5045}}$$

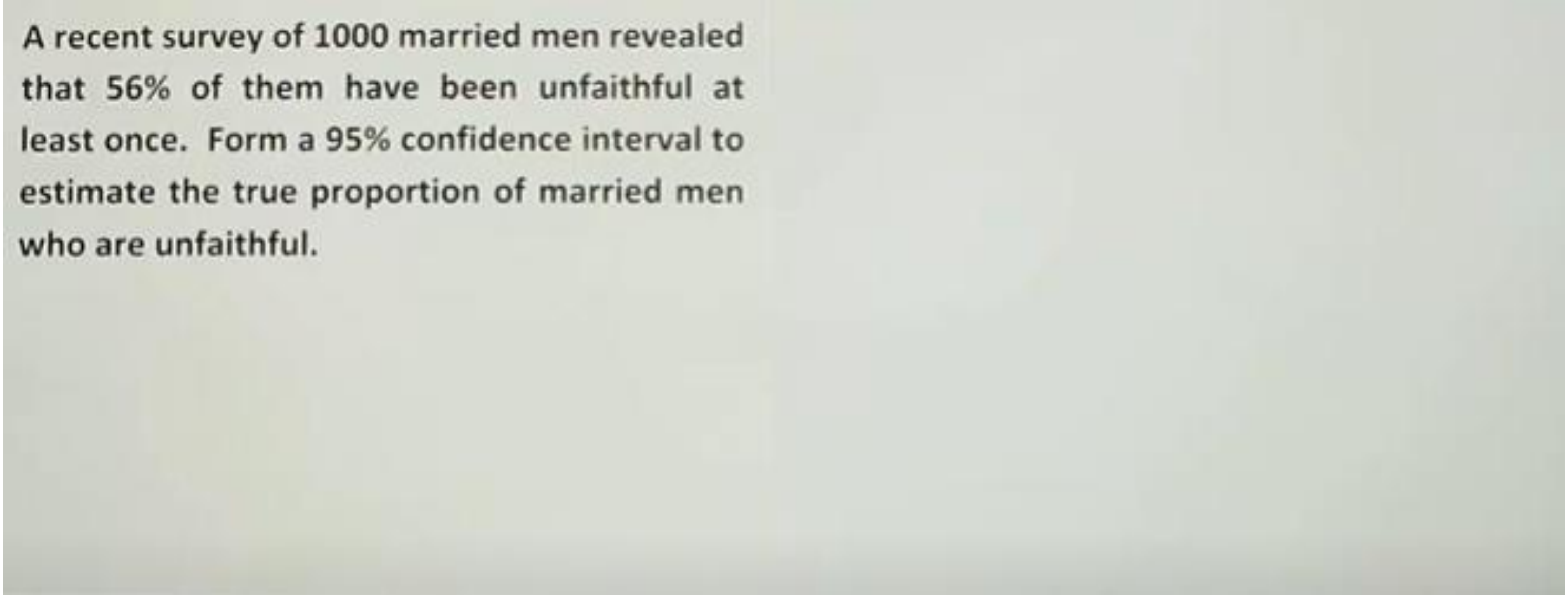
$$0.5322 \pm 0.0138 \qquad = 0.0070$$

$(0.518,\ 0.546)$

We can be 95% confident
that p lies within this interval

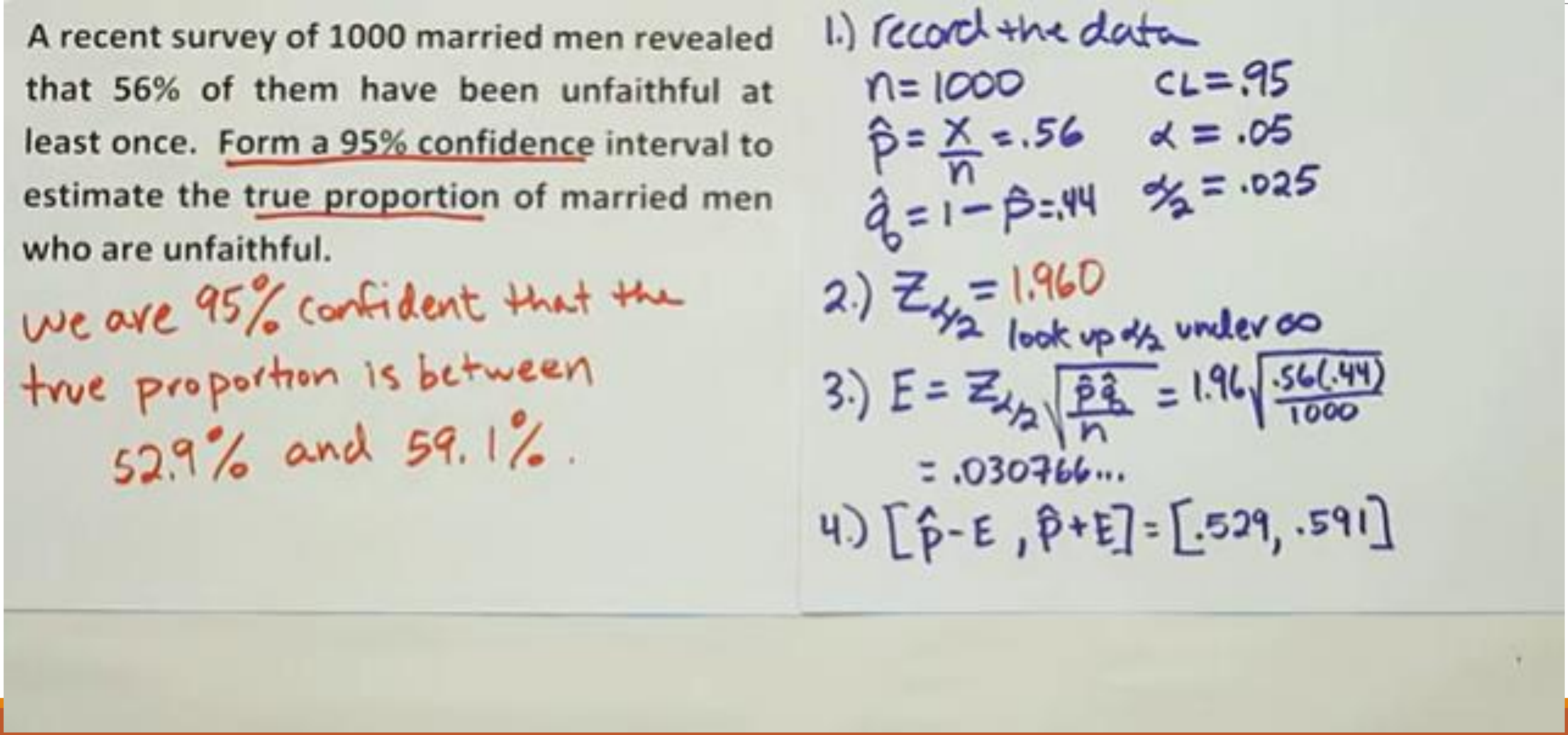
We can be 95% confident that the true proportion of male births lies between 0.518 and 0.546.

CI- Example



A recent survey of 1000 married men revealed that 56% of them have been unfaithful at least once. Form a 95% confidence interval to estimate the true proportion of married men who are unfaithful.

CI



A recent survey of 1000 married men revealed that 56% of them have been unfaithful at least once. Form a 95% confidence interval to estimate the true proportion of married men who are unfaithful.

we are 95% confident that the true proportion is between 52.9% and 59.1%.

1.) record the data

$n = 1000$ $\quad CL = .95$

$\hat{p} = \frac{x}{n} = .56$ $\quad \alpha = .05$

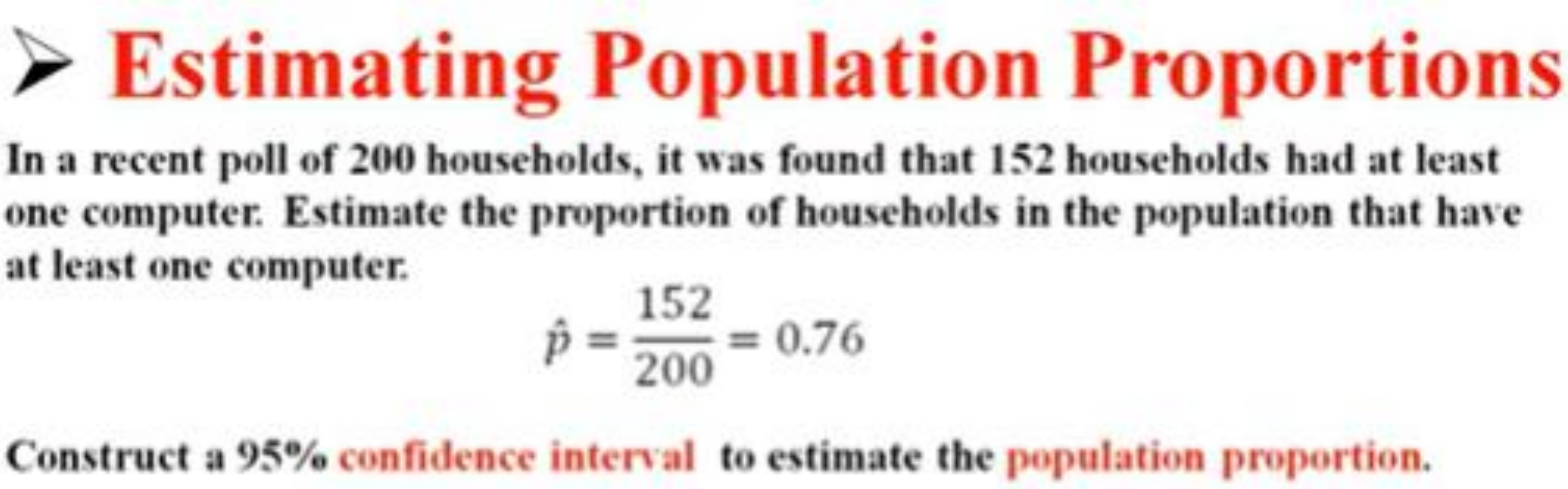
$\hat{q} = 1 - \hat{p} = .44$ $\quad \alpha/2 = .025$

2.) $Z_{\alpha/2} = 1.960$ look up $\alpha/2$ under ∞

3.) $E = Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} = 1.96\sqrt{\frac{.56(.44)}{1000}}$

$= .030766\ldots$

4.) $[\hat{p} - E, \hat{p} + E] = [.529, .591]$

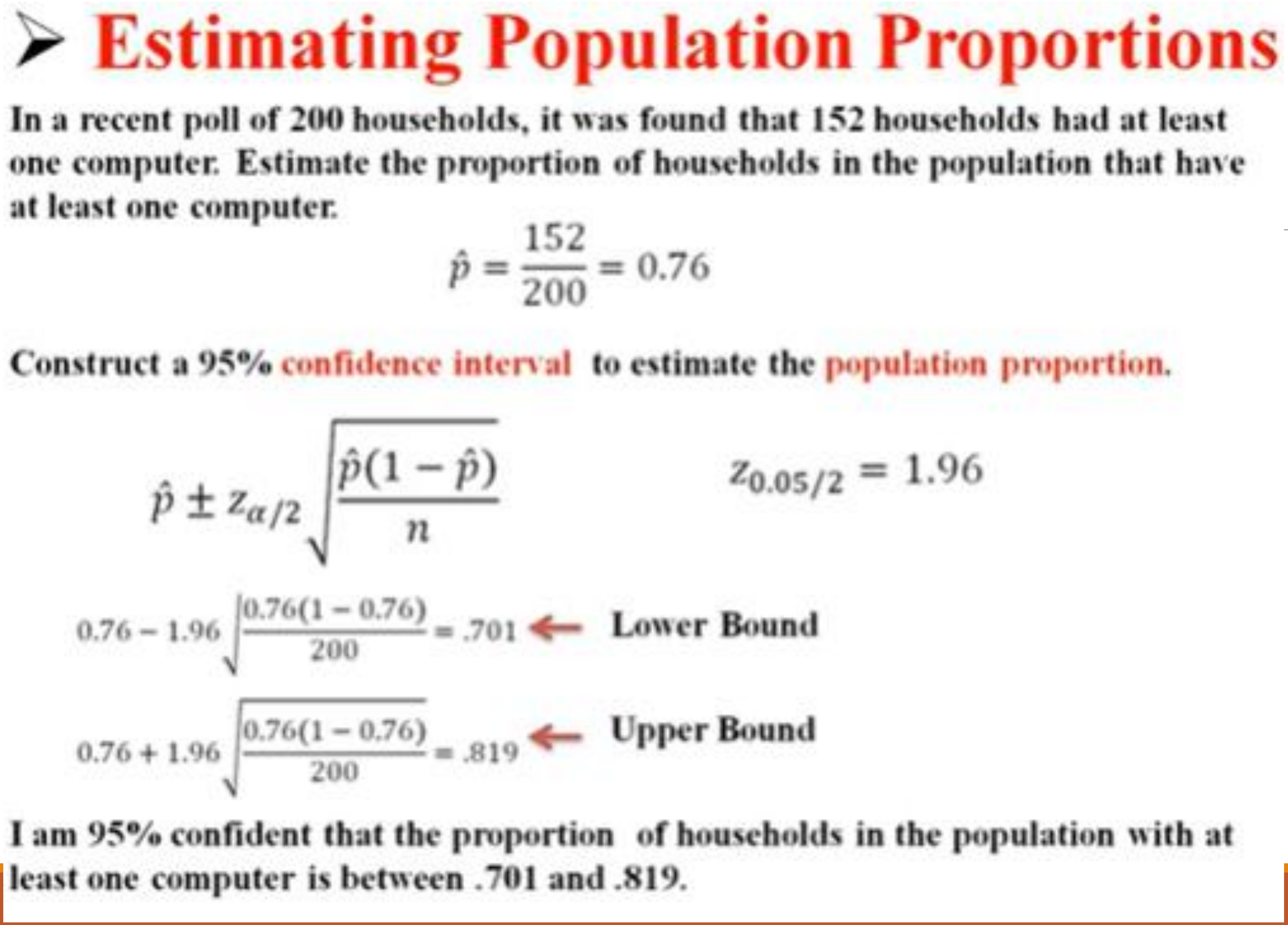


➢ Estimating Population Proportions

In a recent poll of 200 households, it was found that 152 households had at least one computer. Estimate the proportion of households in the population that have at least one computer.

$$\hat{p} = \frac{152}{200} = 0.76$$

Construct a 95% confidence interval to estimate the population proportion.



➢ Estimating Population Proportions

In a recent poll of 200 households, it was found that 152 households had at least one computer. Estimate the proportion of households in the population that have at least one computer.

$$\hat{p} = \frac{152}{200} = 0.76$$

Construct a 95% confidence interval to estimate the population proportion.

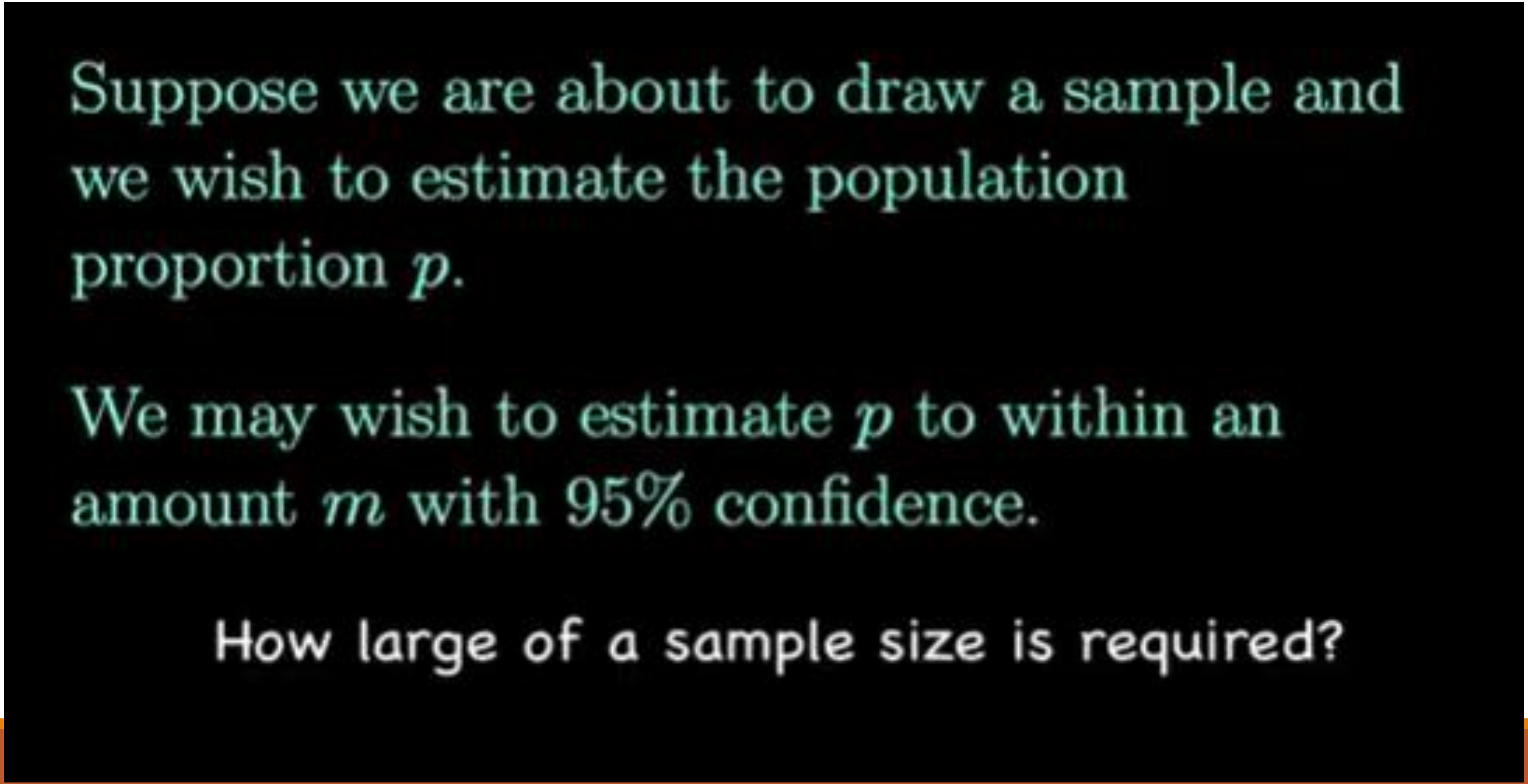
$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad z_{0.05/2} = 1.96$$

$$0.76 - 1.96\sqrt{\frac{0.76(1-0.76)}{200}} = .701 \leftarrow \text{Lower Bound}$$

$$0.76 + 1.96\sqrt{\frac{0.76(1-0.76)}{200}} = .819 \leftarrow \text{Upper Bound}$$

I am 95% confident that the proportion of households in the population with at least one computer is between .701 and .819.

Sampling variability for CI and Required size for this

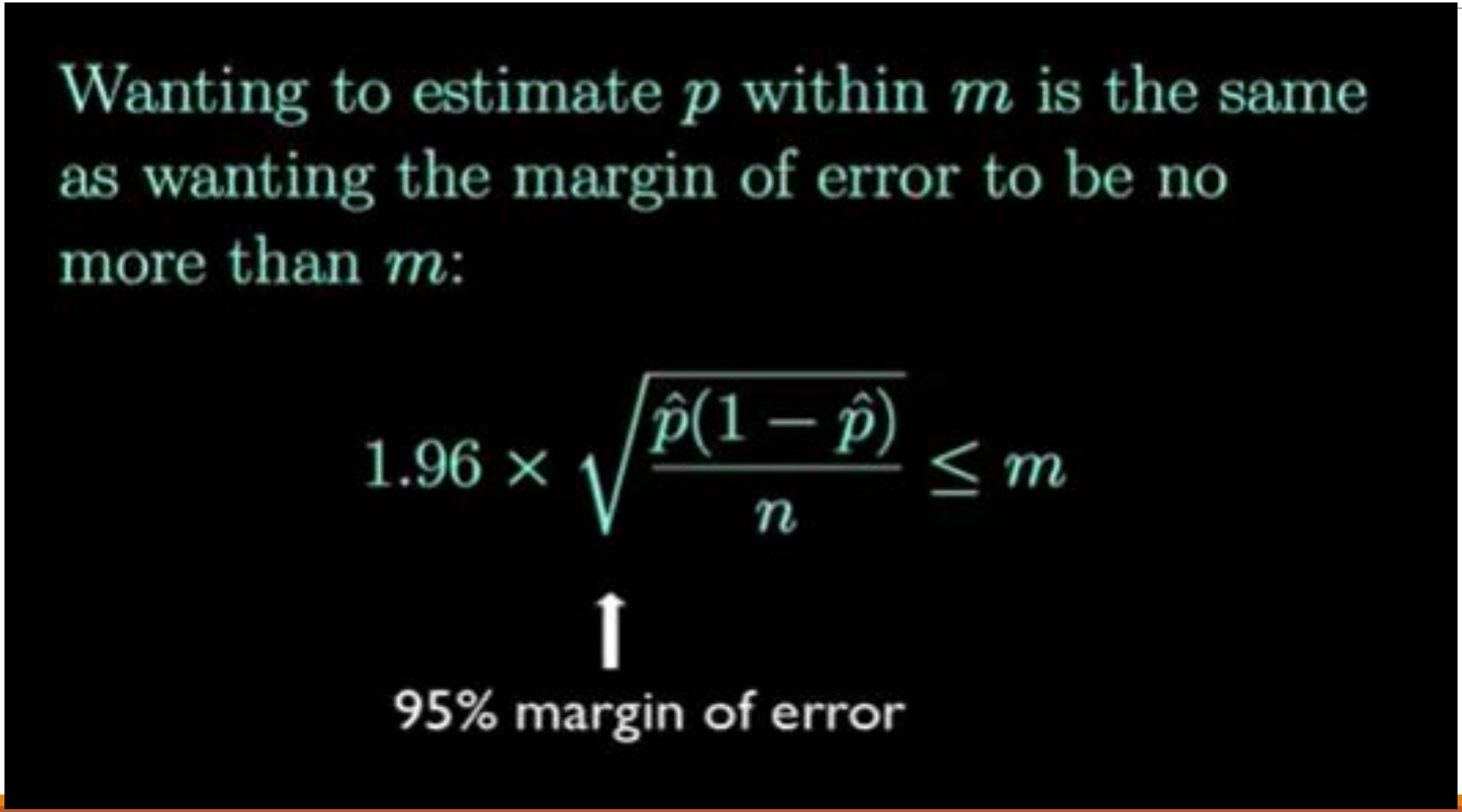


Suppose we are about to draw a sample and we wish to estimate the population proportion p.

We may wish to estimate p to within an amount m with 95% confidence.

How large of a sample size is required?

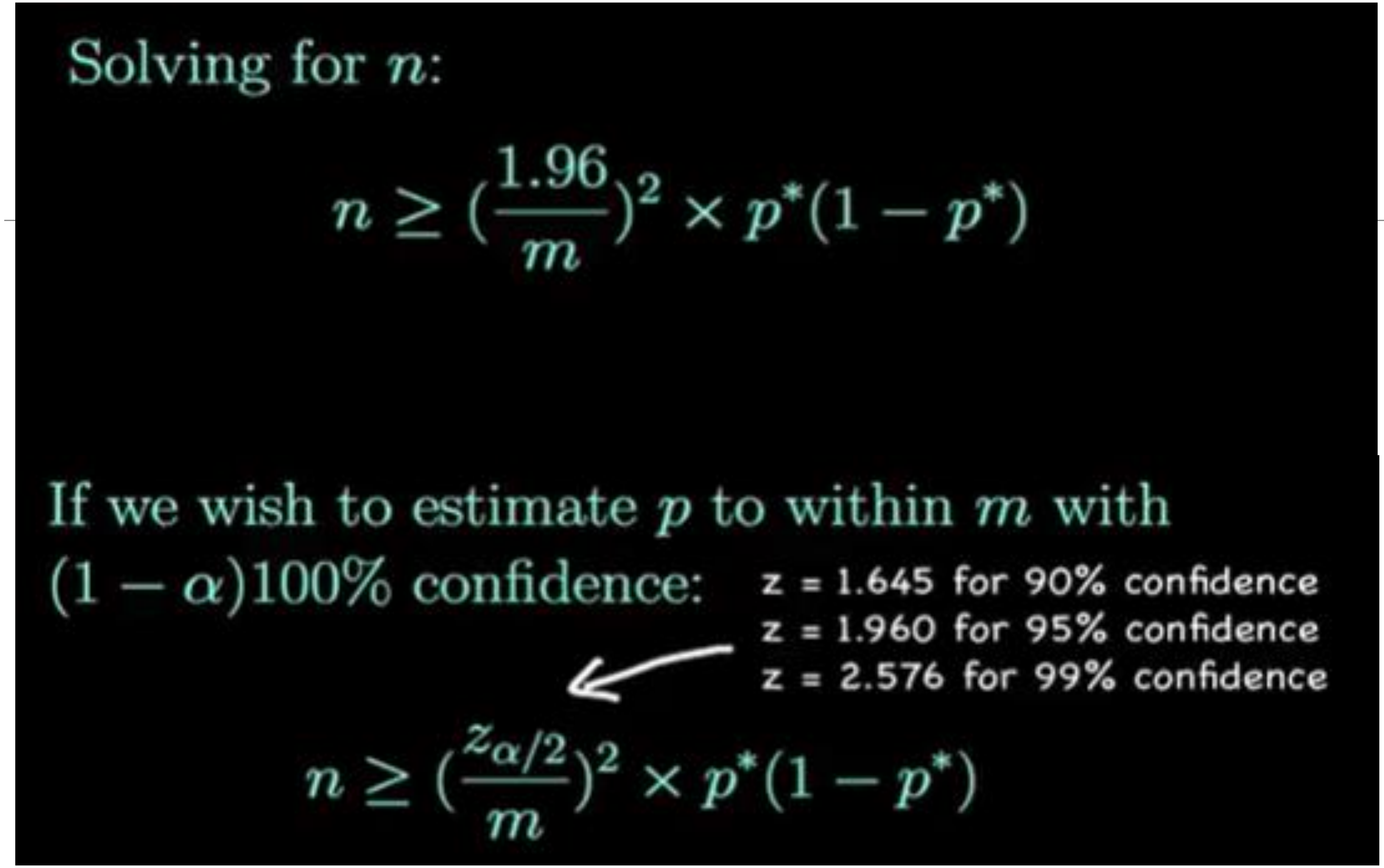
Sampling variability for CI and Required size for this



Wanting to estimate p within m is the same as wanting the margin of error to be no more than m:

$$\underbrace{1.96 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}_{\uparrow} \leq m$$

95% margin of error



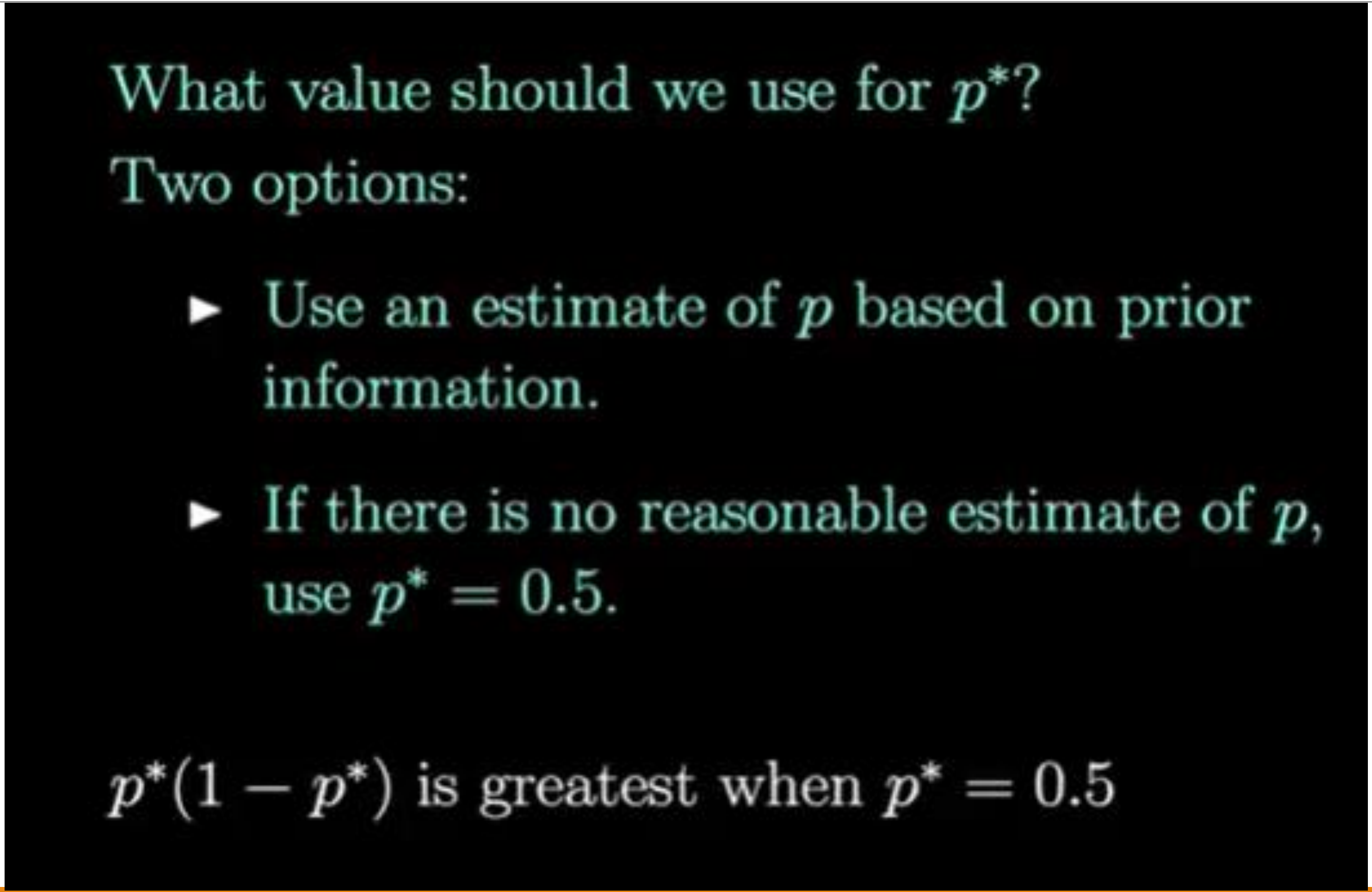
Solving for n:

$$n \geq (\frac{1.96}{m})^2 \times p^*(1-p^*)$$

If we wish to estimate p to within m with $(1-\alpha)100\%$ confidence:

z = 1.645 for 90% confidence
z = 1.960 for 95% confidence
z = 2.576 for 99% confidence

$$n \geq (\frac{z_{\alpha/2}}{m})^2 \times p^*(1-p^*)$$



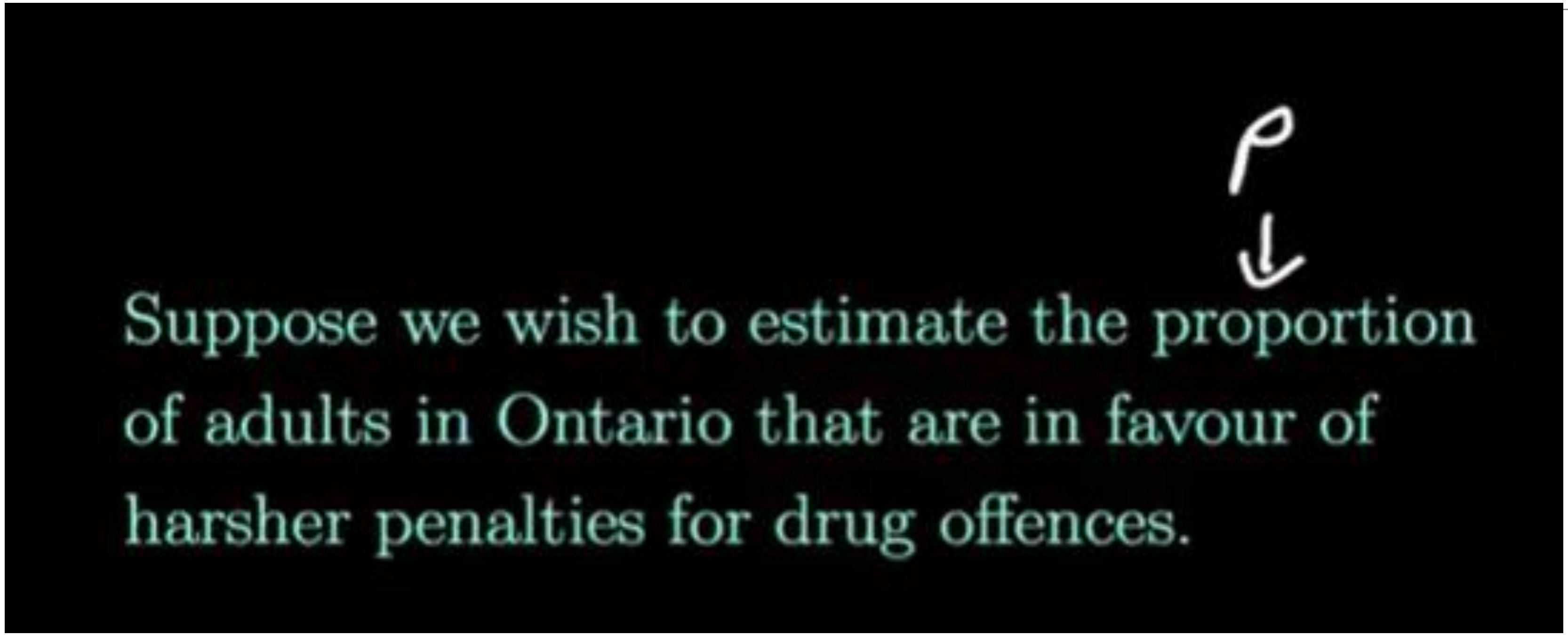
What value should we use for p^*?

Two options:

- Use an estimate of p based on prior information.
- If there is no reasonable estimate of p, use $p^* = 0.5$.

$p^*(1 - p^*)$ is greatest when $p^* = 0.5$

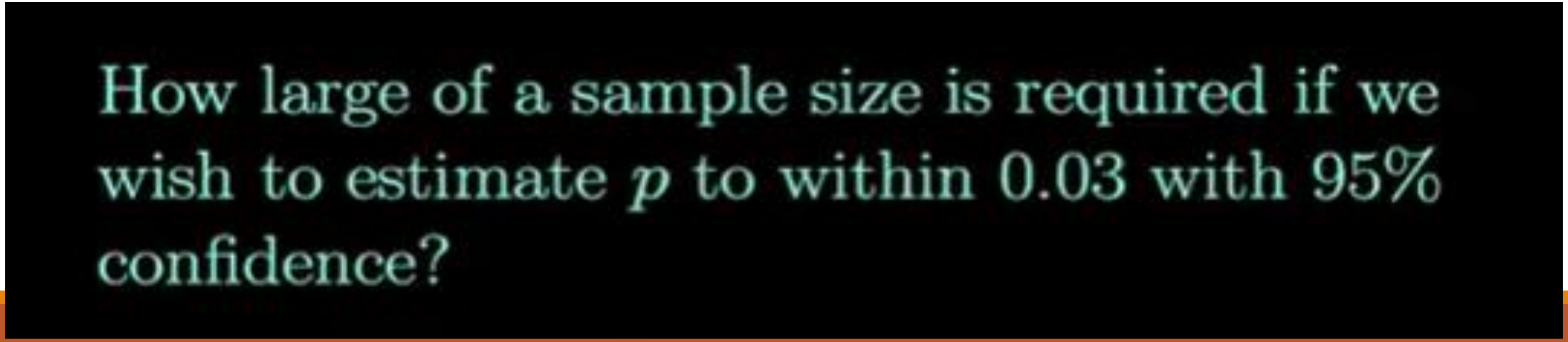
Sampling variability for CI and Required size for this



p

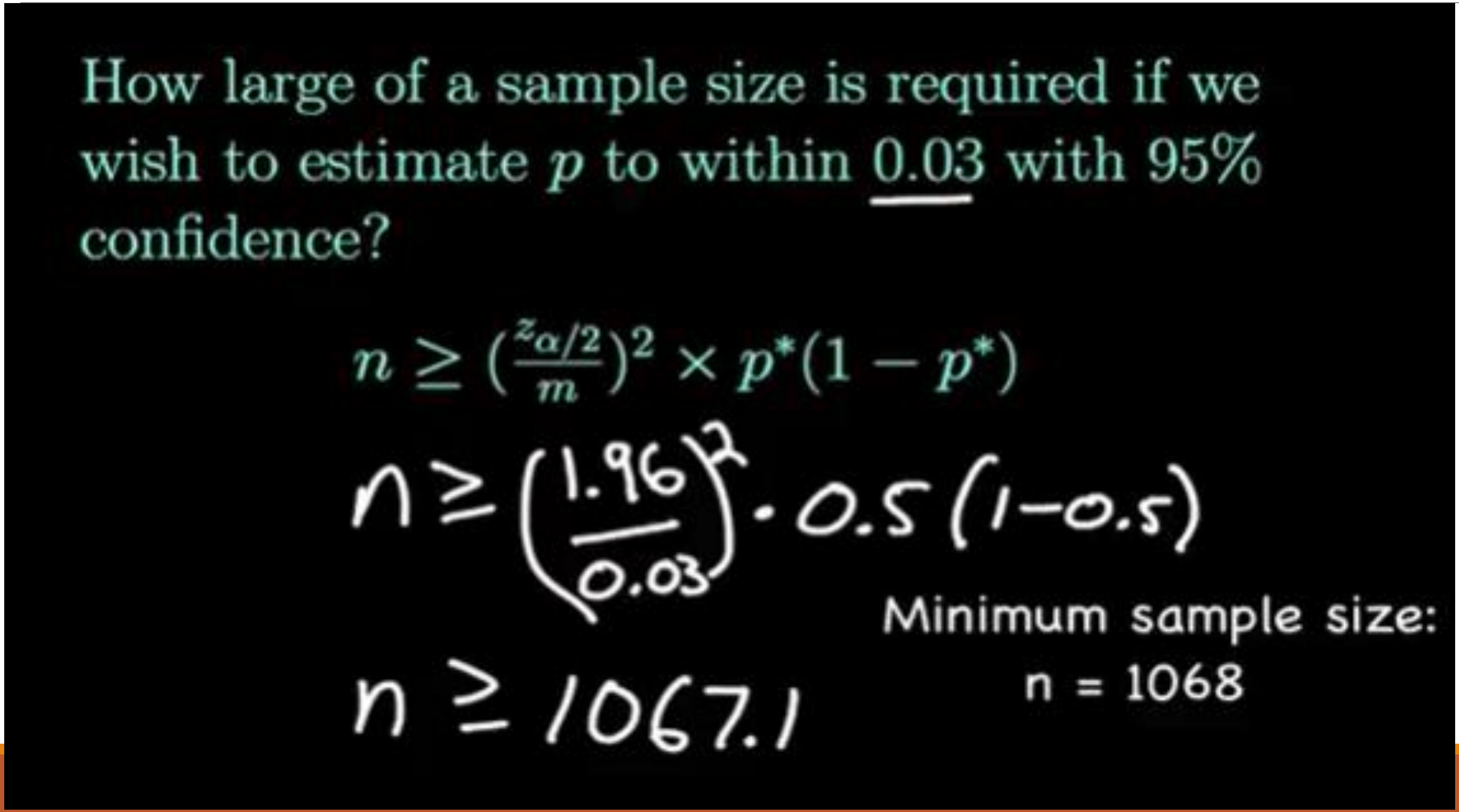
↓

Suppose we wish to estimate the proportion of adults in Ontario that are in favour of harsher penalties for drug offences.



How large of a sample size is required if we wish to estimate p to within 0.03 with 95% confidence?

Sampling variability for CI and Required size for this



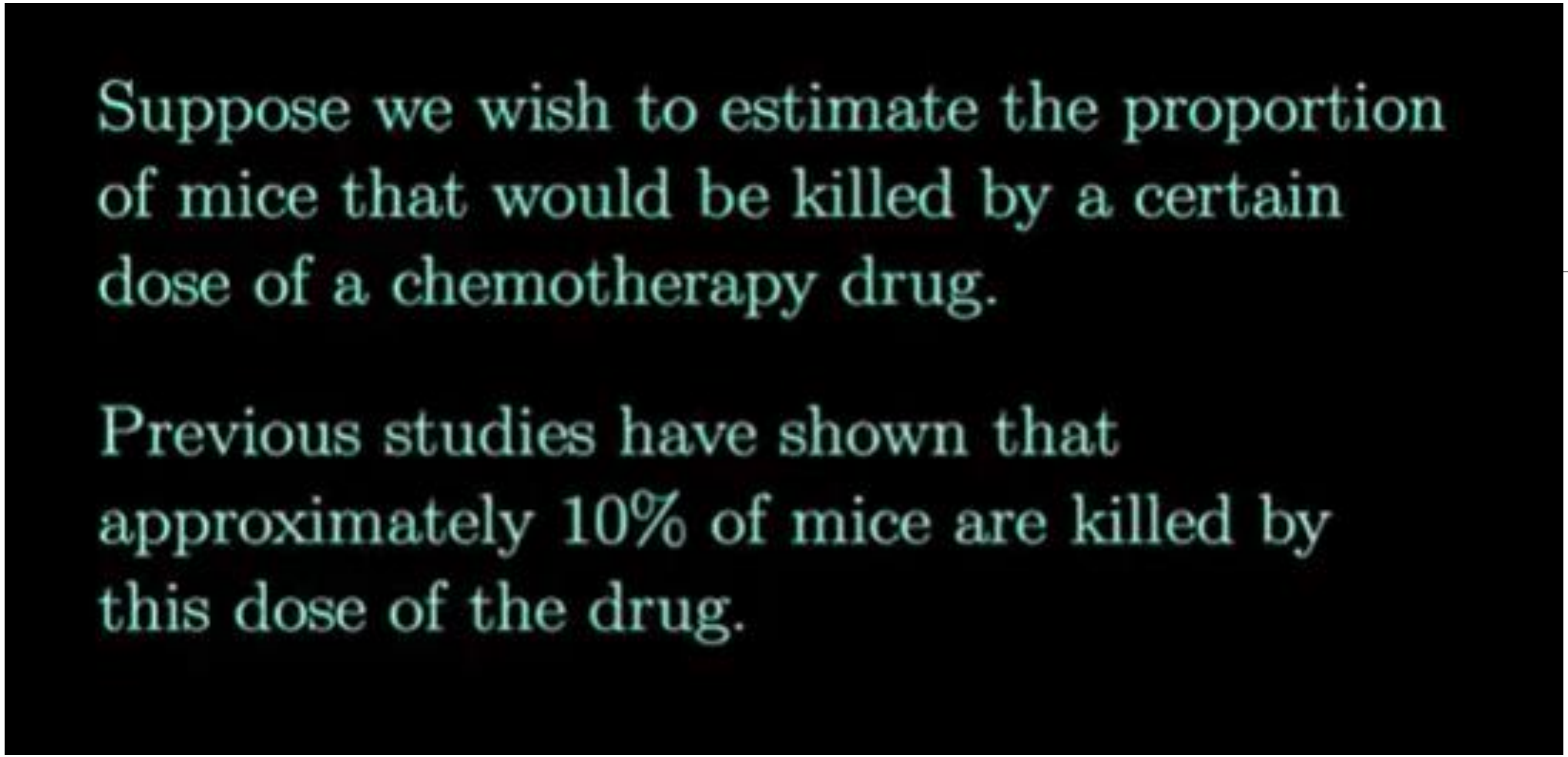
How large of a sample size is required if we wish to estimate p to within $\underline{0.03}$ with 95% confidence?

$$n \geq \left(\frac{z_{\alpha/2}}{m}\right)^2 \times p^*(1-p^*)$$

$$n \geq \left(\frac{1.96}{0.03}\right)^2 \cdot 0.5(1-0.5)$$

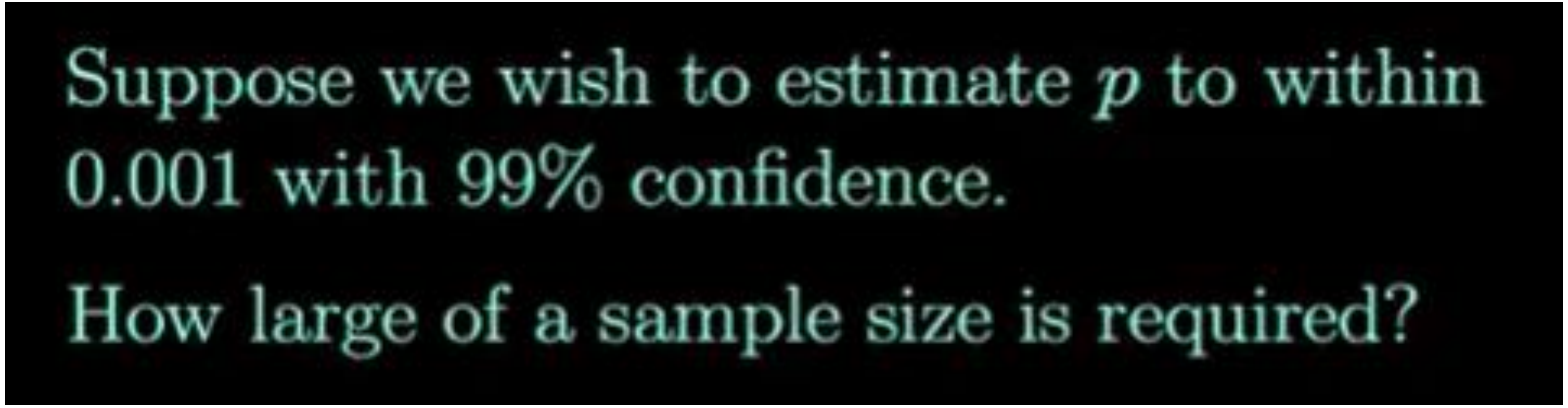
$$n \geq 1067.1$$

Minimum sample size:
n = 1068



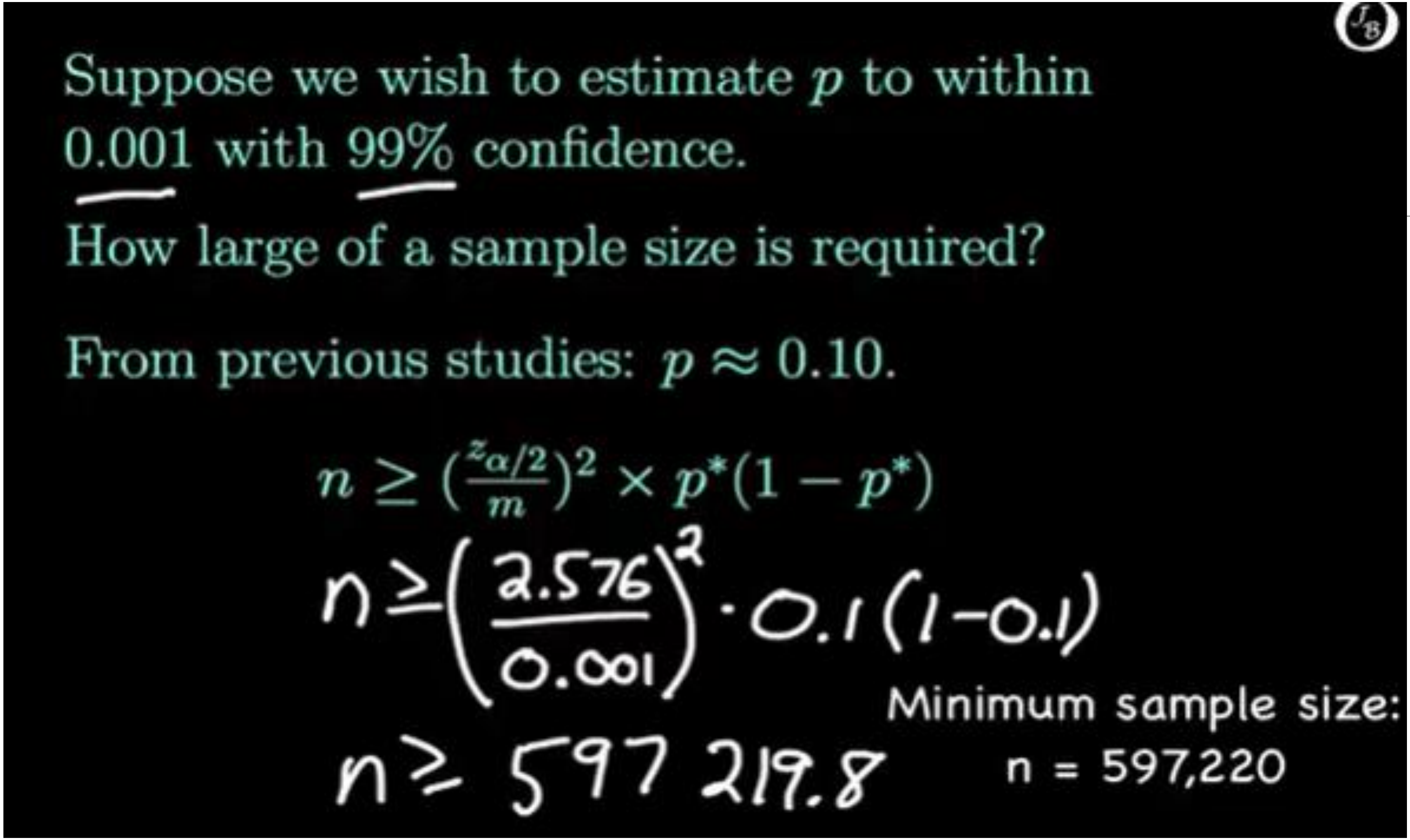
Suppose we wish to estimate the proportion of mice that would be killed by a certain dose of a chemotherapy drug.

Previous studies have shown that approximately 10% of mice are killed by this dose of the drug.



Suppose we wish to estimate p to within 0.001 with 99% confidence.

How large of a sample size is required?



Suppose we wish to estimate p to within 0.001 with 99% confidence.

How large of a sample size is required?

From previous studies: $p \approx 0.10$.

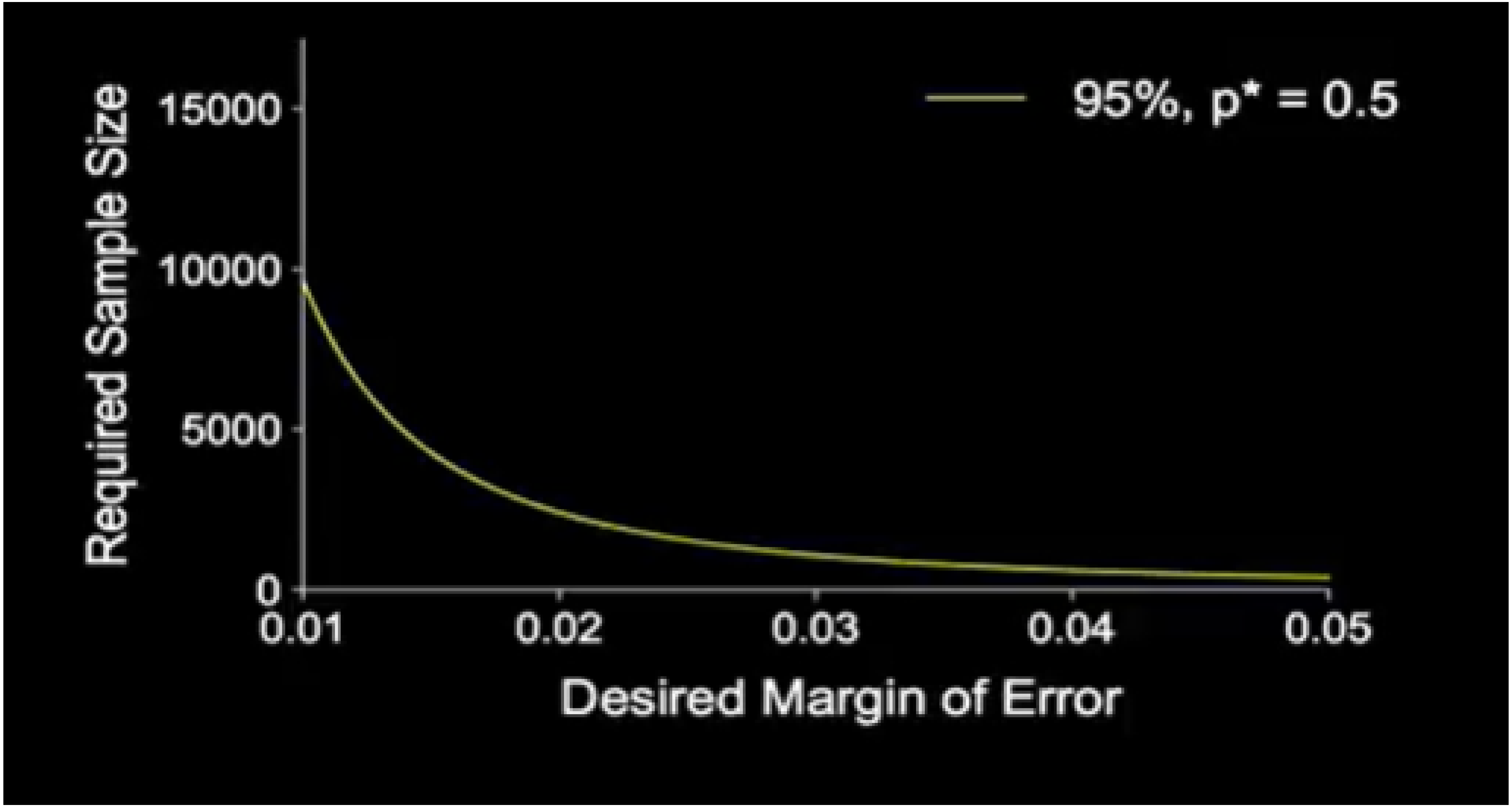
$$n \geq \left(\frac{z_{\alpha/2}}{m}\right)^2 \times p^*(1 - p^*)$$

$$n \geq \left(\frac{2.576}{0.001}\right)^2 \cdot 0.1(1-0.1)$$

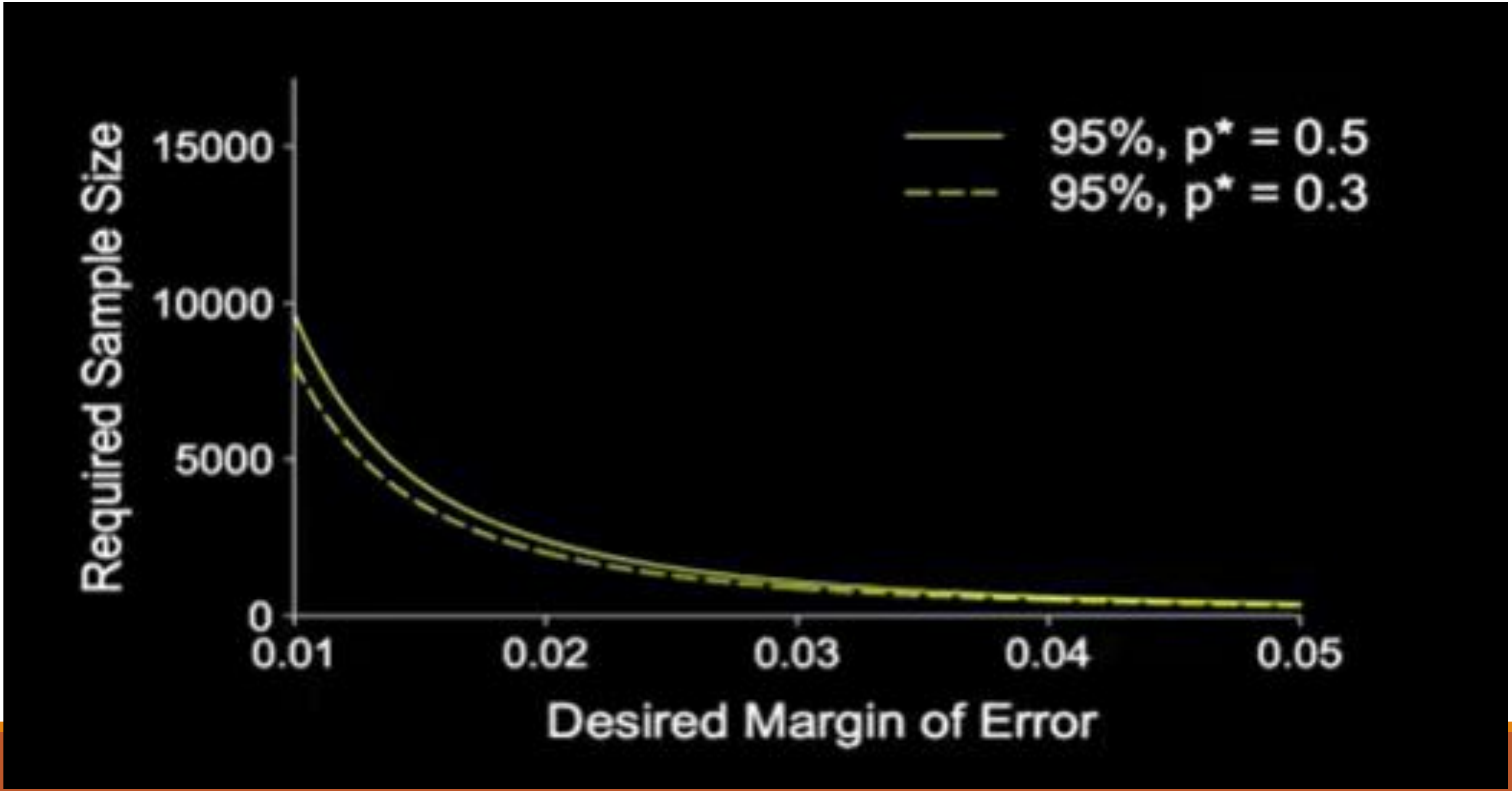
$$n \geq 597\,219.8$$

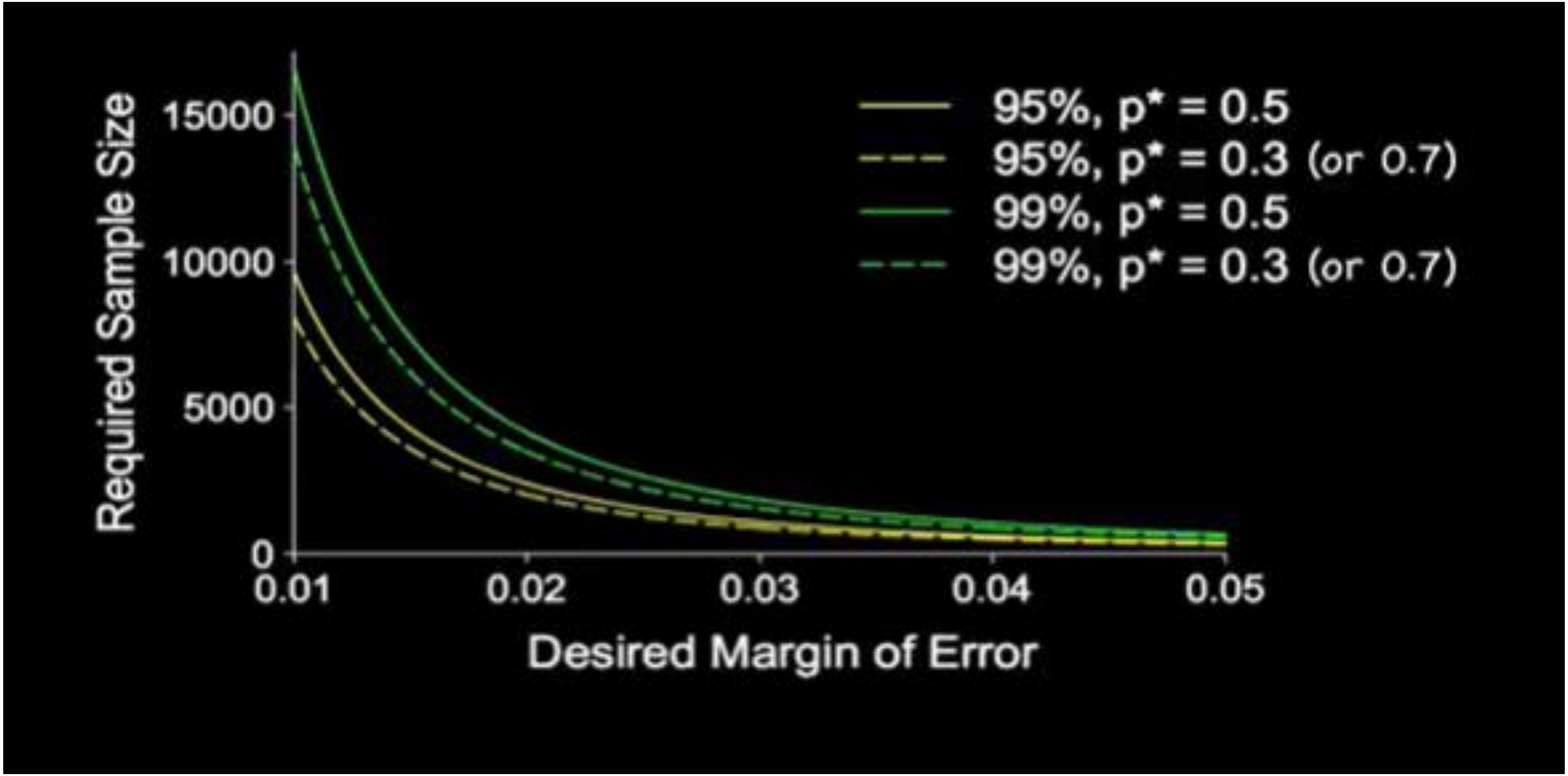
Minimum sample size:

n = 597,220

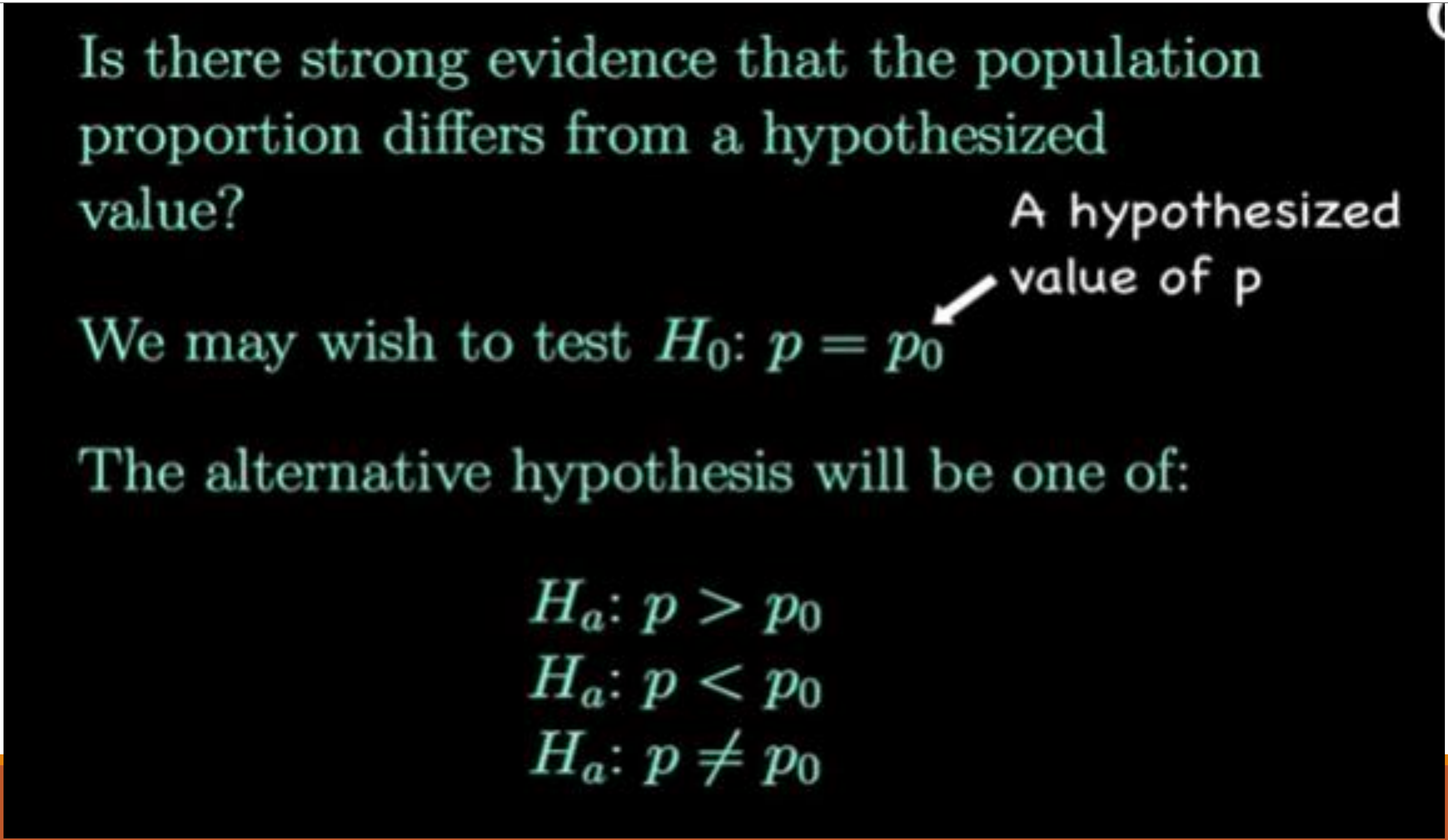

95%, p* = 0.5
Required Sample Size
15000
10000
5000
0
0.01
0.02
0.03
0.04
0.05
Desired Margin of Error

Sampling variability for CI and Required size for this




Required Sample Size
15000
10000
5000
0
0.01
0.02
0.03
0.04
0.05
Desired Margin of Error
95%, p* = 0.5
95%, p* = 0.3 (or 0.7)
99%, p* = 0.5
99%, p* = 0.3 (or 0.7)

Hypothesis Testing



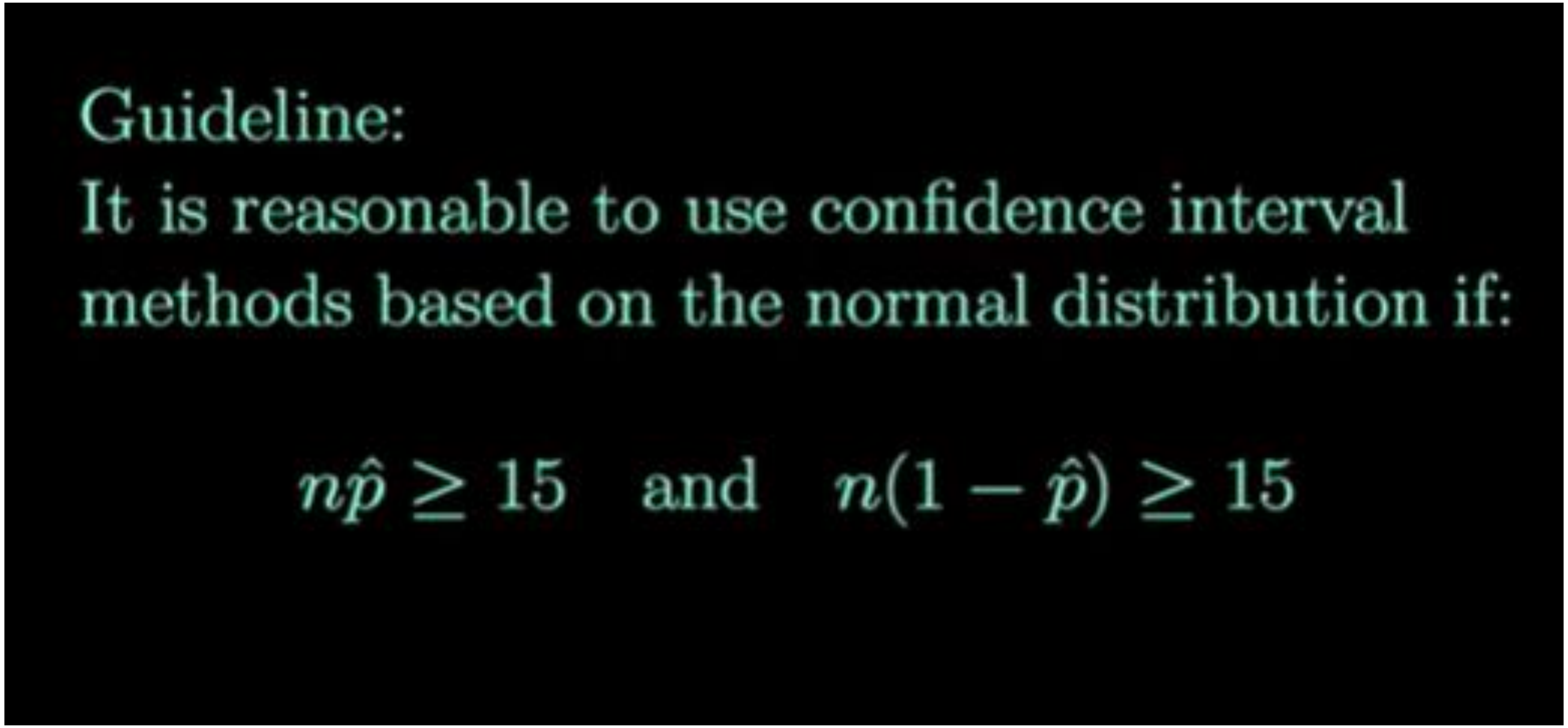
Is there strong evidence that the population proportion differs from a hypothesized value?

We may wish to test H_0: $p = p_0$

A hypothesized value of p

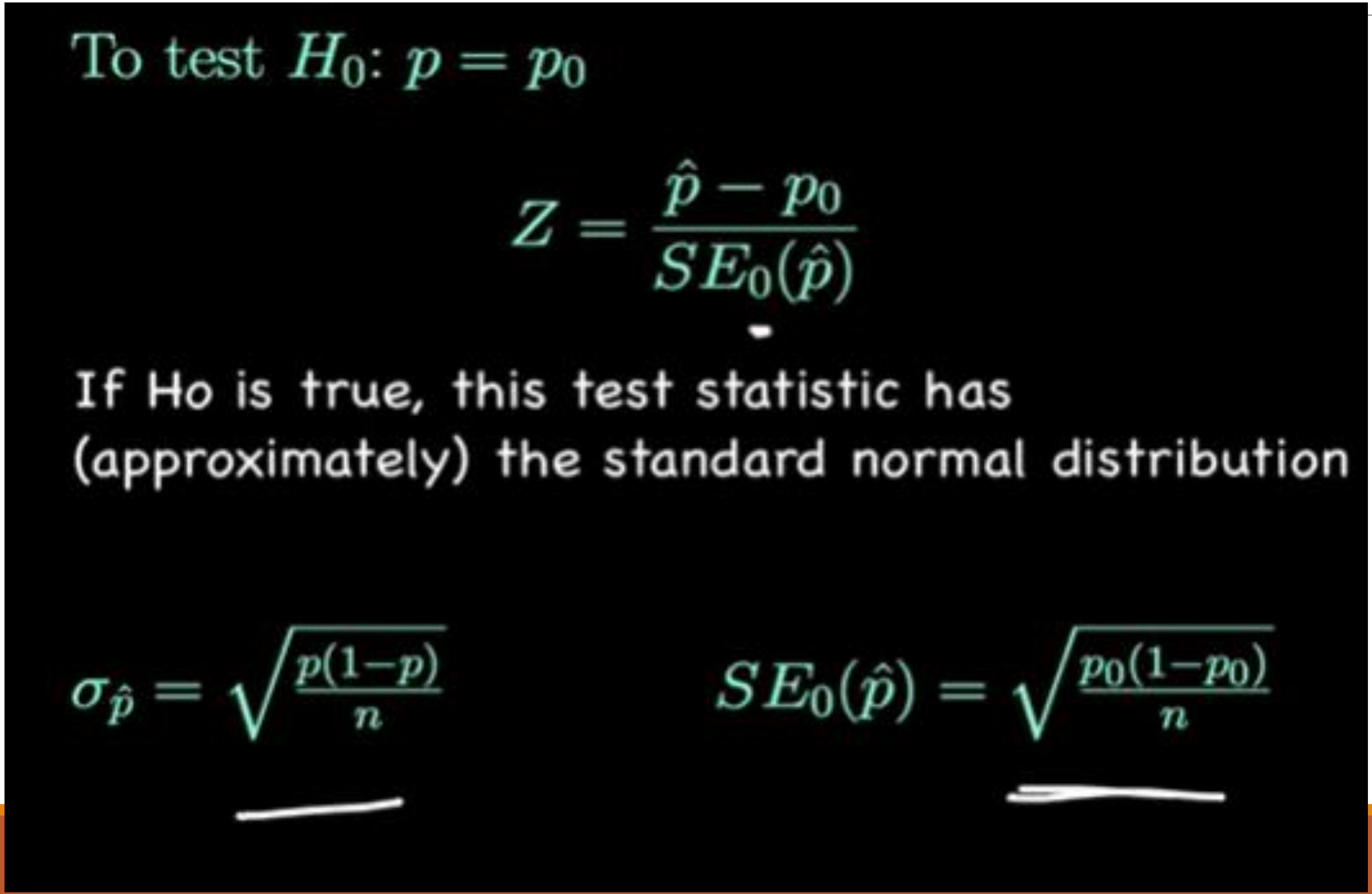
The alternative hypothesis will be one of:

$$H_a\colon p > p_0$$
$$H_a\colon p < p_0$$
$$H_a\colon p \neq p_0$$



Guideline:
It is reasonable to use confidence interval methods based on the normal distribution if:

$$n\hat{p} \geq 15 \quad \text{and} \quad n(1-\hat{p}) \geq 15$$

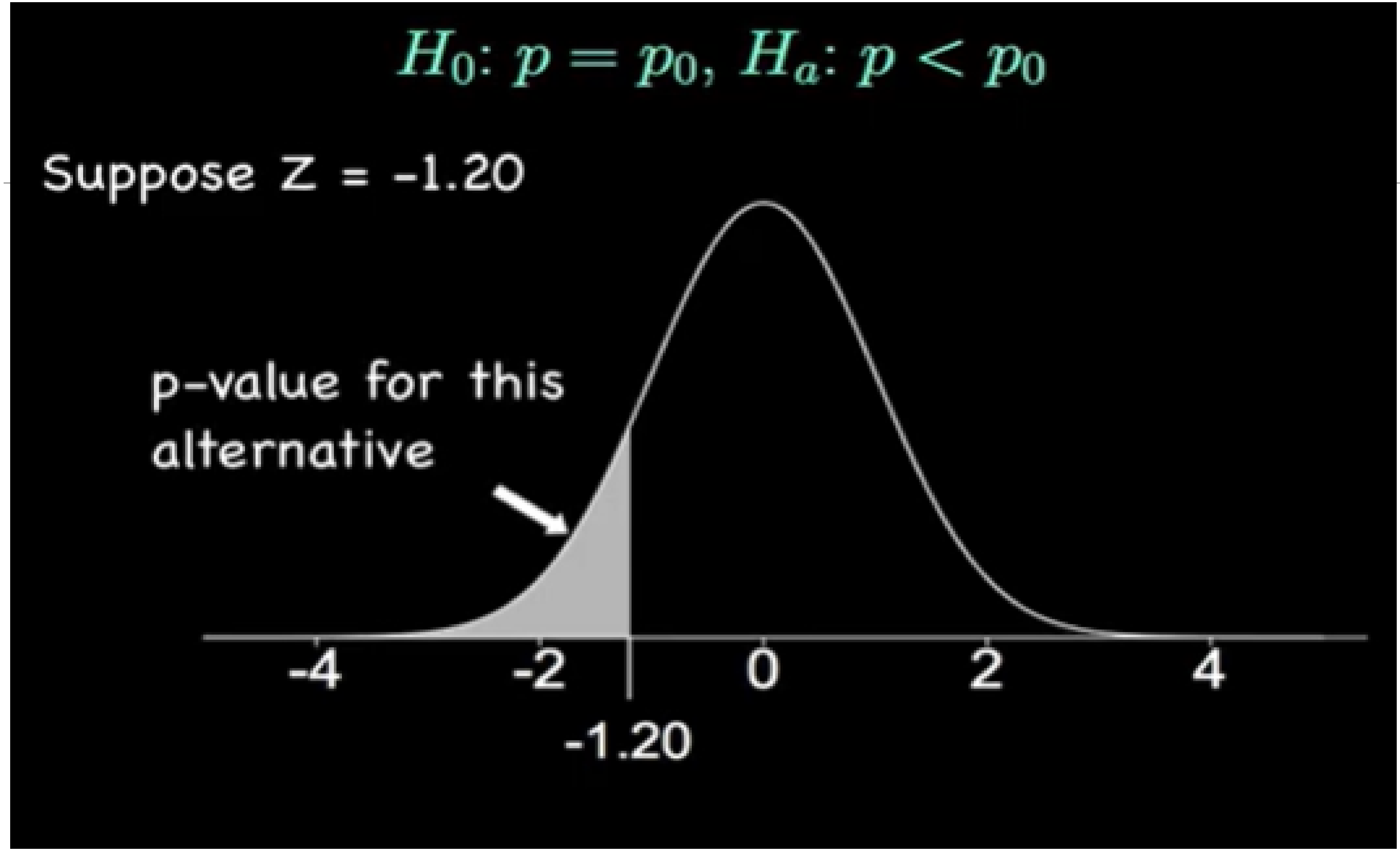


To test H_0: $p = p_0$

$$Z = \frac{\hat{p} - p_0}{SE_0(\hat{p})}$$

If Ho is true, this test statistic has (approximately) the standard normal distribution

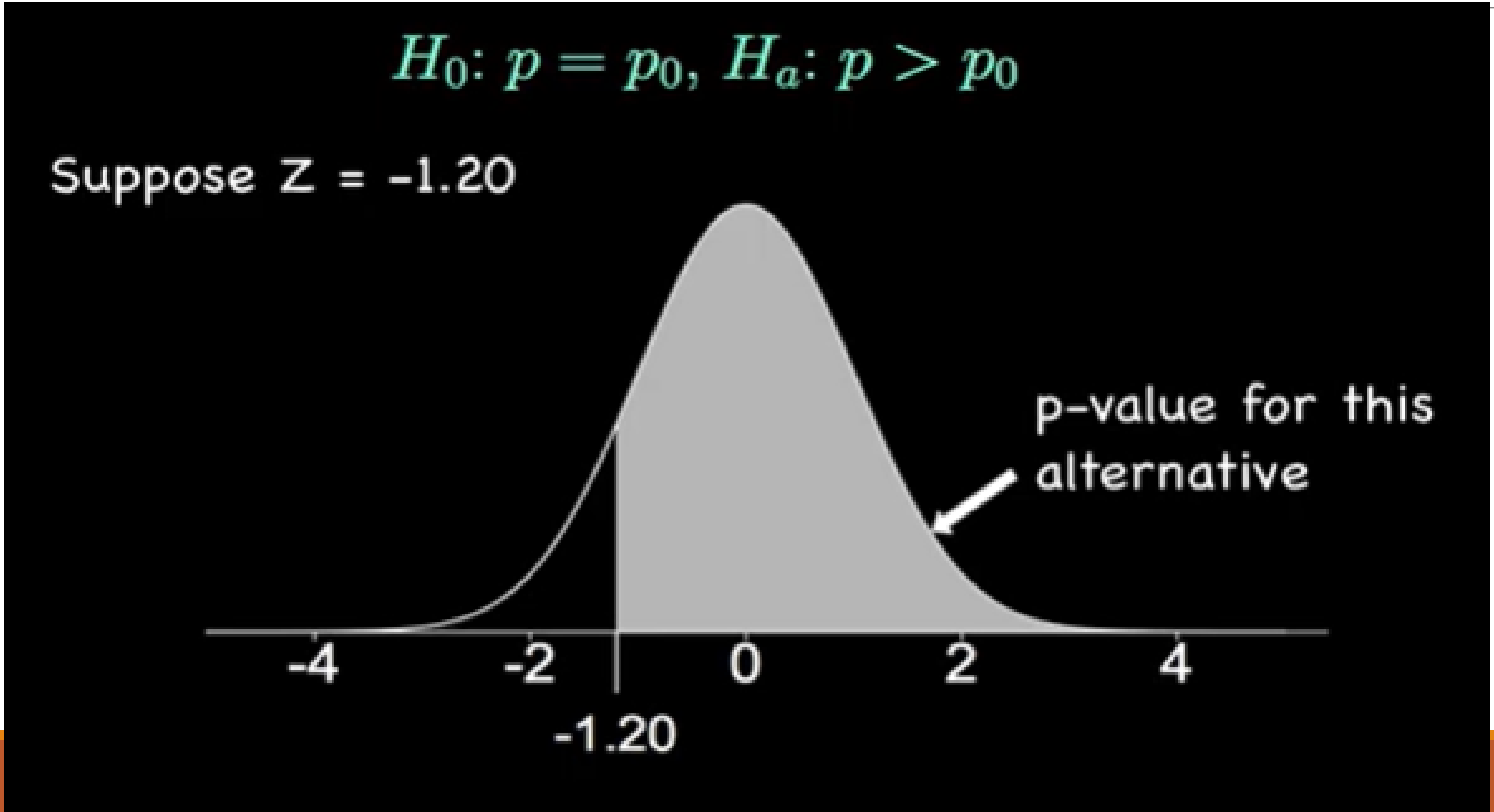
$$\sigma_{\hat{p}} = \underline{\sqrt{\frac{p(1-p)}{n}}} \qquad SE_0(\hat{p}) = \underline{\sqrt{\frac{p_0(1-p_0)}{n}}}$$



$H_0\colon p = p_0,\ H_a\colon p < p_0$

Suppose Z = -1.20

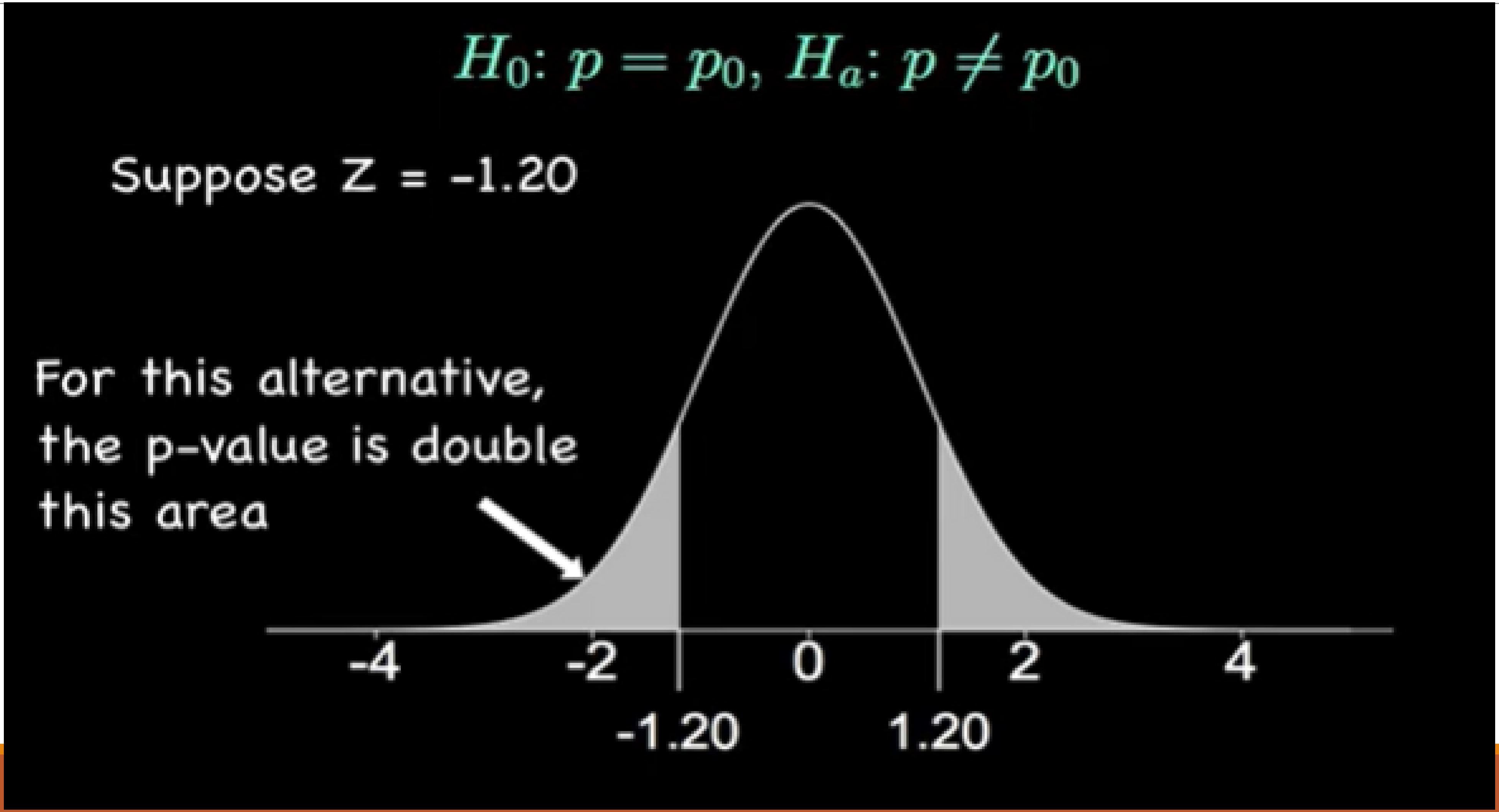
p-value for this alternative



$H_0: p = p_0,\ H_a: p > p_0$

Suppose Z = -1.20

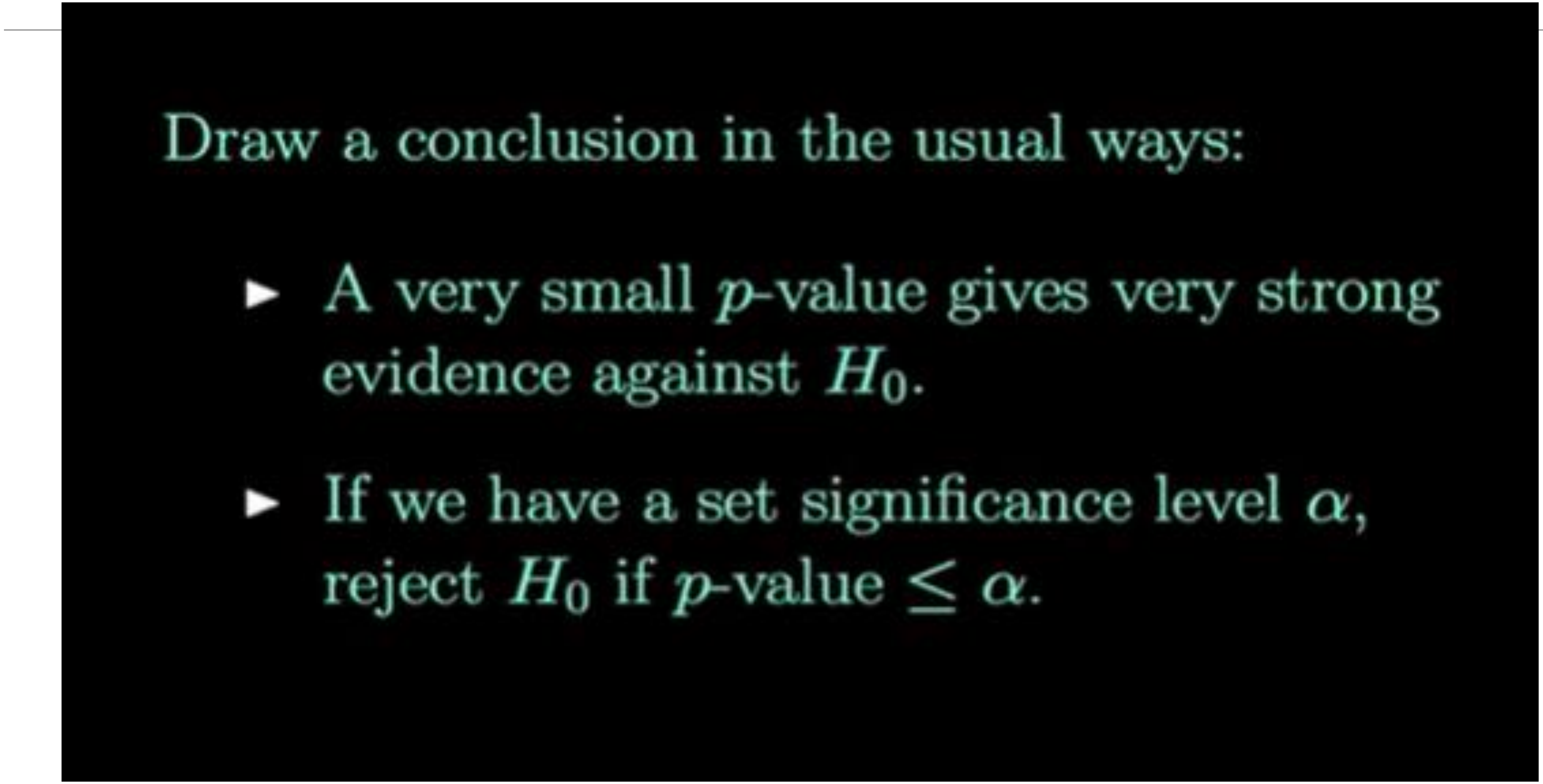
p-value for this alternative



$H_0\colon p = p_0,\ H_a\colon p \neq p_0$

Suppose Z = -1.20

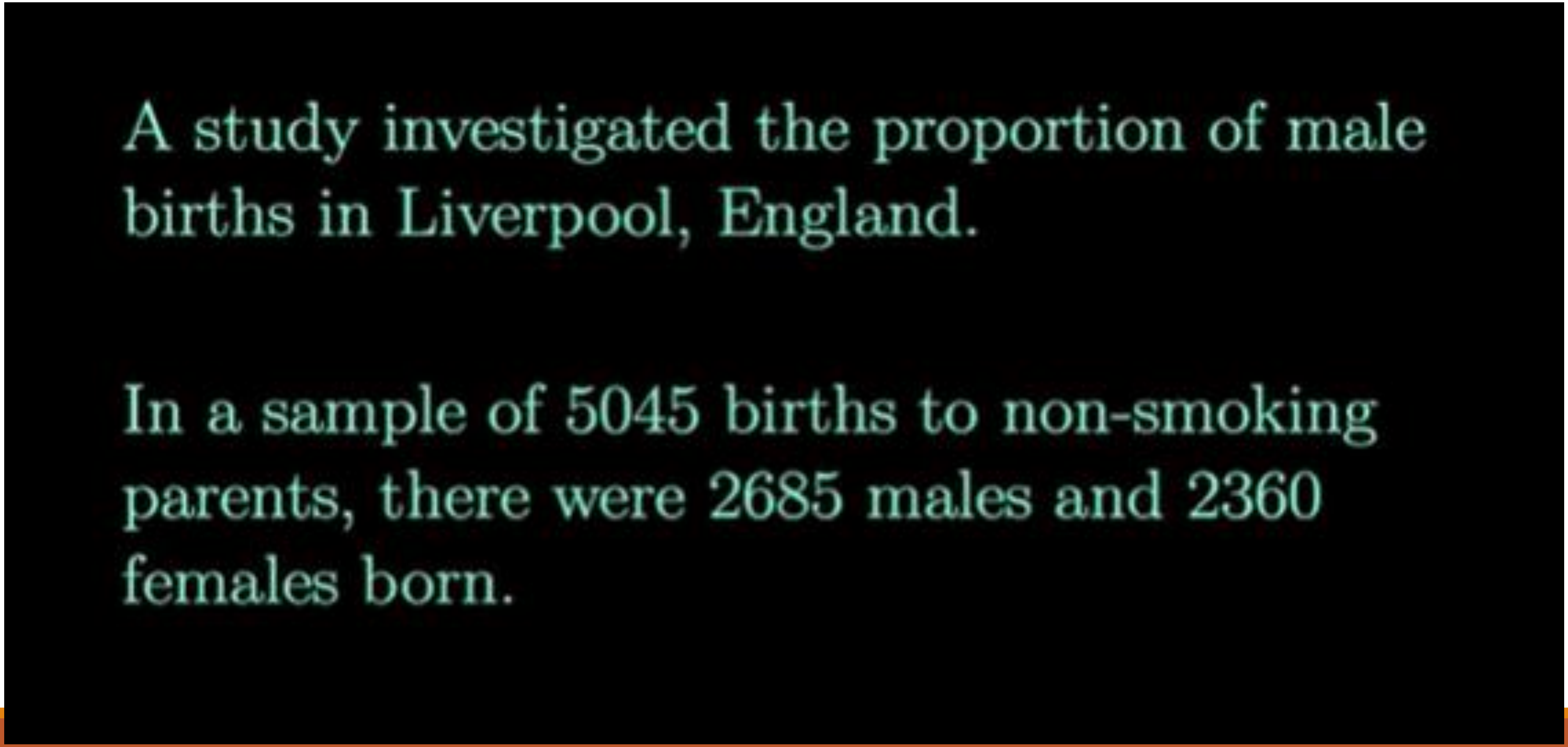
For this alternative, the p-value is double this area



Draw a conclusion in the usual ways:

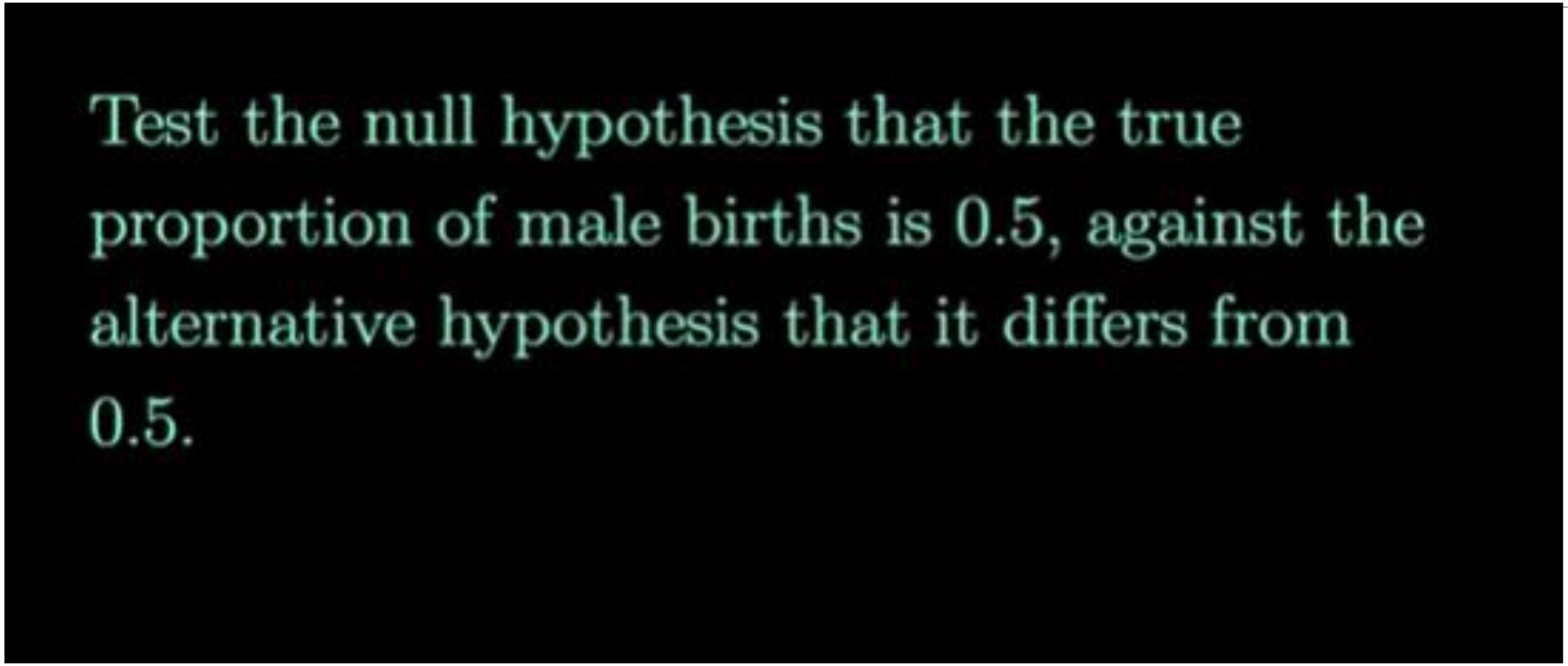
- A very small p-value gives very strong evidence against H_0.
- If we have a set significance level α, reject H_0 if p-value $\leq \alpha$.

Sampling Distribution of Sample Proportion

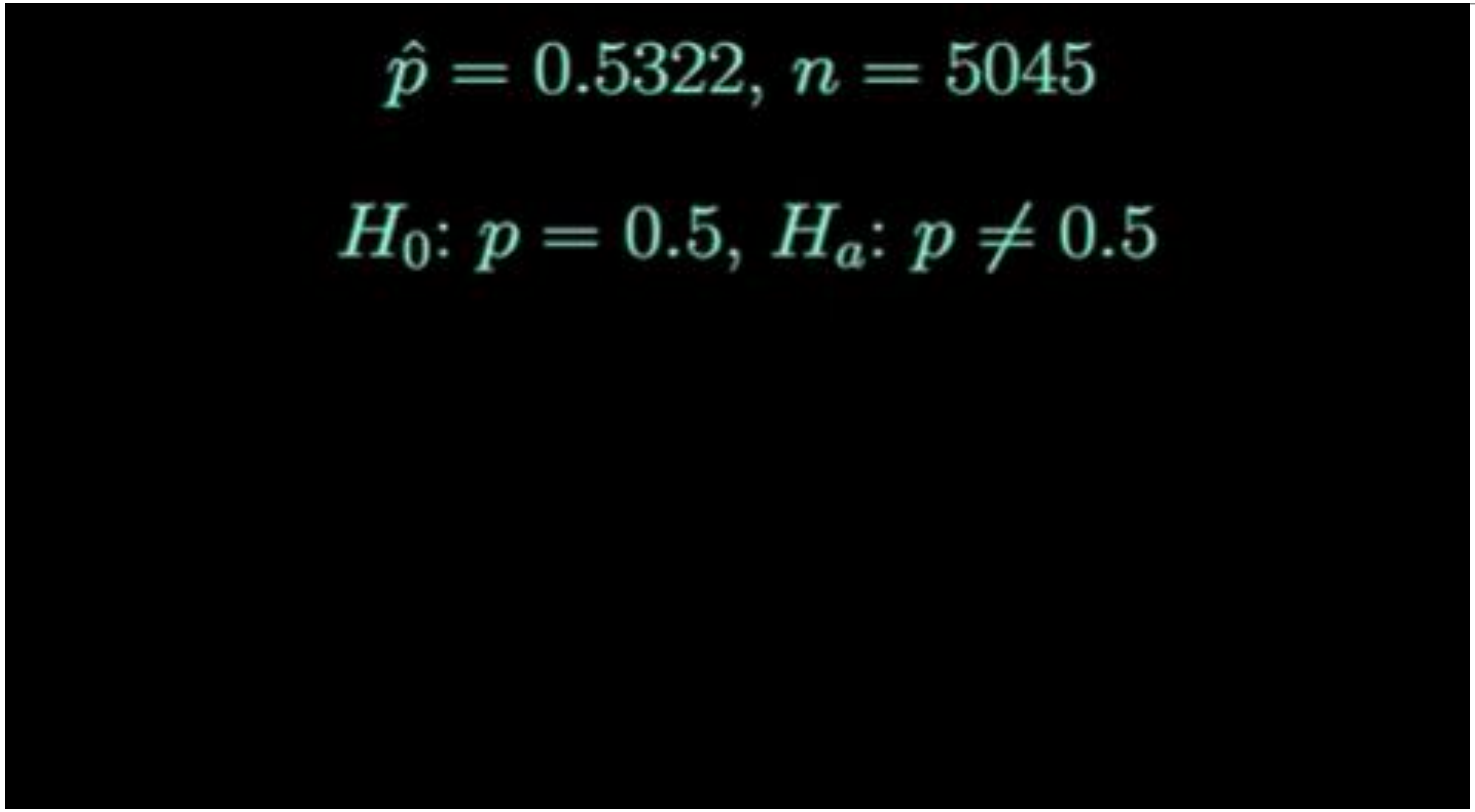


A study investigated the proportion of male births in Liverpool, England.

In a sample of 5045 births to non-smoking parents, there were 2685 males and 2360 females born.

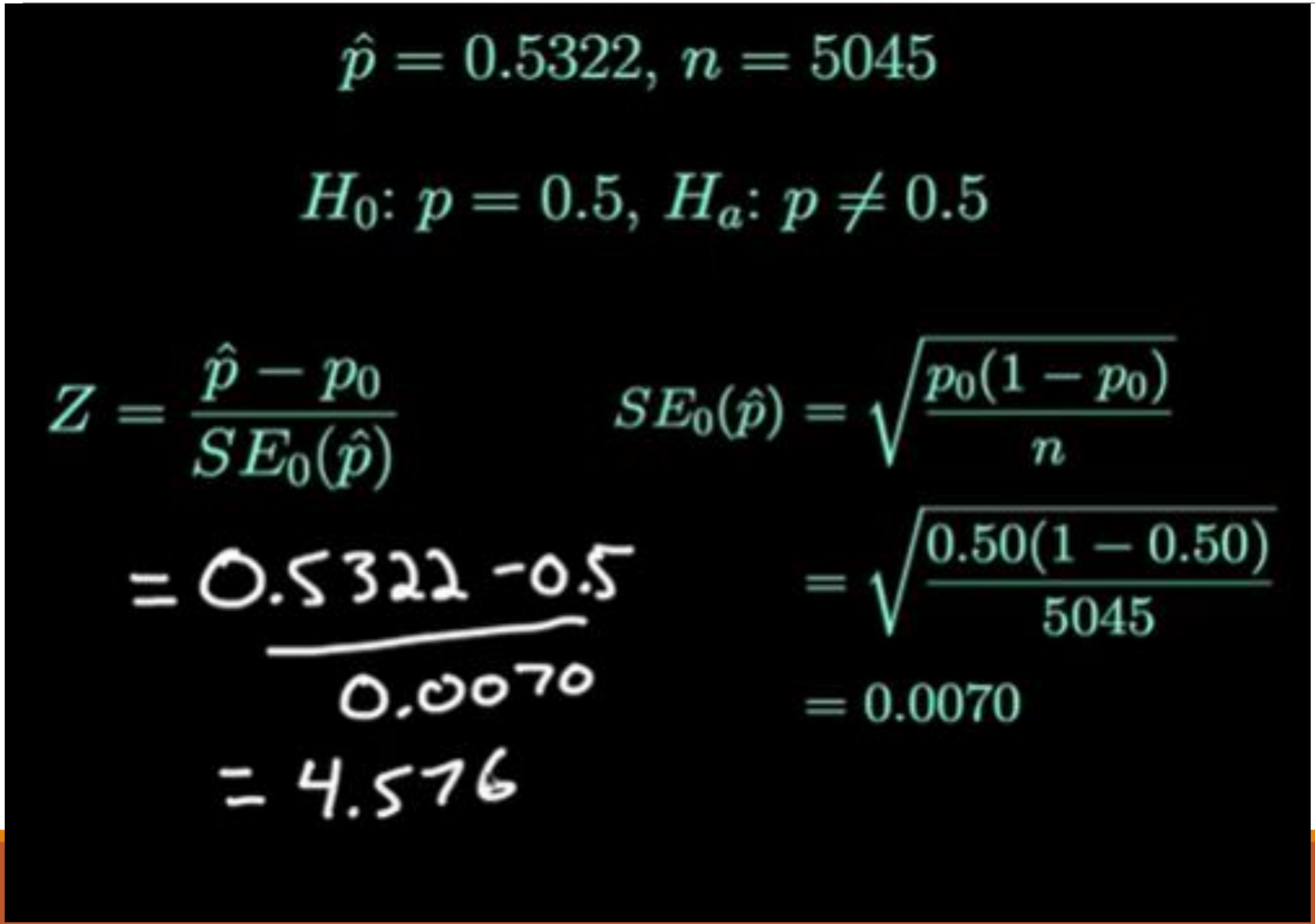


Test the null hypothesis that the true proportion of male births is 0.5, against the alternative hypothesis that it differs from 0.5.



$$\hat{p} = 0.5322,\ n = 5045$$

$$H_0\colon p = 0.5,\ H_a\colon p \neq 0.5$$



$$\hat{p} = 0.5322,\ n = 5045$$

$$H_0\colon p = 0.5,\ H_a\colon p \neq 0.5$$

$$Z = \frac{\hat{p} - p_0}{SE_0(\hat{p})}$$

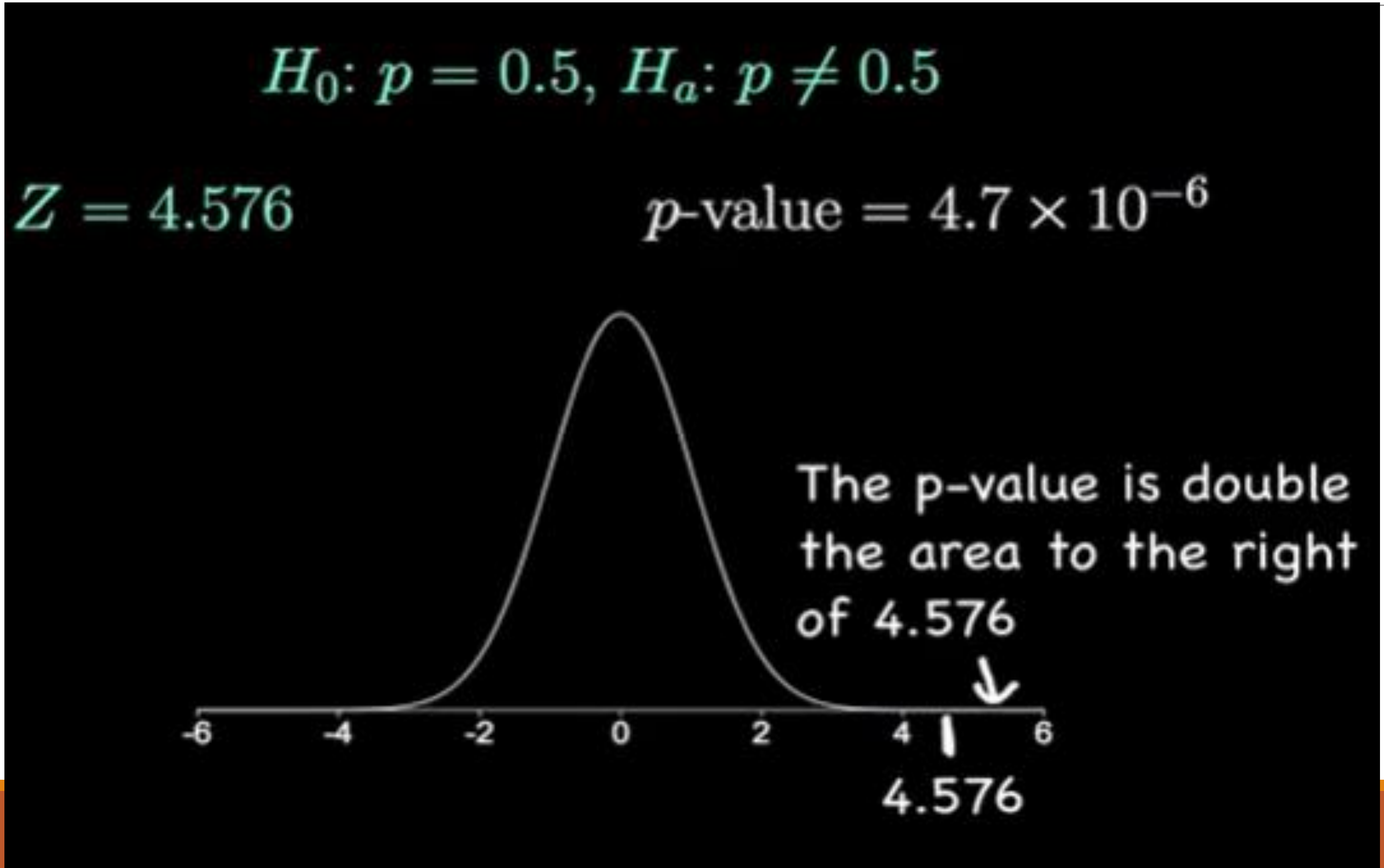
$$= \frac{0.5322 - 0.5}{0.0070}$$

$$= 4.576$$

$$SE_0(\hat{p}) = \sqrt{\frac{p_0(1 - p_0)}{n}}$$

$$= \sqrt{\frac{0.50(1 - 0.50)}{5045}}$$

$$= 0.0070$$

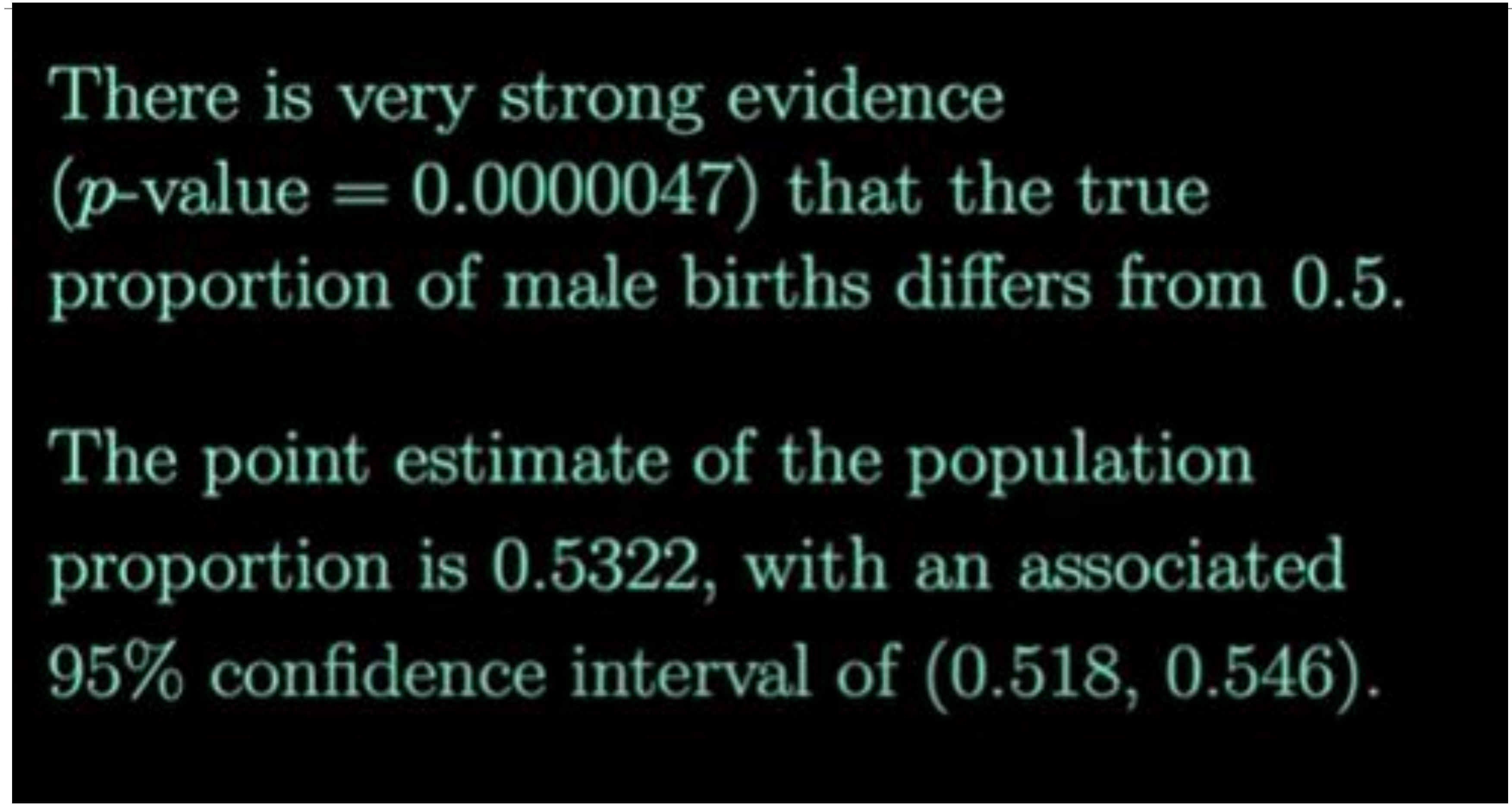


H_0: $p = 0.5$, H_a: $p \neq 0.5$

$Z = 4.576$

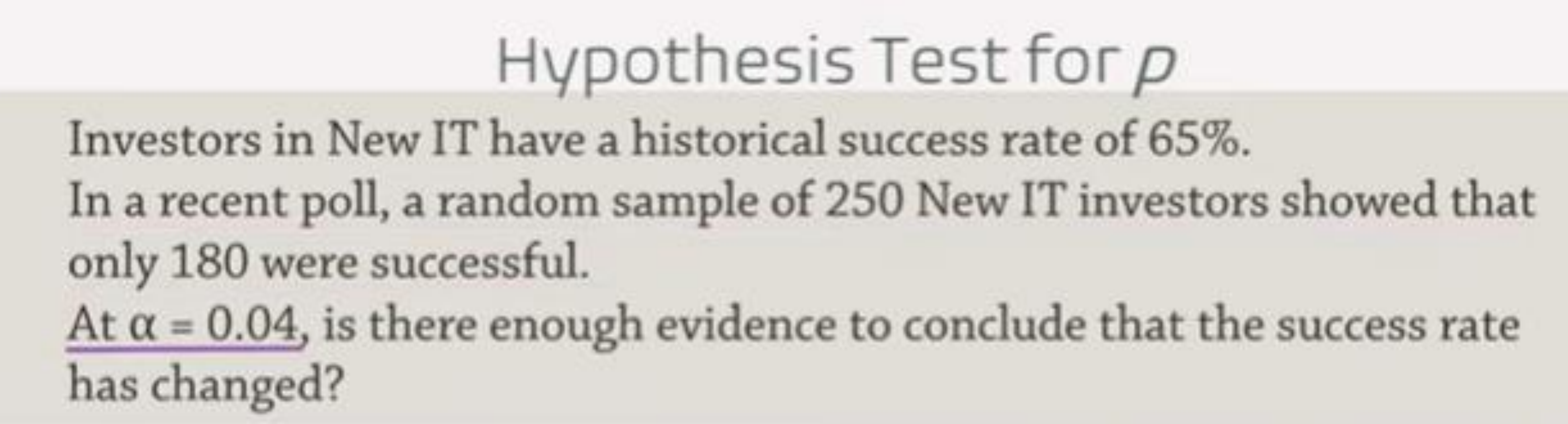
p-value $= 4.7 \times 10^{-6}$

The p-value is double the area to the right of 4.576



There is very strong evidence (p-value = 0.0000047) that the true proportion of male births differs from 0.5.

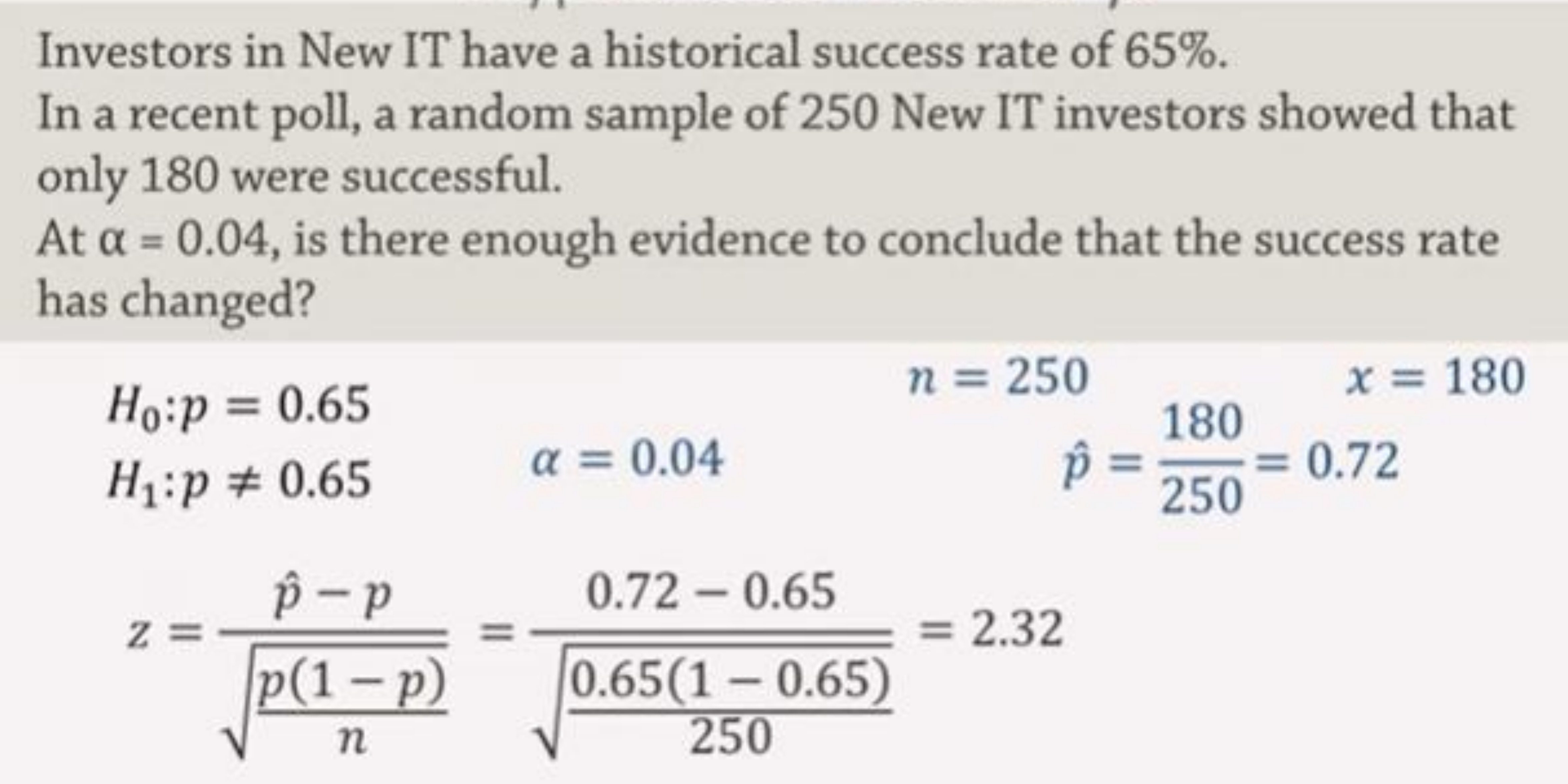
The point estimate of the population proportion is 0.5322, with an associated 95% confidence interval of (0.518, 0.546).



Hypothesis Test for *p*

Investors in New IT have a historical success rate of 65%.
In a recent poll, a random sample of 250 New IT investors showed that only 180 were successful.
At $\alpha = 0.04$, is there enough evidence to conclude that the success rate has changed?

Hypothesis Test for *p*



Investors in New IT have a historical success rate of 65%.
In a recent poll, a random sample of 250 New IT investors showed that only 180 were successful.
At α = 0.04, is there enough evidence to conclude that the success rate has changed?

$H_0 : p = 0.65$

$H_1 : p \neq 0.65$

$\alpha = 0.04$

$n = 250$ $\quad x = 180$

$$\hat{p} = \frac{180}{250} = 0.72$$

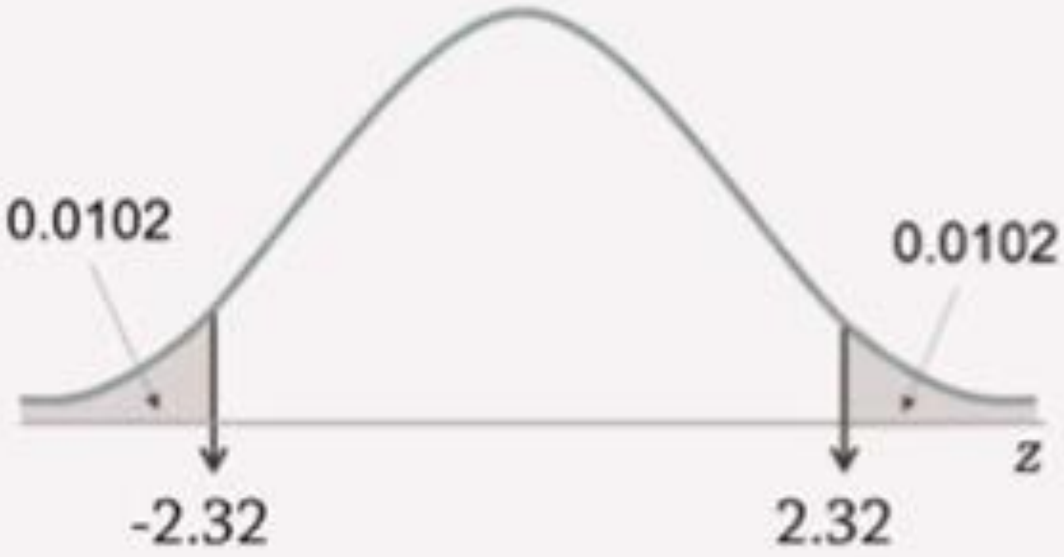
$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.72 - 0.65}{\sqrt{\frac{0.65(1-0.65)}{250}}} = 2.32$$

$H_0: p = 0.65$

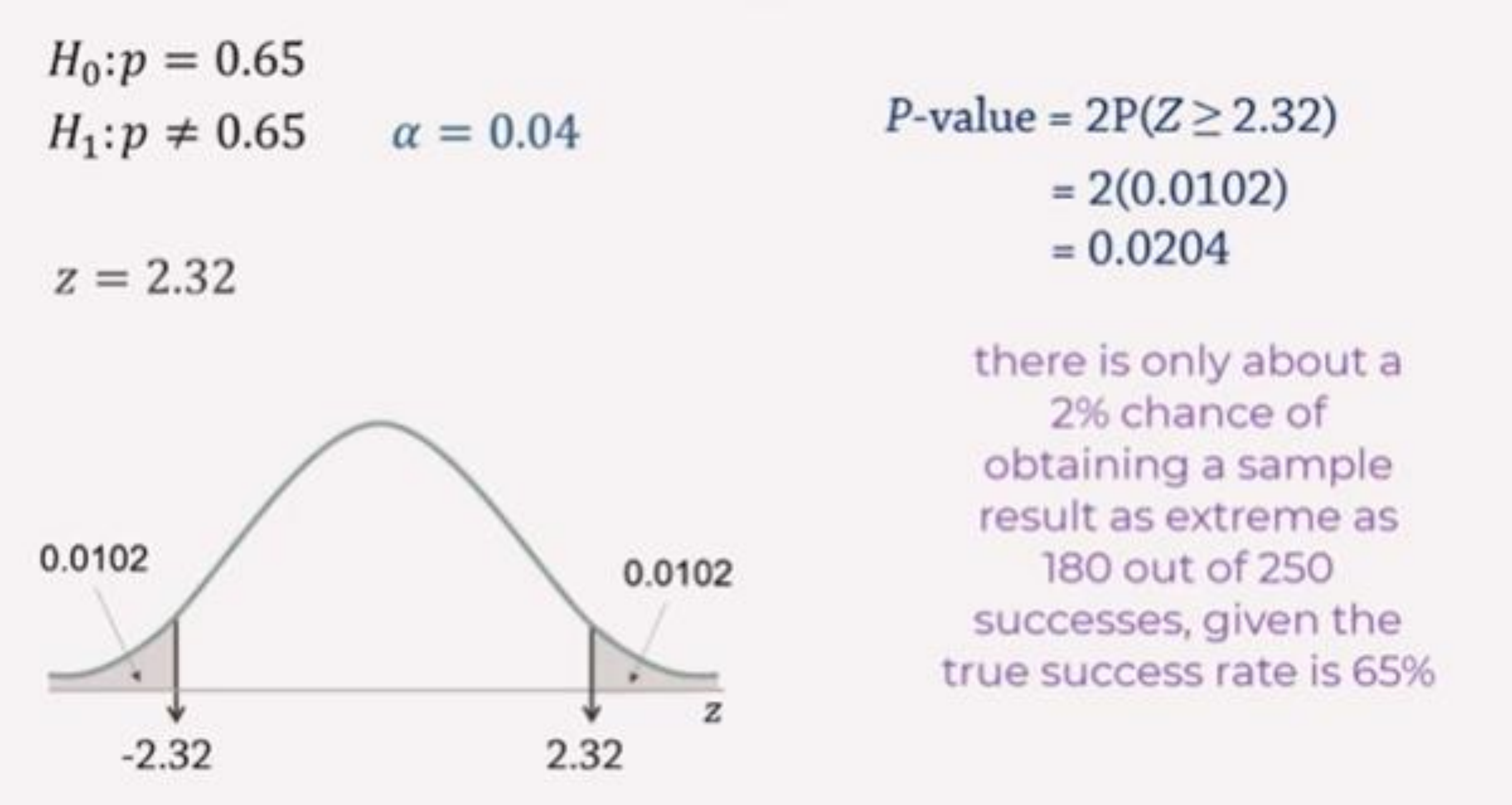
$H_1: p \neq 0.65$ $\alpha = 0.04$

$z = 2.32$

z	0.00	0.01	0.02	0.03	0.04	0.05
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960
...	...	...	...	...	...	...



$$\begin{aligned} P\text{-value} &= 2P(Z \geq 2.32) \\ &= 2(0.0102) \\ &= \boxed{0.0204} \end{aligned}$$



$H_0 : p = 0.65$

$H_1 : p \neq 0.65$ $\quad \alpha = 0.04$

$z = 2.32$

0.0102
0.0102
z
-2.32
2.32

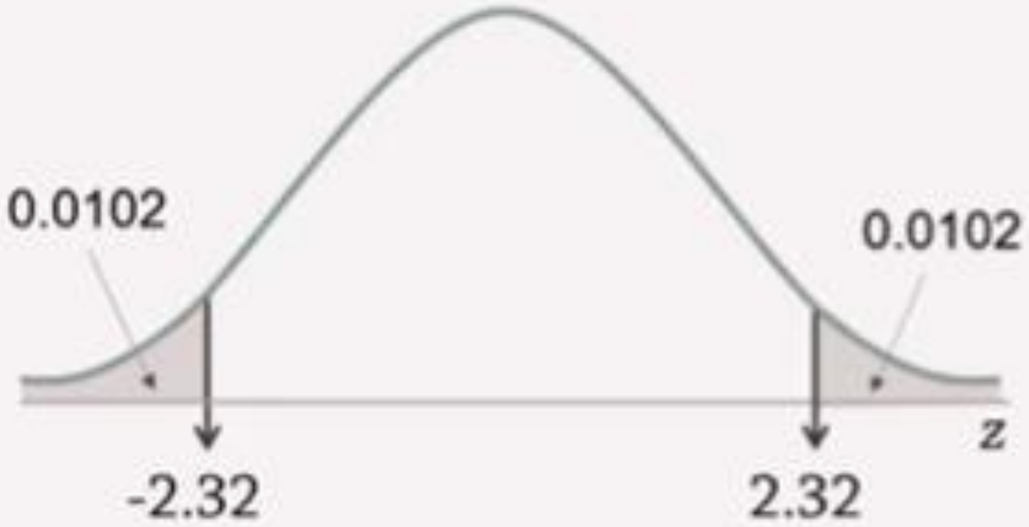
$$P\text{-value} = 2P(Z \geq 2.32)$$
$$= 2(0.0102)$$
$$= 0.0204$$

there is only about a 2% chance of obtaining a sample result as extreme as 180 out of 250 successes, given the true success rate is 65%

$H_0 : p = 0.65$

$H_1 : p \neq 0.65 \qquad \alpha = 0.04$

$z = 2.32$



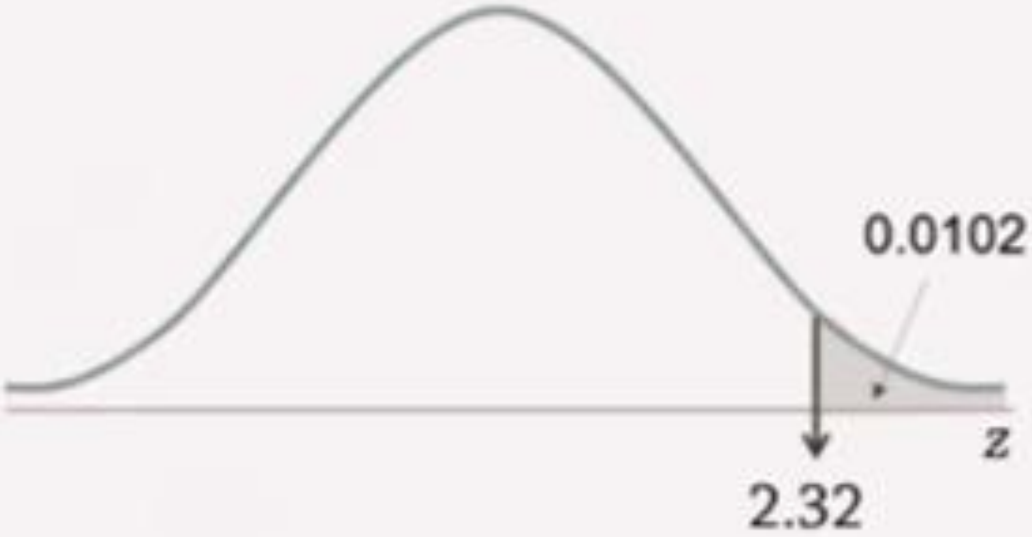
Since P-value = 0.0204 < α

Reject H_0

There is enough evidence to conclude that the population proportion differs from 0.65
or that the success rate has changed.

Is there enough evidence to conclude that the success rate has increased from 65%?

$z = 2.32$



$H_1 : p > 0.65$

Is there enough evidence to conclude that the success rate has increased from 65%?

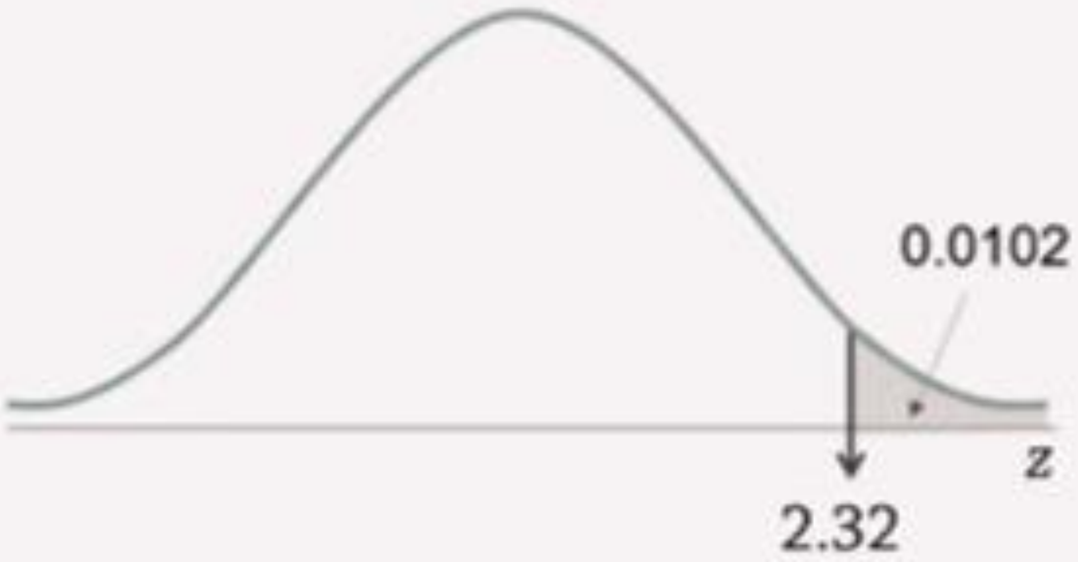
$z = 2.32$

$\alpha = 0.04$

P-value = 0.0102 < α

Reject H_0

There is sufficient evidence to conclude that the success rate has increased.



$H_1 : p > 0.65$

Is there enough evidence to conclude that the success rate has increased from 65%?

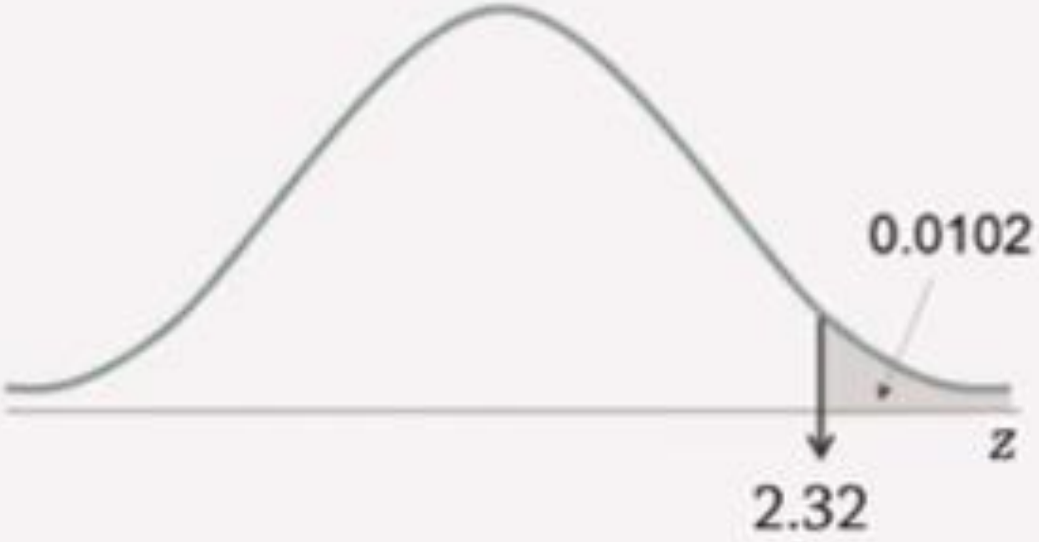
$z = 2.32$

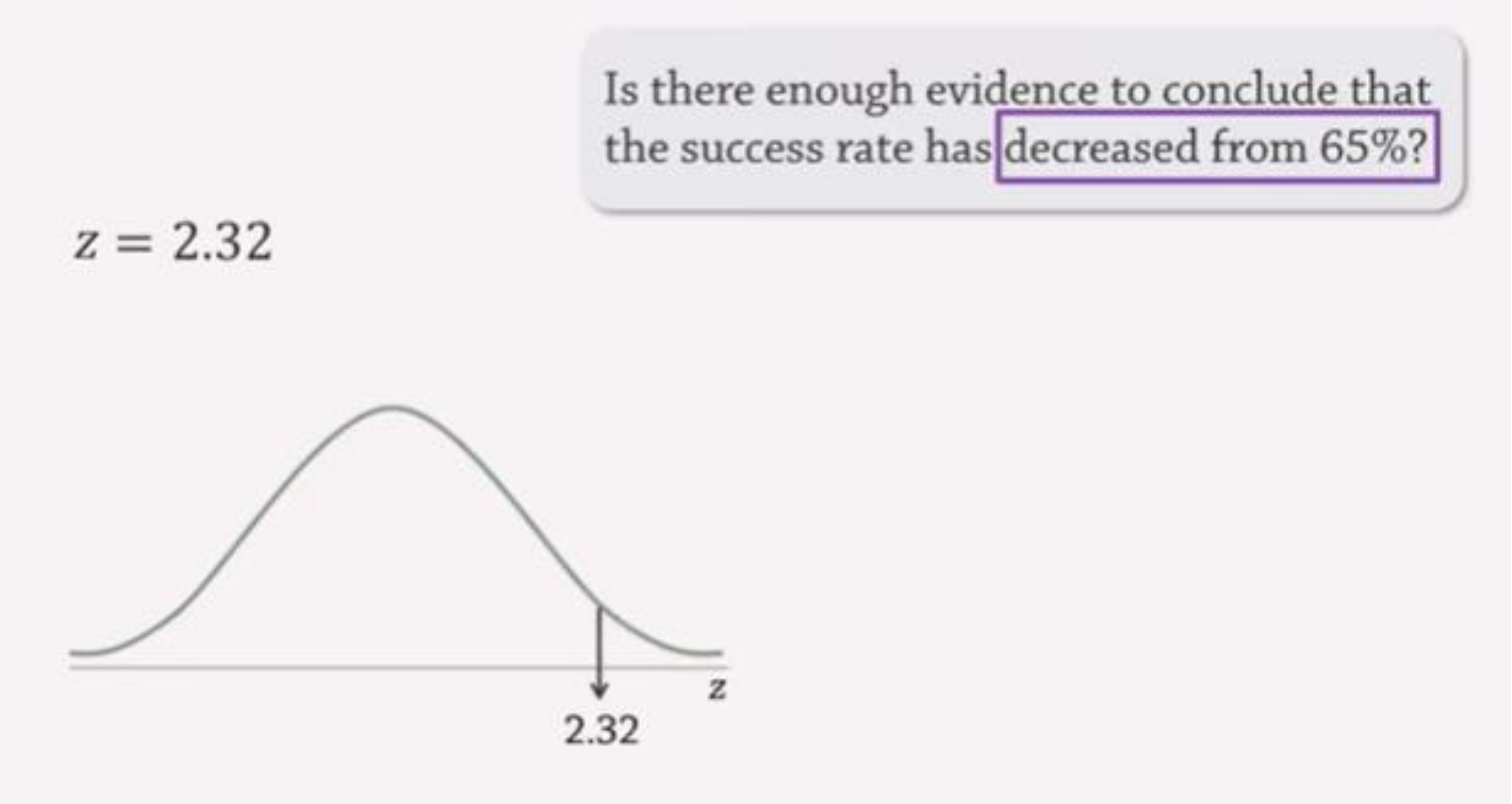
$\alpha = 0.01$

P-value = 0.0102 > α

Fail to Reject H_0

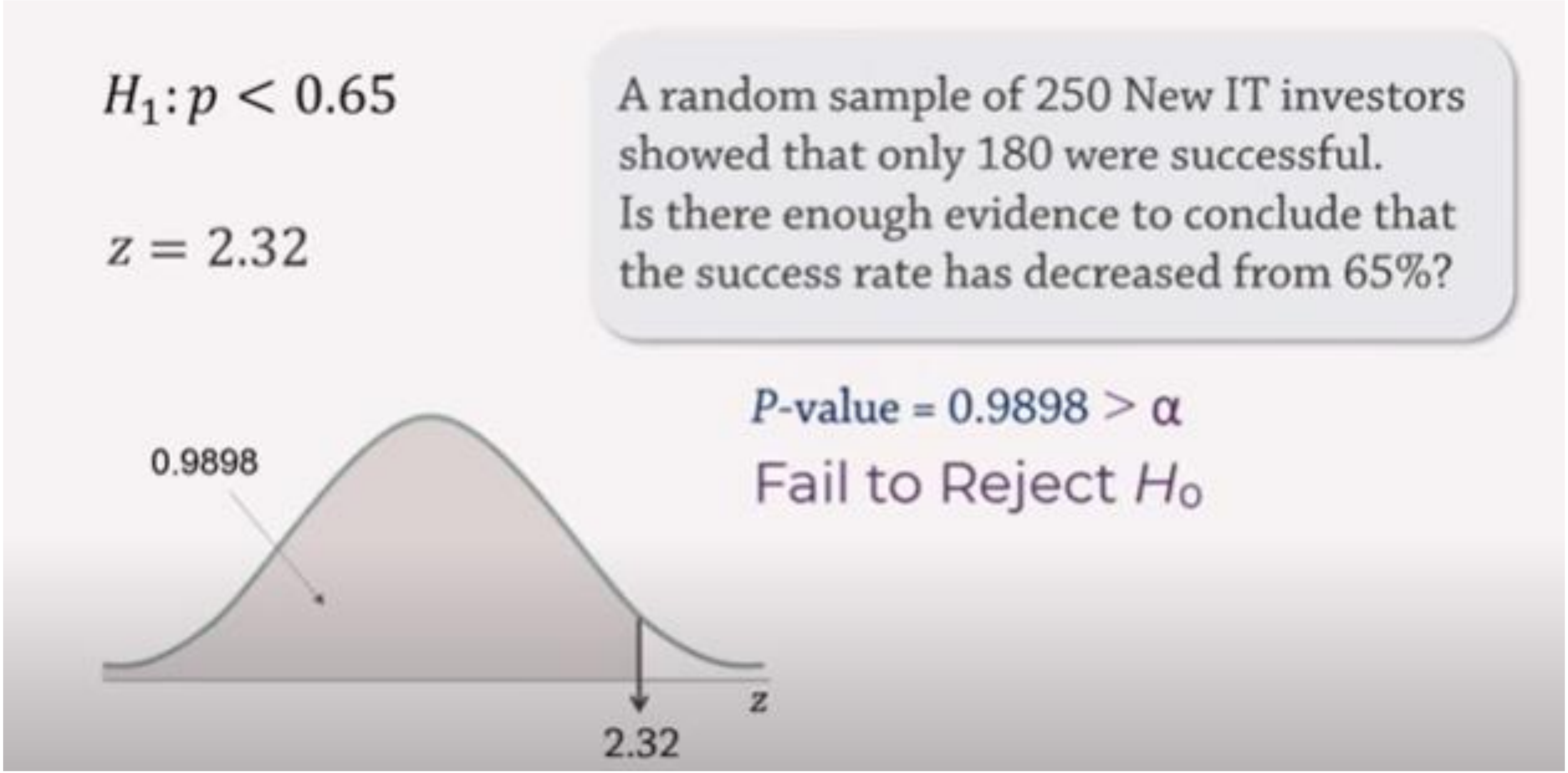
There is insufficient evidence to conclude that the success rate has increased.





Is there enough evidence to conclude that the success rate has decreased from 65%?

$z = 2.32$



$H_1\!: p < 0.65$

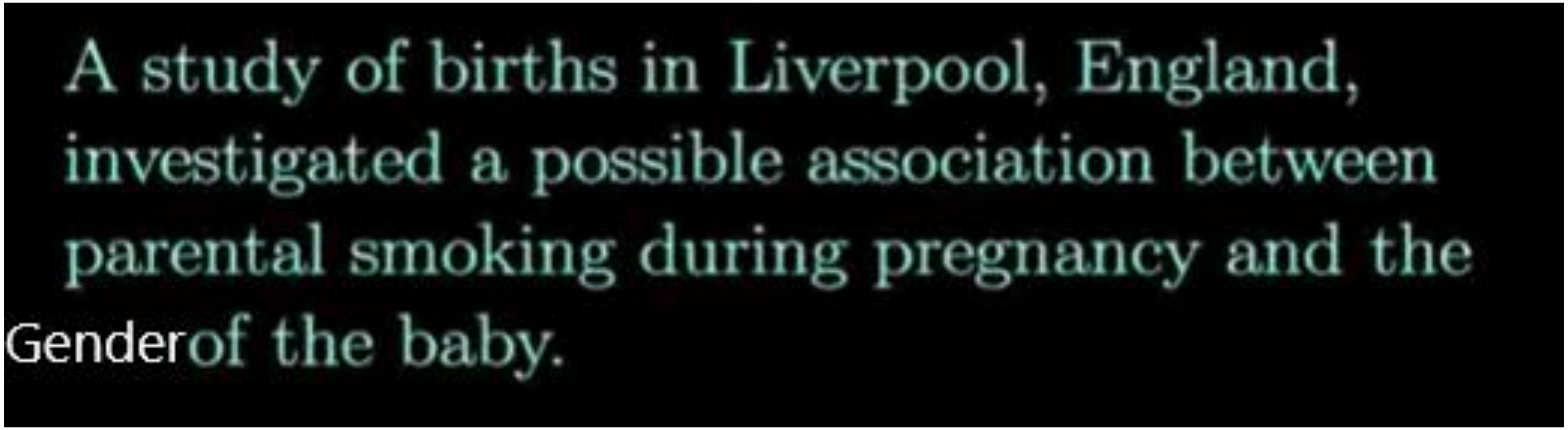
$z = 2.32$

A random sample of 250 New IT investors showed that only 180 were successful. Is there enough evidence to conclude that the success rate has decreased from 65%?

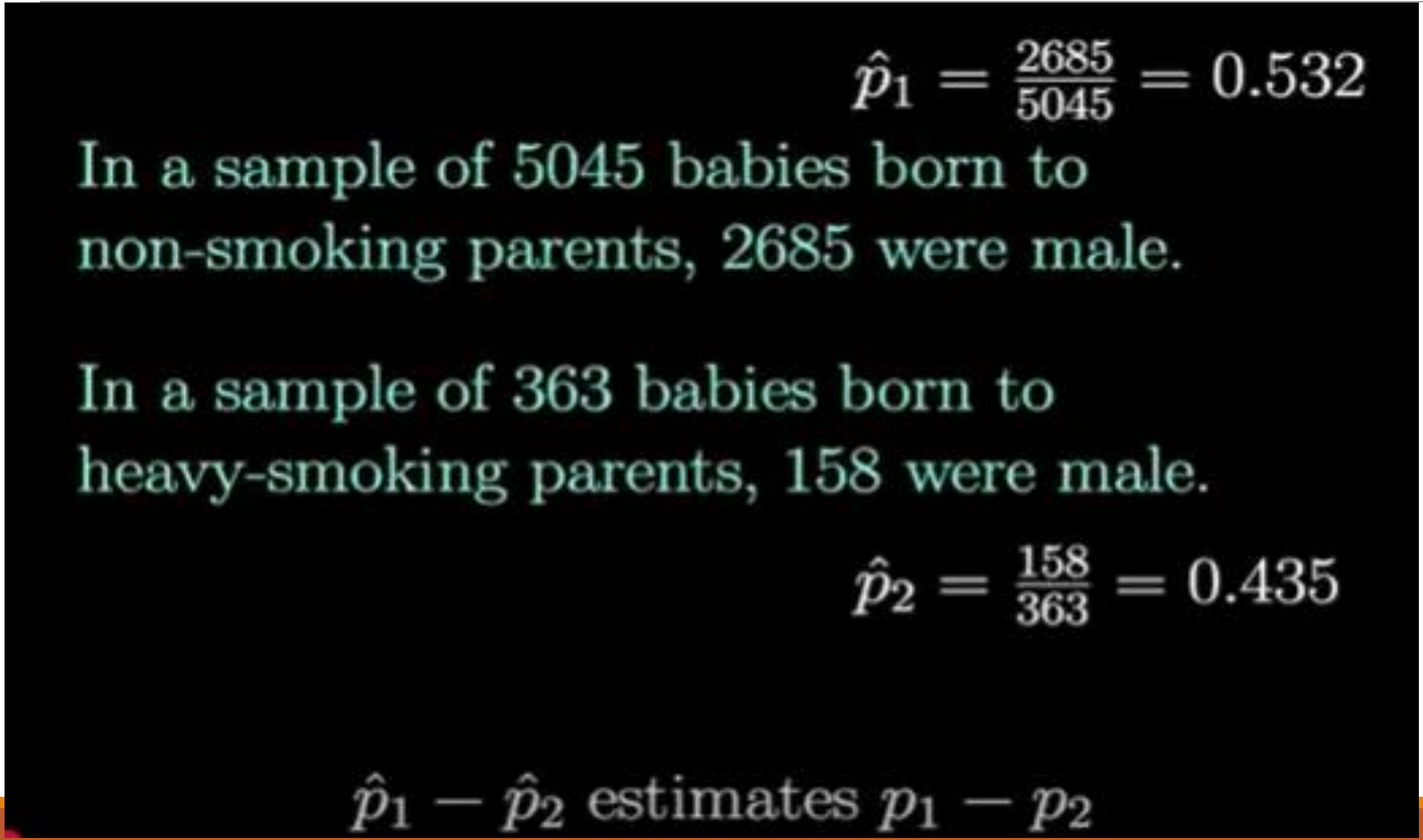
P-value = 0.9898 > α

Fail to Reject H_0

Hypothesis Test for Comparing Two Proportions



A study of births in Liverpool, England, investigated a possible association between parental smoking during pregnancy and the Genderof the baby.



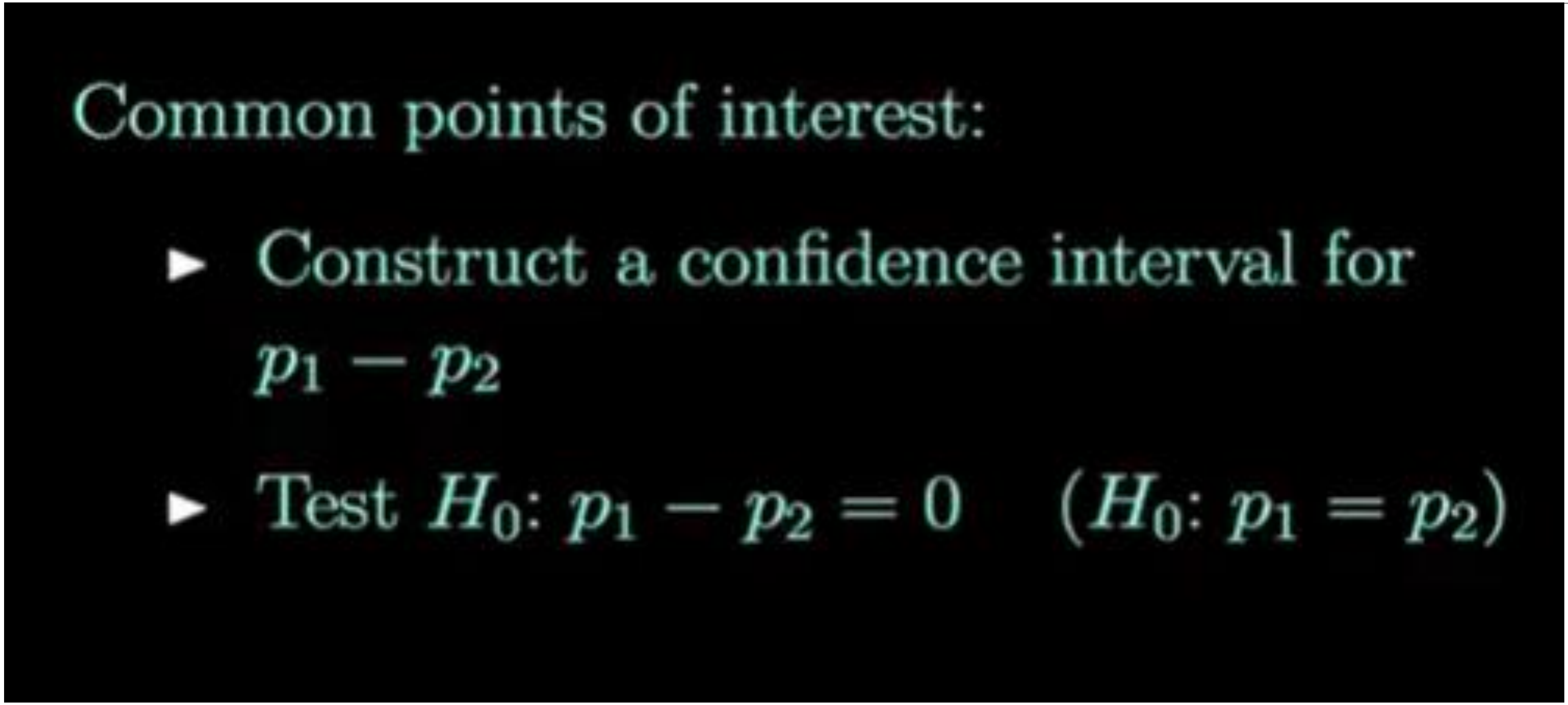
$$\hat{p}_1 = \frac{2685}{5045} = 0.532$$

In a sample of 5045 babies born to non-smoking parents, 2685 were male.

In a sample of 363 babies born to heavy-smoking parents, 158 were male.

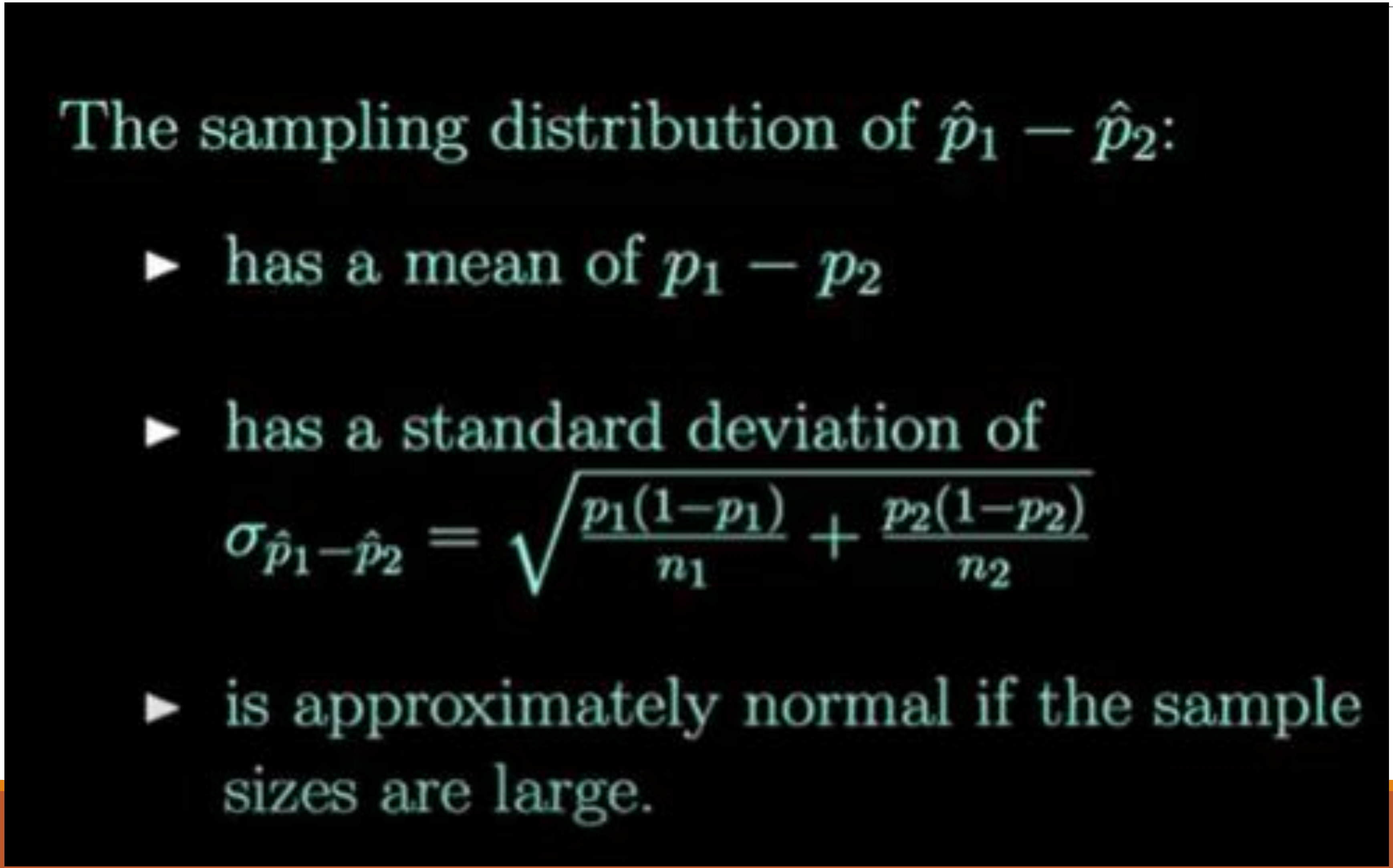
$$\hat{p}_2 = \frac{158}{363} = 0.435$$

$\hat{p}_1 - \hat{p}_2$ estimates $p_1 - p_2$



Common points of interest:

- Construct a confidence interval for $p_1 - p_2$
- Test H_0: $p_1 - p_2 = 0$ $\quad$ (H_0: $p_1 = p_2$)

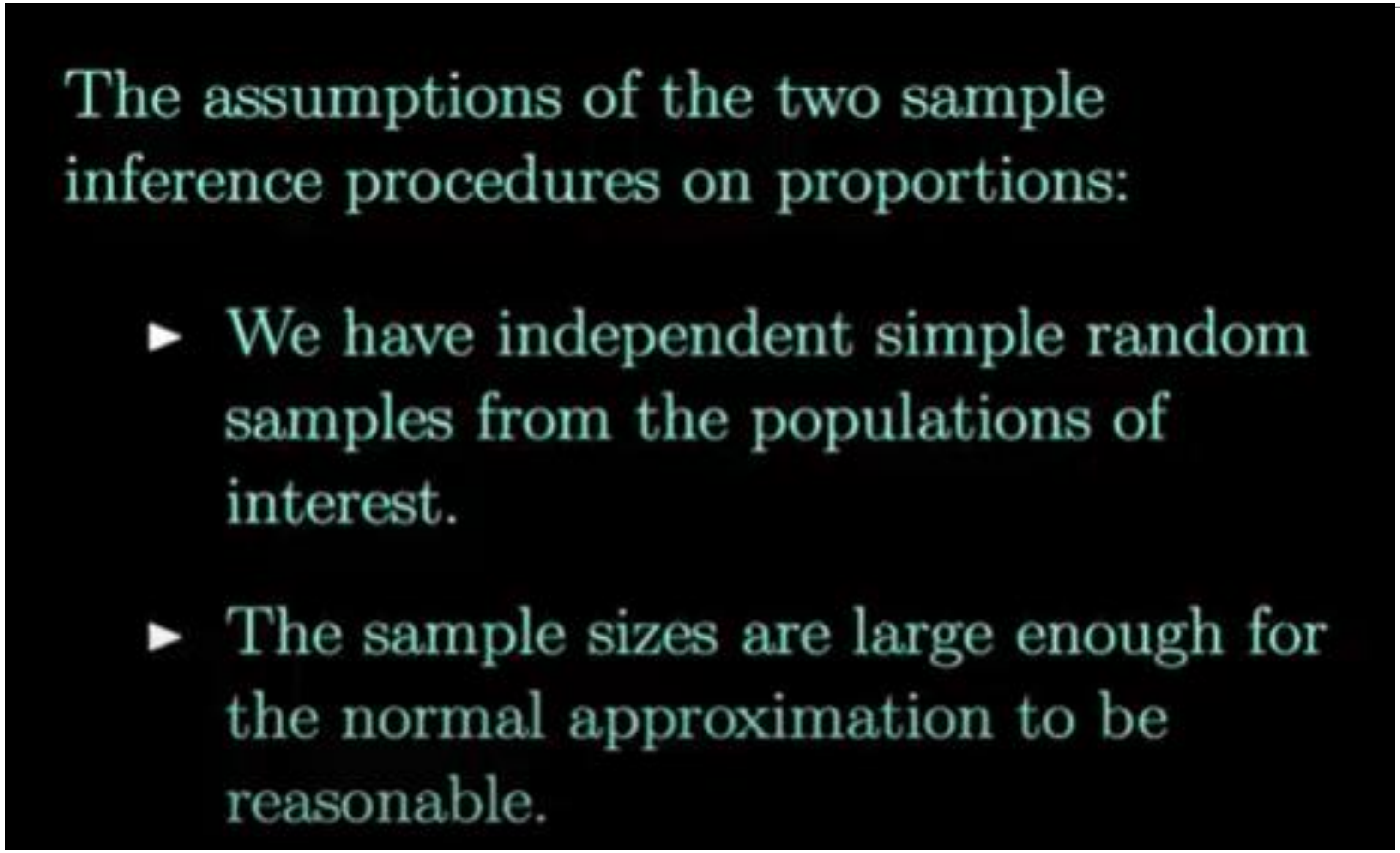


The sampling distribution of $\hat{p}_1 - \hat{p}_2$:

- has a mean of $p_1 - p_2$
- has a standard deviation of

$$\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

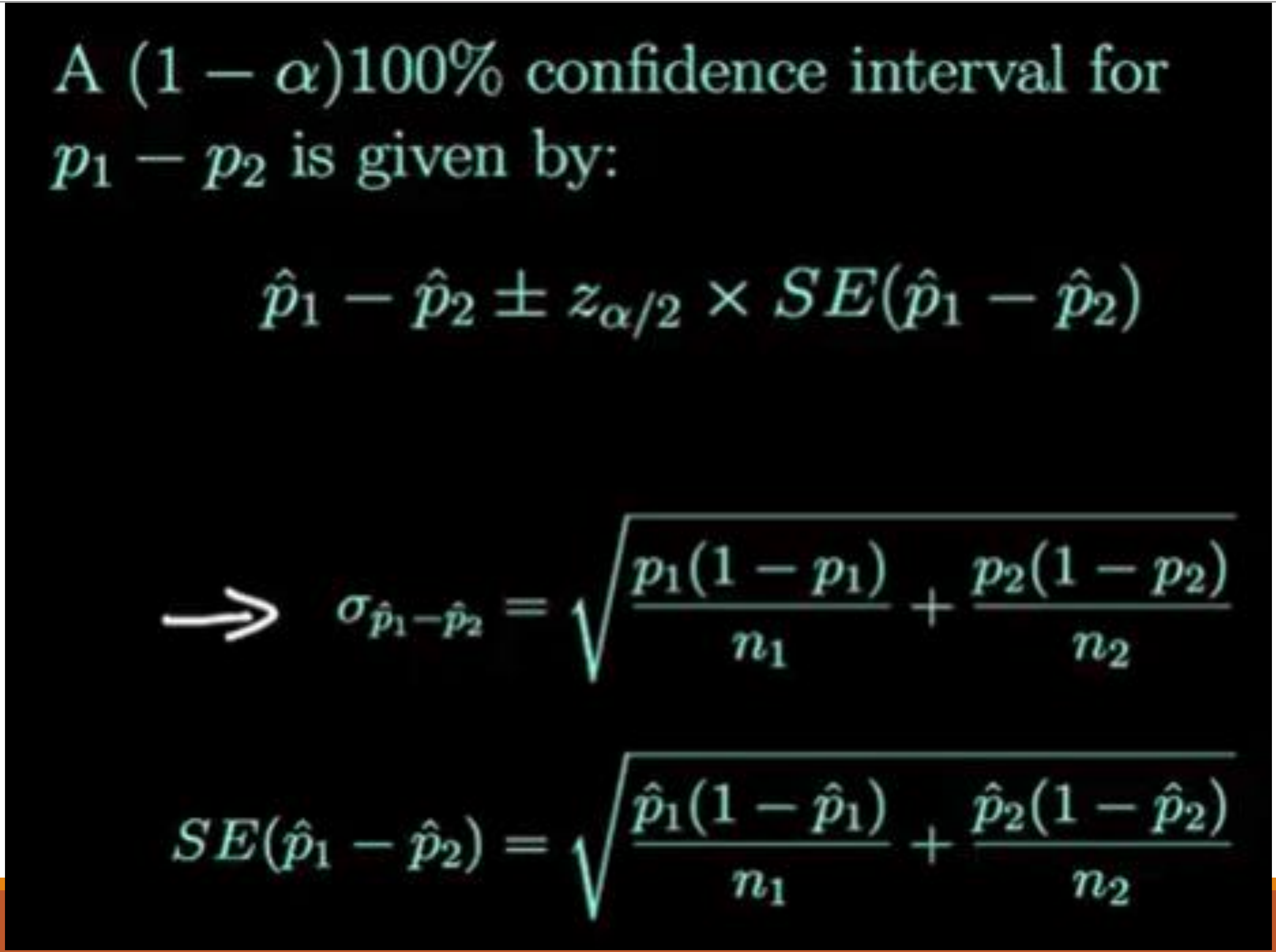
- is approximately normal if the sample sizes are large.



The assumptions of the two sample inference procedures on proportions:

- We have independent simple random samples from the populations of interest.
- The sample sizes are large enough for the normal approximation to be reasonable.

CI



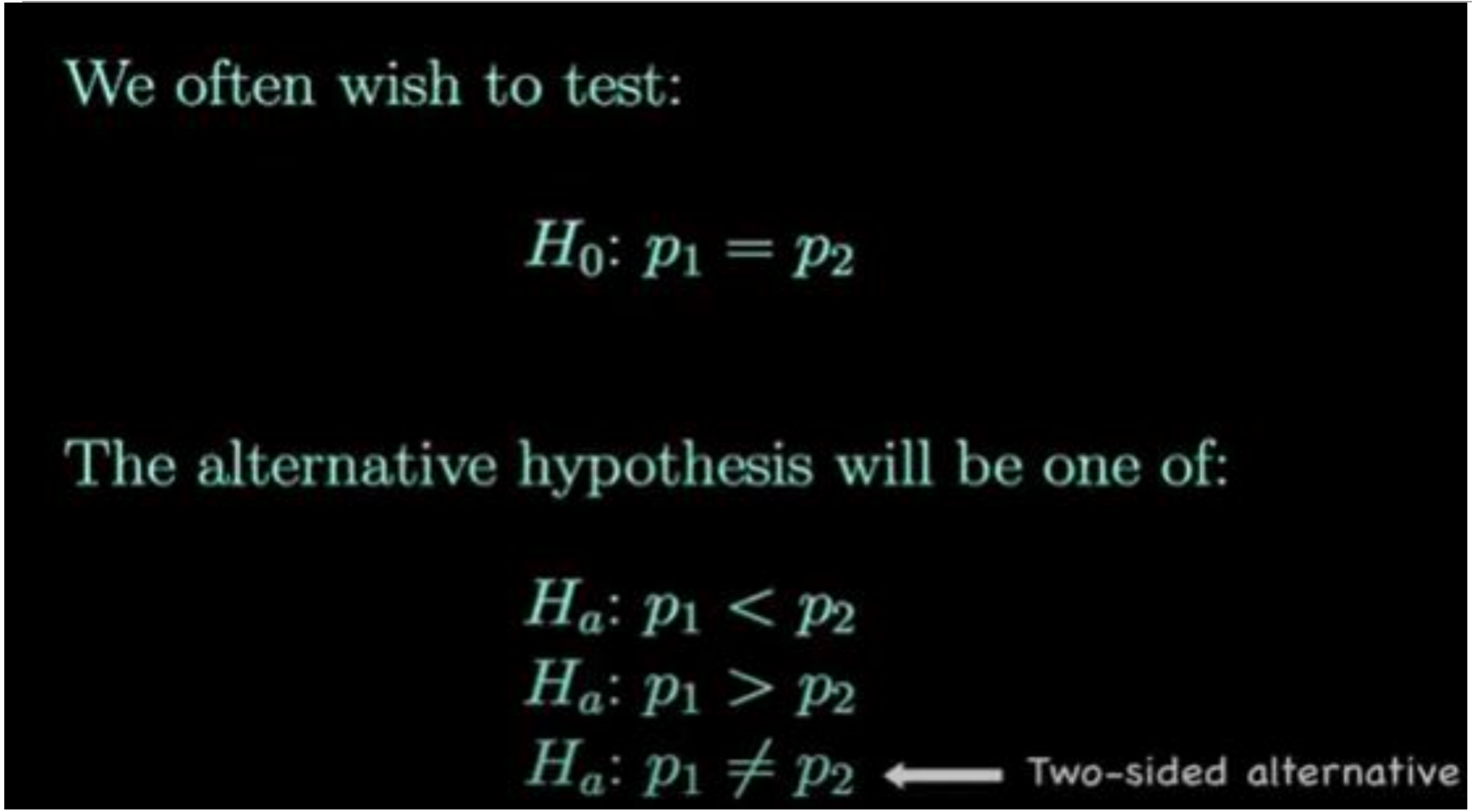
A $(1-\alpha)100\%$ confidence interval for $p_1 - p_2$ is given by:

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \times SE(\hat{p}_1 - \hat{p}_2)$$

$$\longrightarrow \quad \sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Hypotheis testing for two proportions



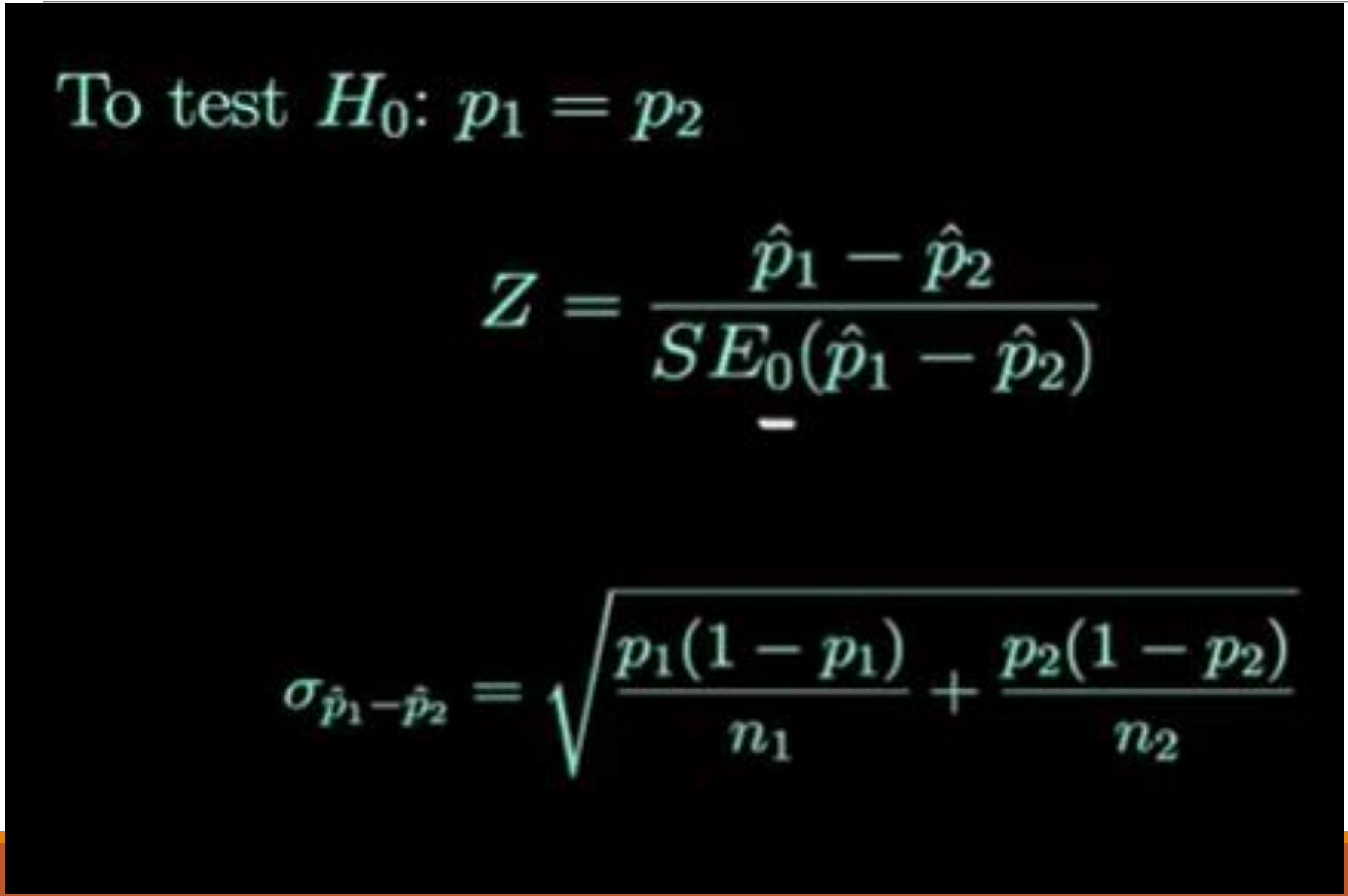
We often wish to test:

$$H_0: p_1 = p_2$$

The alternative hypothesis will be one of:

$$H_a: p_1 < p_2$$
$$H_a: p_1 > p_2$$
$$H_a: p_1 \neq p_2$$

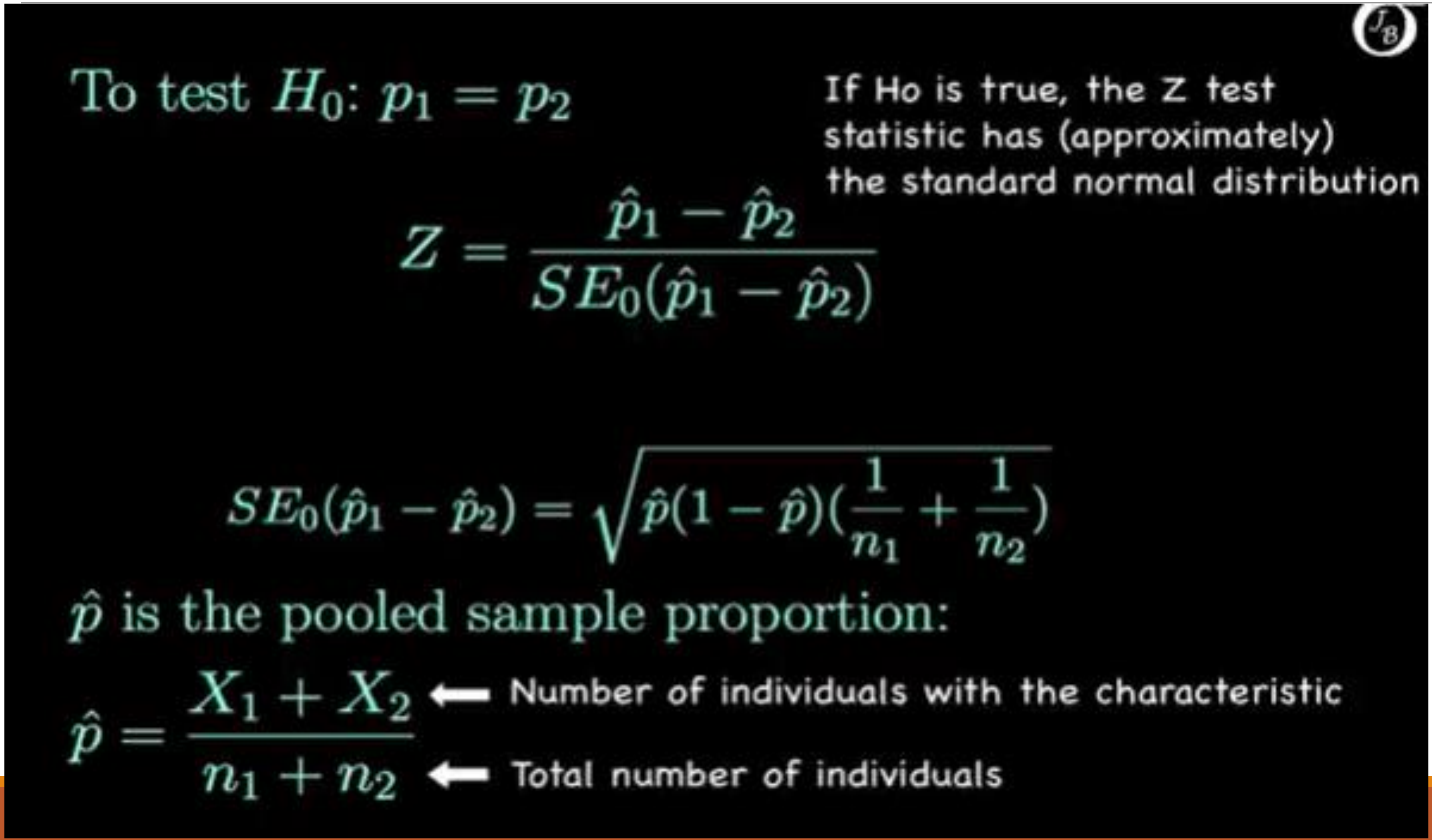
← Two-sided alternative



To test H_0: $p_1 = p_2$

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{SE_0(\hat{p}_1 - \hat{p}_2)}$$

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}$$



To test H_0: $p_1 = p_2$

If Ho is true, the Z test statistic has (approximately) the standard normal distribution

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{SE_0(\hat{p}_1 - \hat{p}_2)}$$

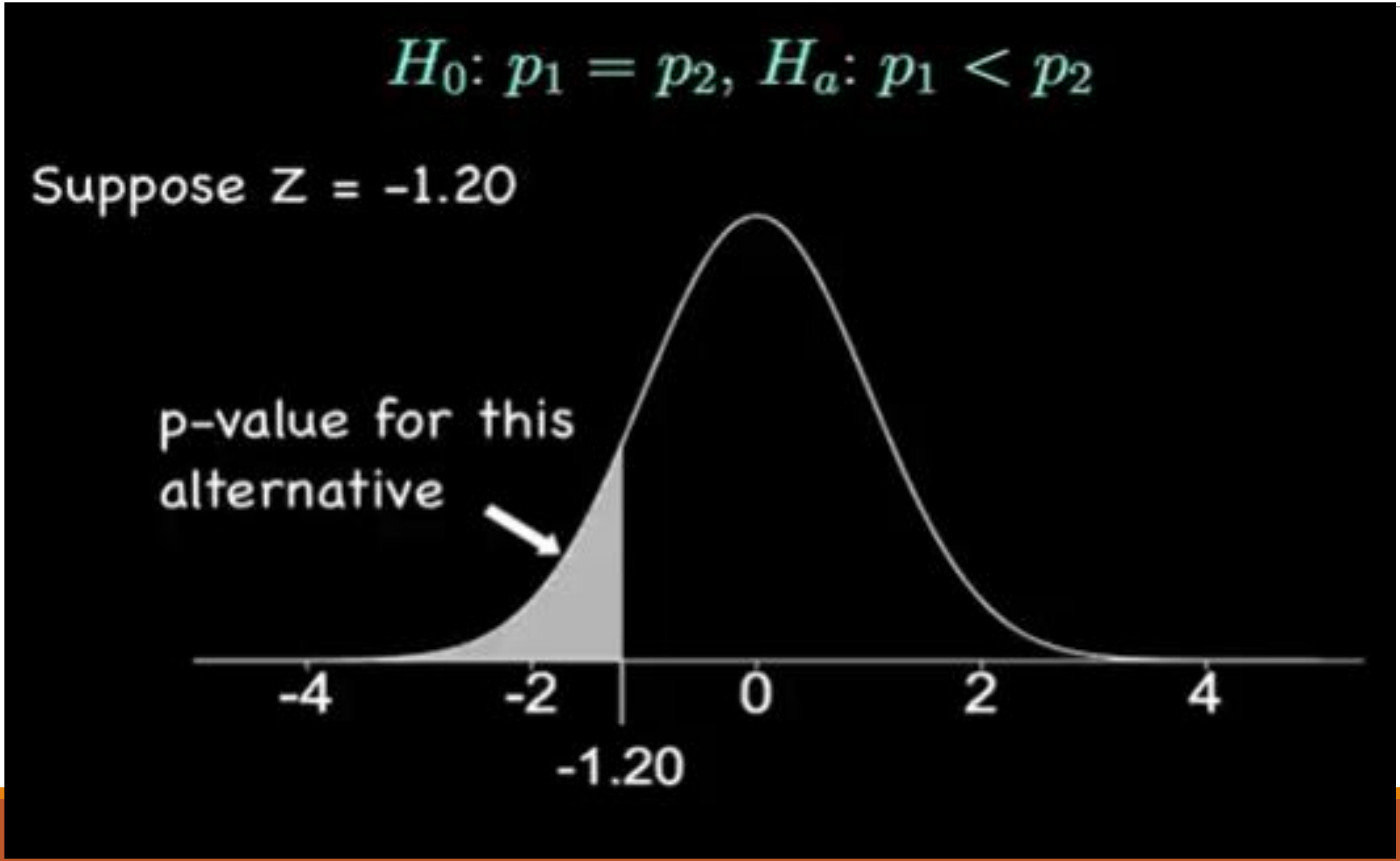
$$SE_0(\hat{p}_1 - \hat{p}_2) = \sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}$$

$\hat{p}$ is the pooled sample proportion:

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

← Number of individuals with the characteristic

← Total number of individuals



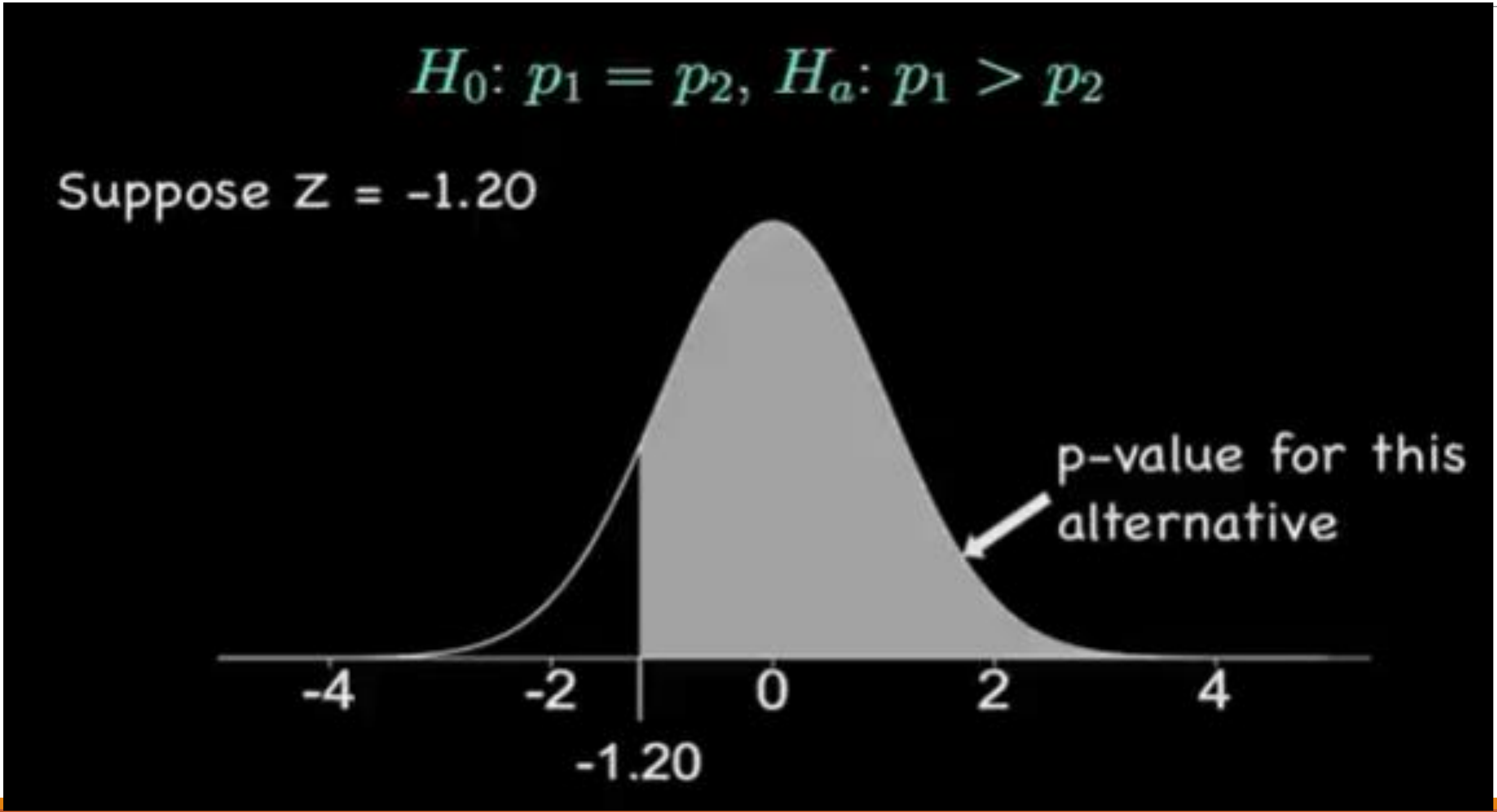
H_0: $p_1 = p_2$, H_a: $p_1 < p_2$

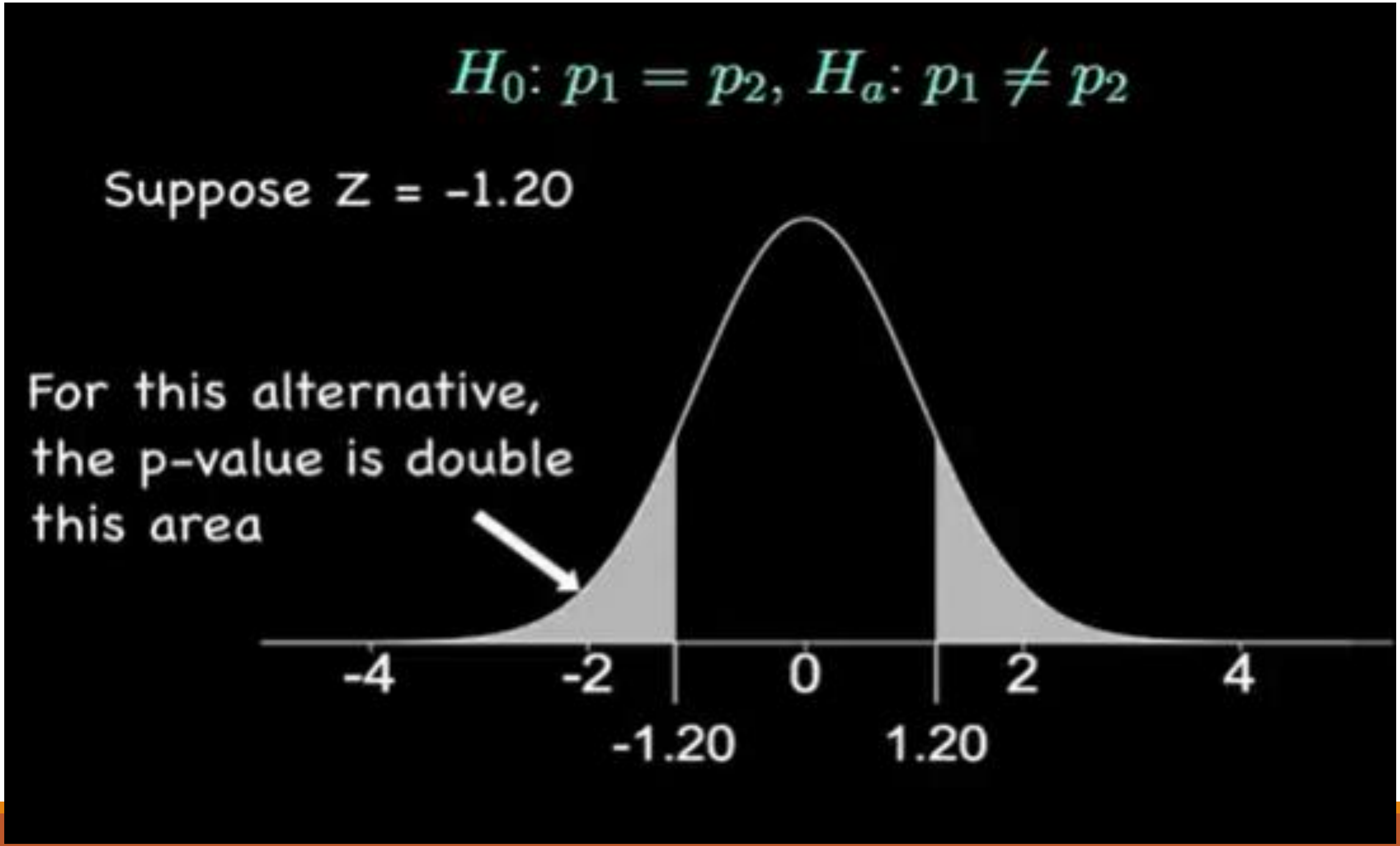
Suppose Z = -1.20

p-value for this alternative

H_0: $p_1 = p_2$, H_a: $p_1 > p_2$

Suppose Z = -1.20

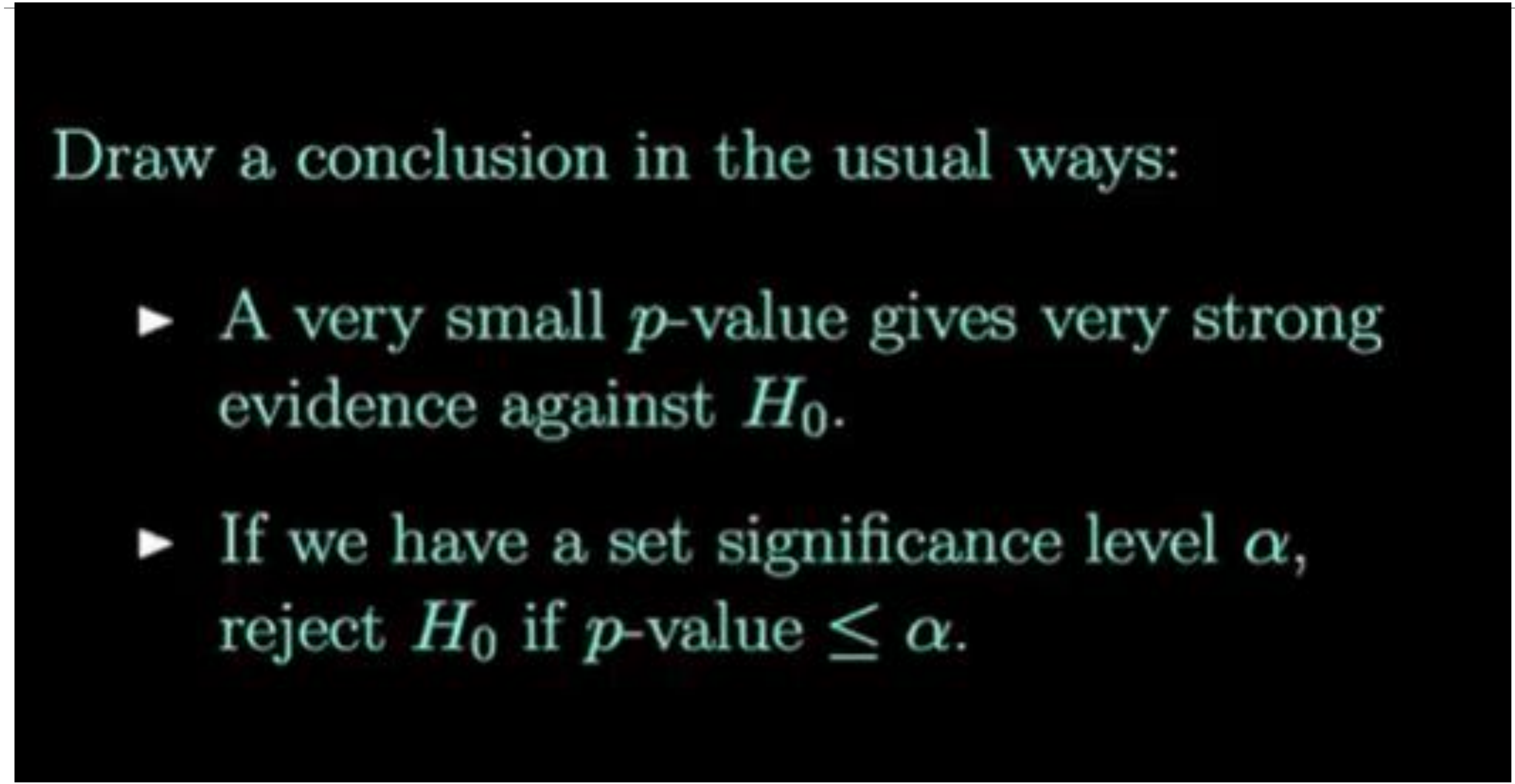




H_0: $p_1 = p_2$, H_a: $p_1 \neq p_2$

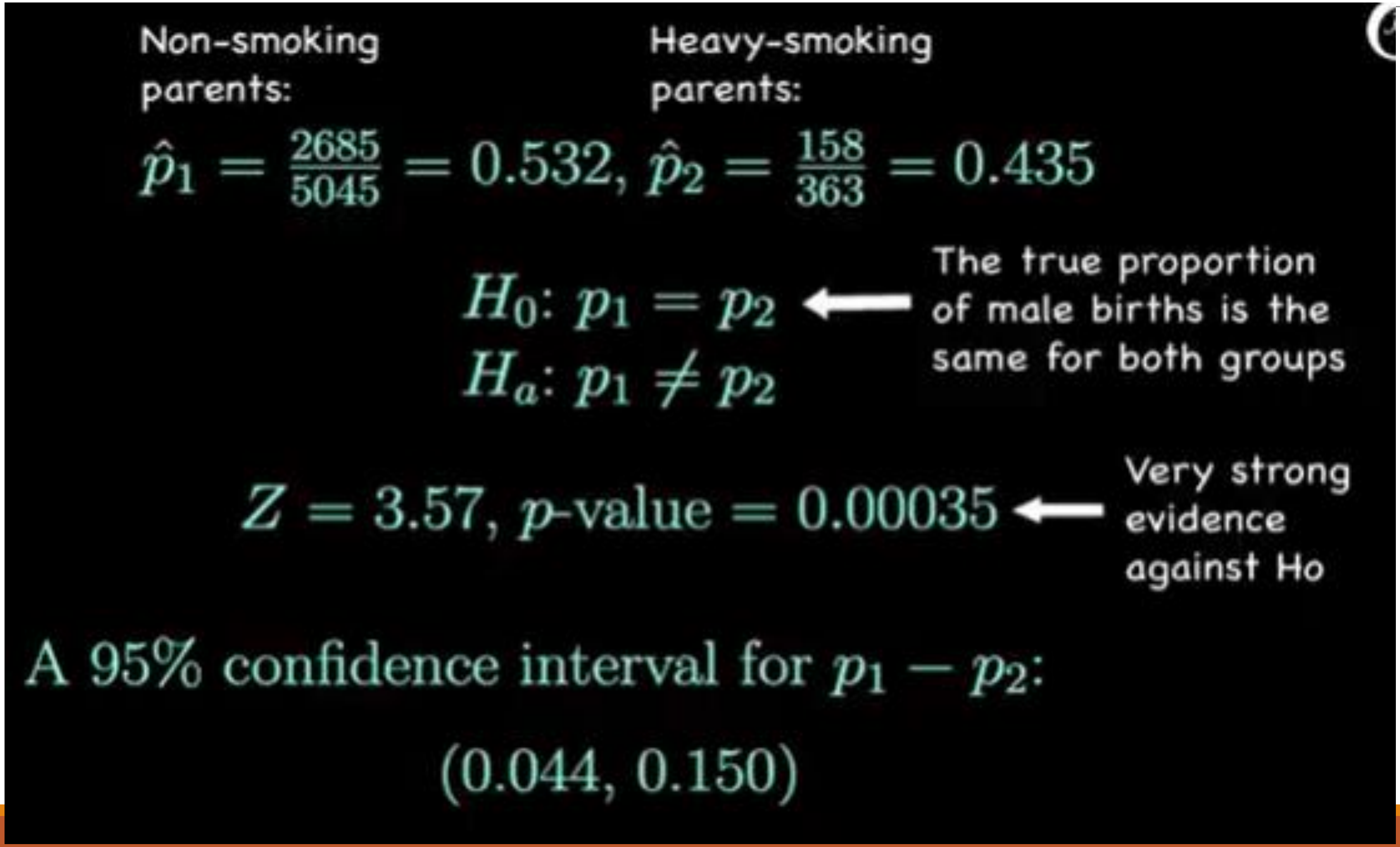
Suppose Z = -1.20

For this alternative, the p-value is double this area



Draw a conclusion in the usual ways:

- A very small p-value gives very strong evidence against H_0.
- If we have a set significance level α, reject H_0 if p-value $\leq \alpha$.



Non-smoking parents:

Heavy-smoking parents:

$$\hat{p}_1 = \frac{2685}{5045} = 0.532, \hat{p}_2 = \frac{158}{363} = 0.435$$

$$H_0\colon p_1 = p_2$$

$$H_a\colon p_1 \neq p_2$$

The true proportion of male births is the same for both groups

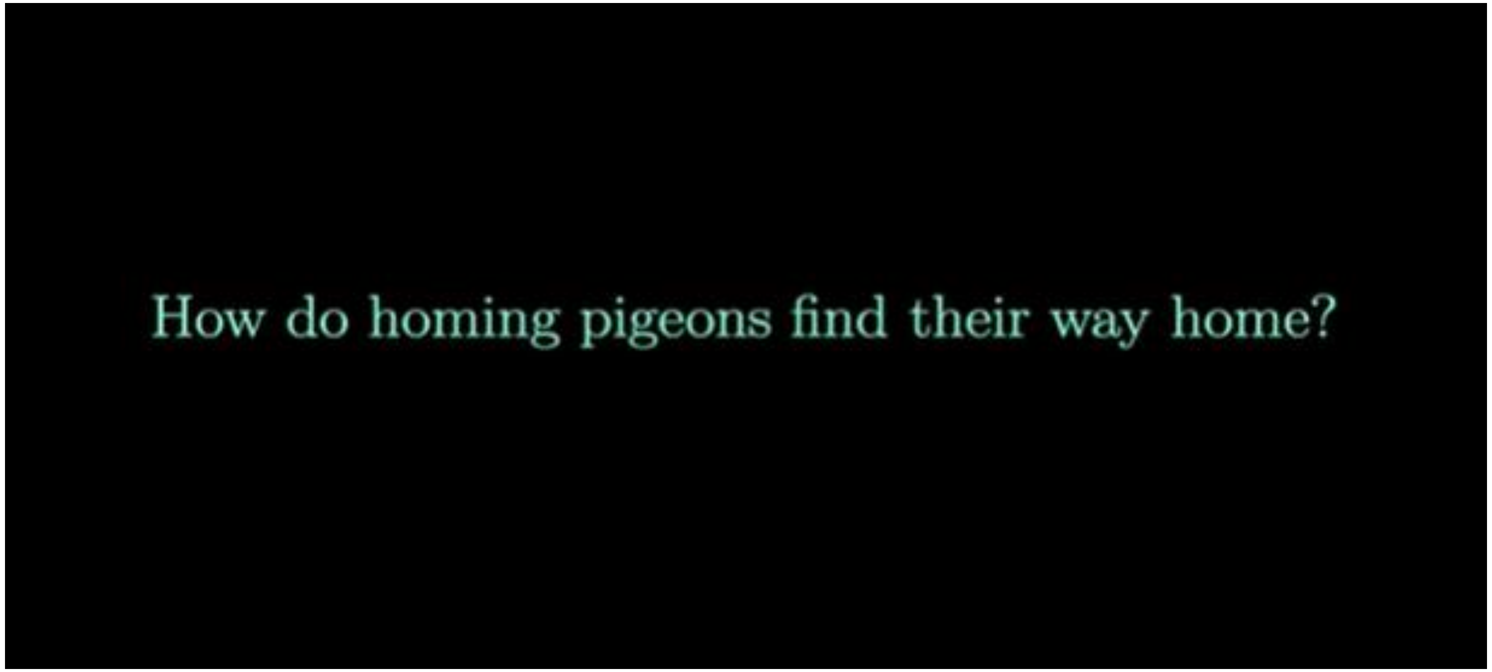
$$Z = 3.57, p\text{-value} = 0.00035$$

Very strong evidence against Ho

A 95% confidence interval for $p_1 - p_2$:

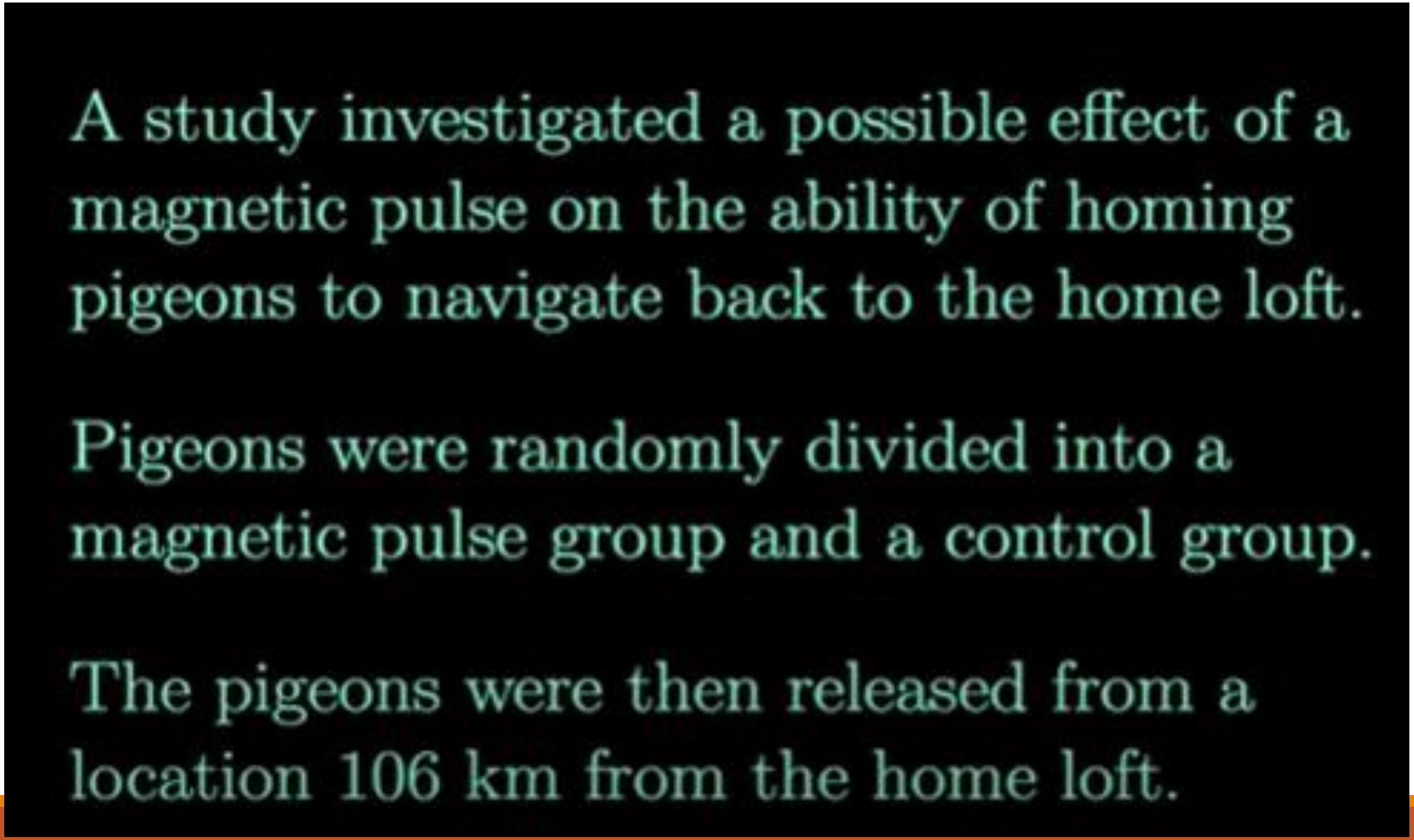
$$(0.044, 0.150)$$

Example



How do homing pigeons find their way home?

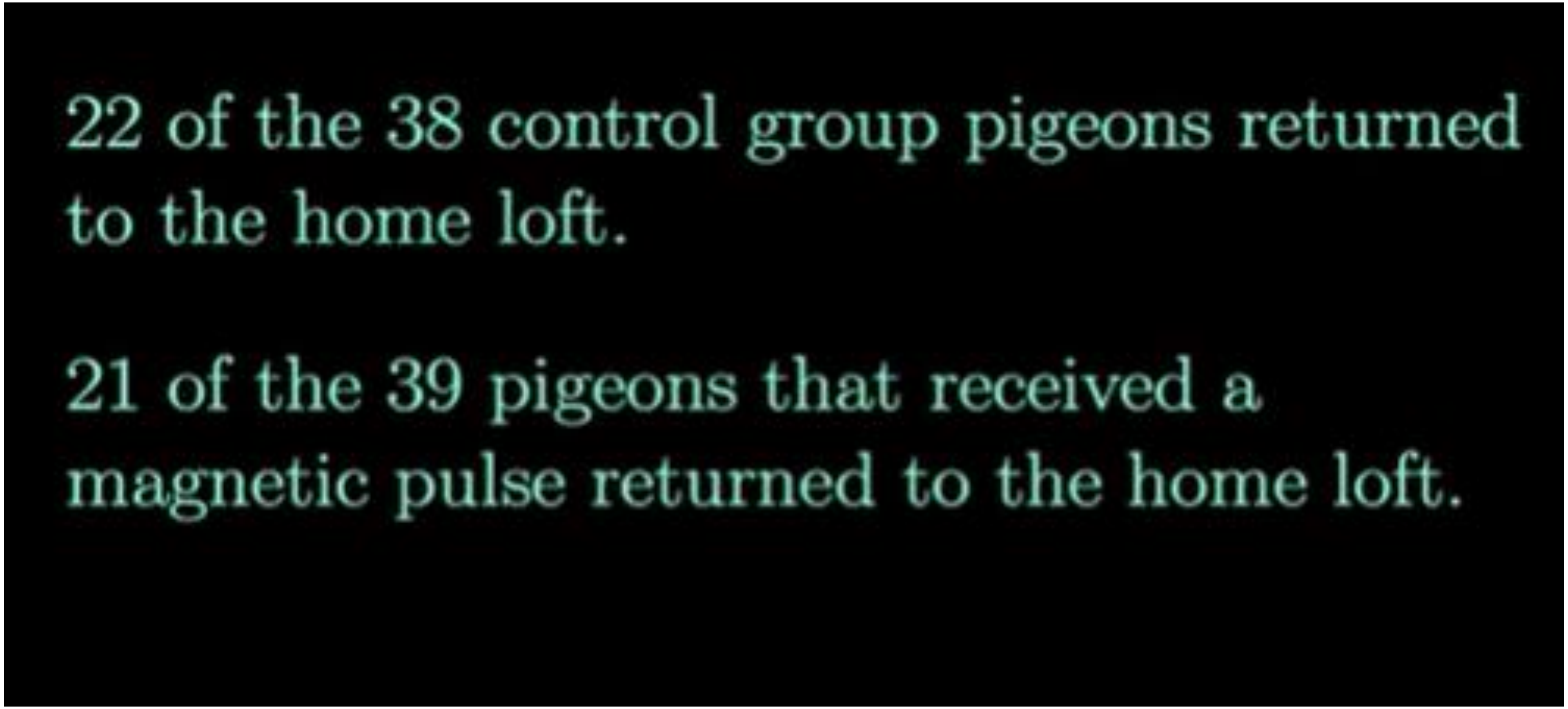
Example



A study investigated a possible effect of a magnetic pulse on the ability of homing pigeons to navigate back to the home loft.

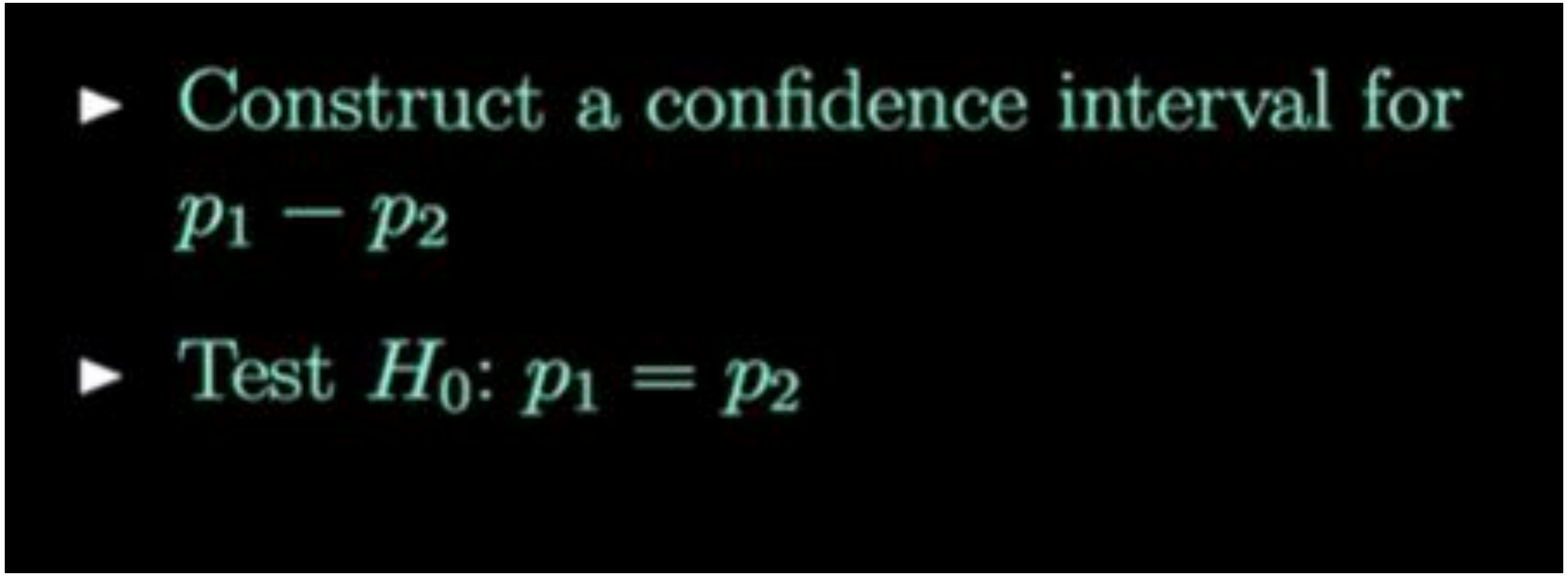
Pigeons were randomly divided into a magnetic pulse group and a control group.

The pigeons were then released from a location 106 km from the home loft.

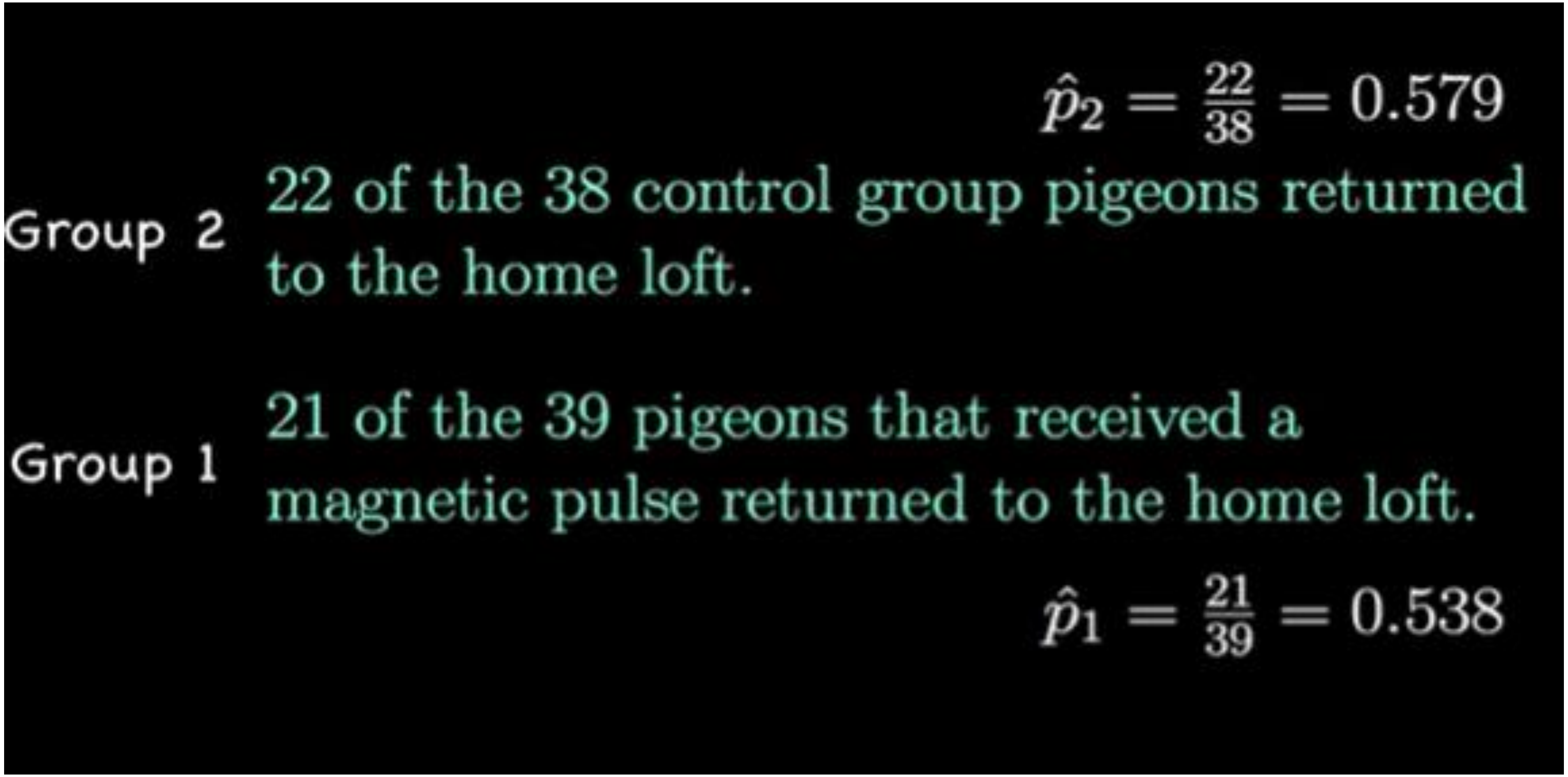


22 of the 38 control group pigeons returned to the home loft.

21 of the 39 pigeons that received a magnetic pulse returned to the home loft.



- Construct a confidence interval for $p_1 - p_2$
- Test H_0: $p_1 = p_2$



$$\hat{p}_2 = \frac{22}{38} = 0.579$$

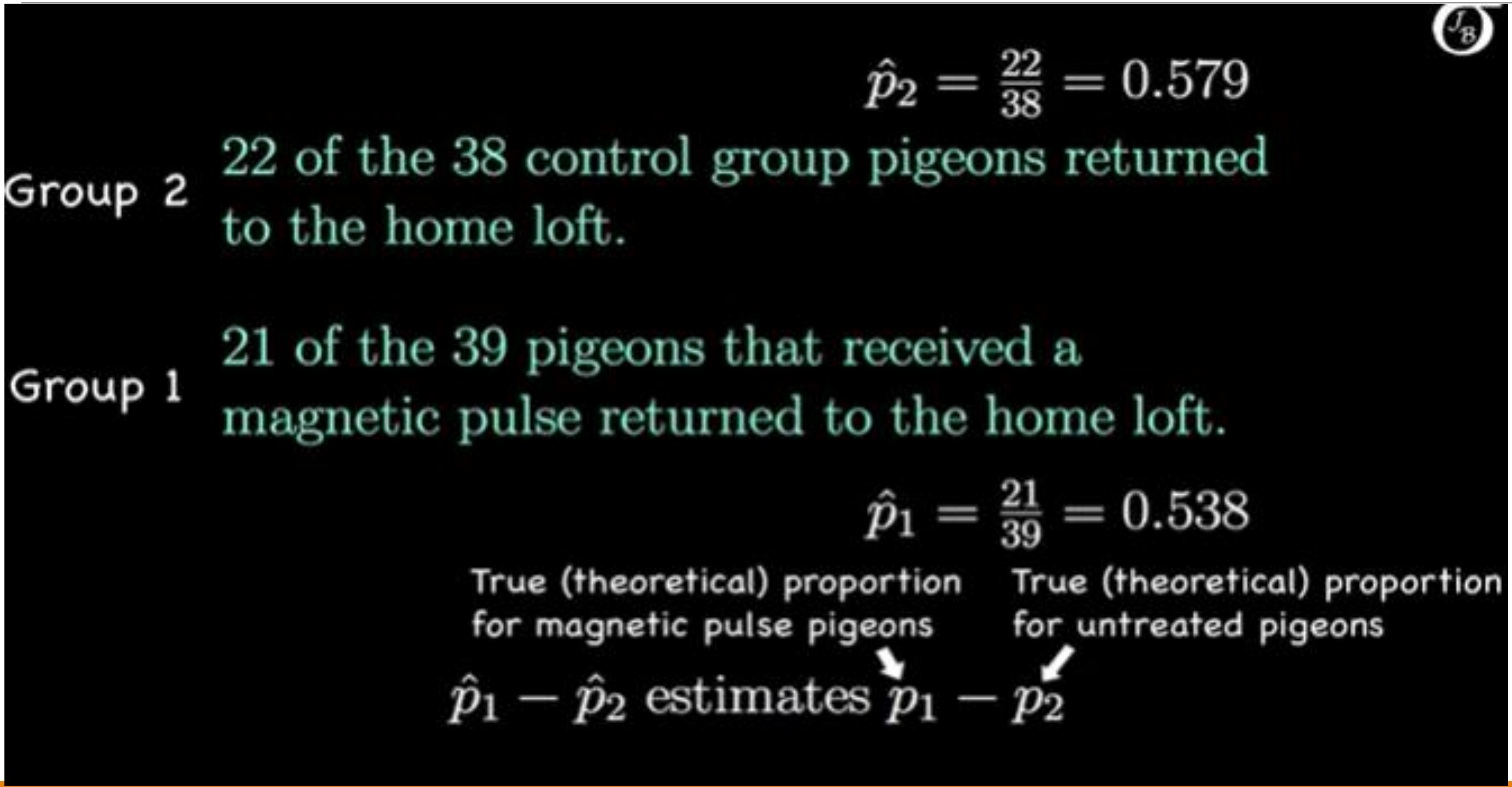
Group 2

22 of the 38 control group pigeons returned to the home loft.

Group 1

21 of the 39 pigeons that received a magnetic pulse returned to the home loft.

$$\hat{p}_1 = \frac{21}{39} = 0.538$$



$$\hat{p}_2 = \frac{22}{38} = 0.579$$

Group 2 22 of the 38 control group pigeons returned to the home loft.

Group 1 21 of the 39 pigeons that received a magnetic pulse returned to the home loft.

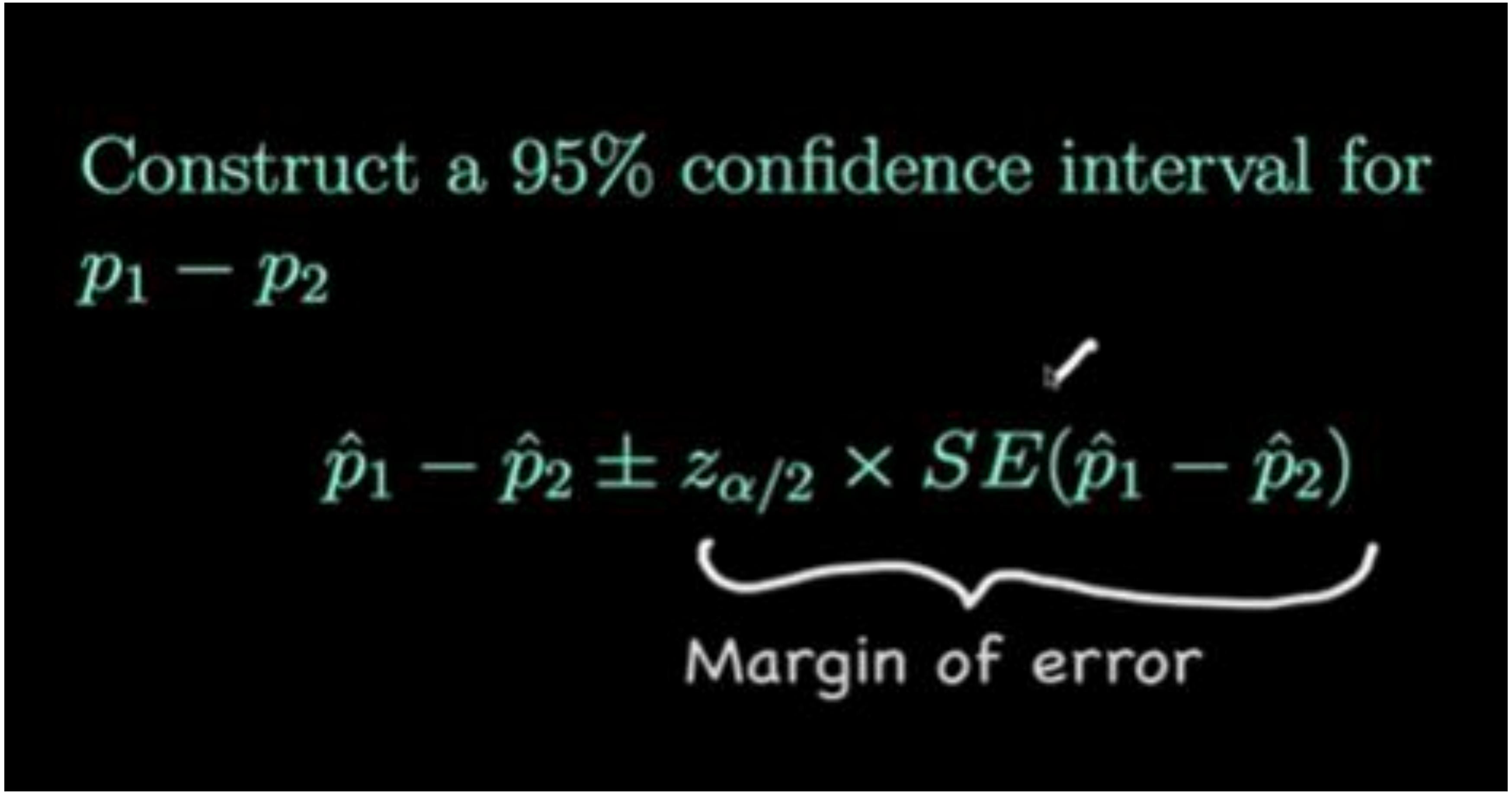
$$\hat{p}_1 = \frac{21}{39} = 0.538$$

True (theoretical) proportion for magnetic pulse pigeons

True (theoretical) proportion for untreated pigeons

$\hat{p}_1 - \hat{p}_2$ estimates $p_1 - p_2$

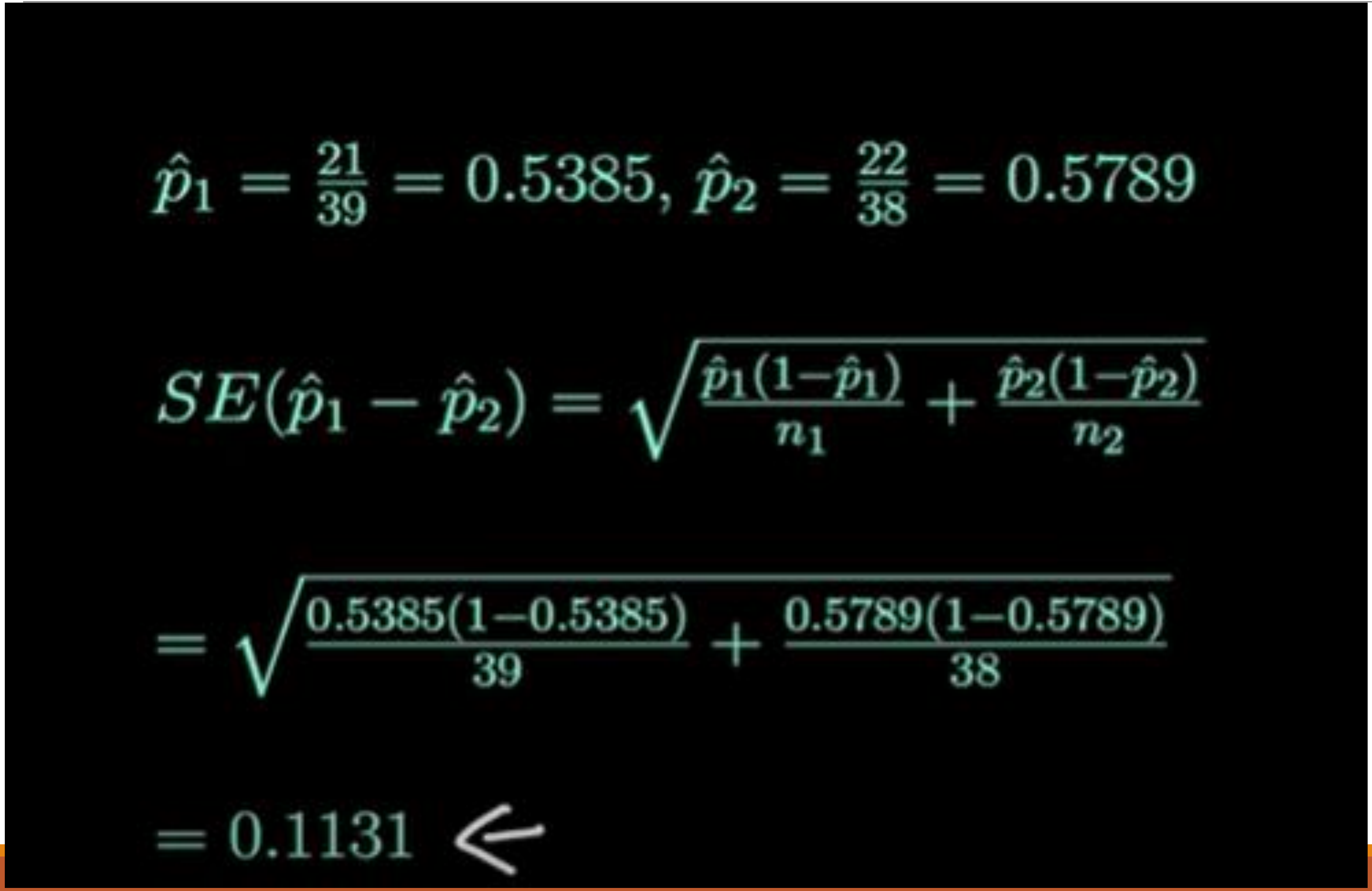
CI



Construct a 95% confidence interval for $p_1 - p_2$

$$\hat{p}_1 - \hat{p}_2 \pm \underbrace{z_{\alpha/2} \times SE(\hat{p}_1 - \hat{p}_2)}_{\text{Margin of error}}$$

CI



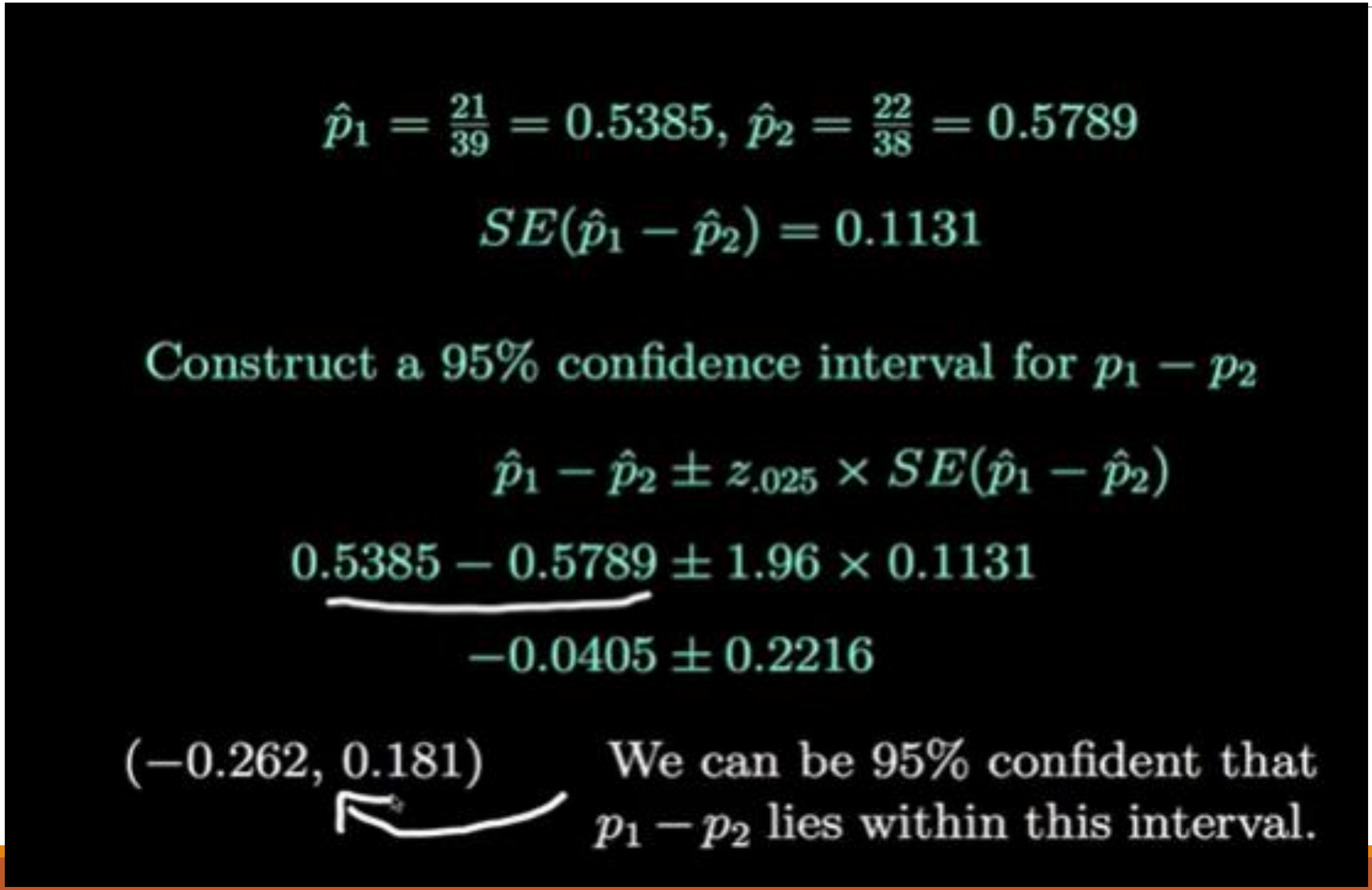
$$\hat{p}_1 = \frac{21}{39} = 0.5385,\ \hat{p}_2 = \frac{22}{38} = 0.5789$$

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$= \sqrt{\frac{0.5385(1-0.5385)}{39} + \frac{0.5789(1-0.5789)}{38}}$$

$$= 0.1131 \leftarrow$$

CI



$$\hat{p}_1 = \tfrac{21}{39} = 0.5385,\ \hat{p}_2 = \tfrac{22}{38} = 0.5789$$

$$SE(\hat{p}_1 - \hat{p}_2) = 0.1131$$

Construct a 95% confidence interval for $p_1 - p_2$

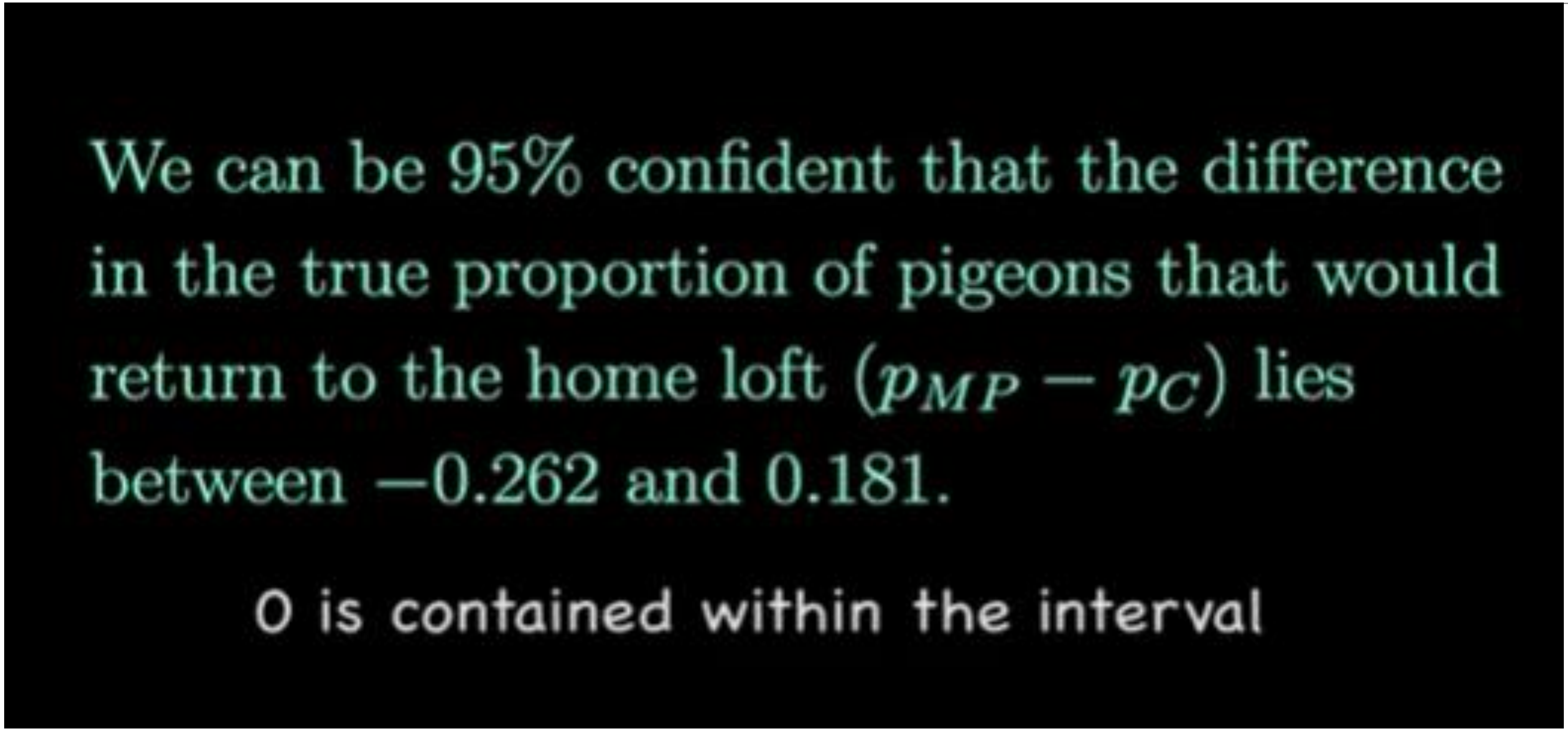
$$\hat{p}_1 - \hat{p}_2 \pm z_{.025} \times SE(\hat{p}_1 - \hat{p}_2)$$

$$\underline{0.5385 - 0.5789} \pm 1.96 \times 0.1131$$

$$-0.0405 \pm 0.2216$$

$(-0.262,\ 0.181)$ We can be 95% confident that $p_1 - p_2$ lies within this interval.

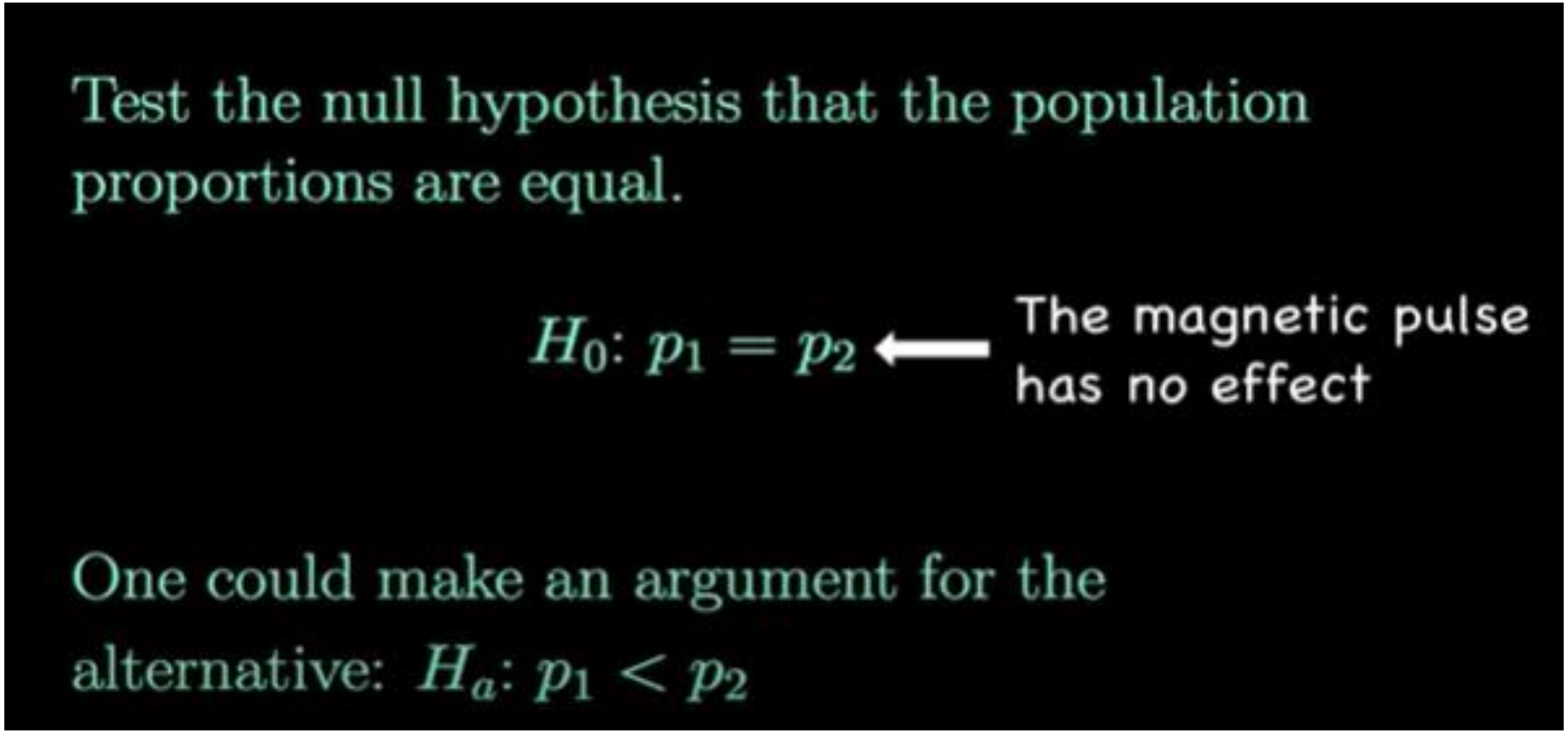
CI



We can be 95% confident that the difference in the true proportion of pigeons that would return to the home loft $(p_{MP} - p_C)$ lies between −0.262 and 0.181.

0 is contained within the interval

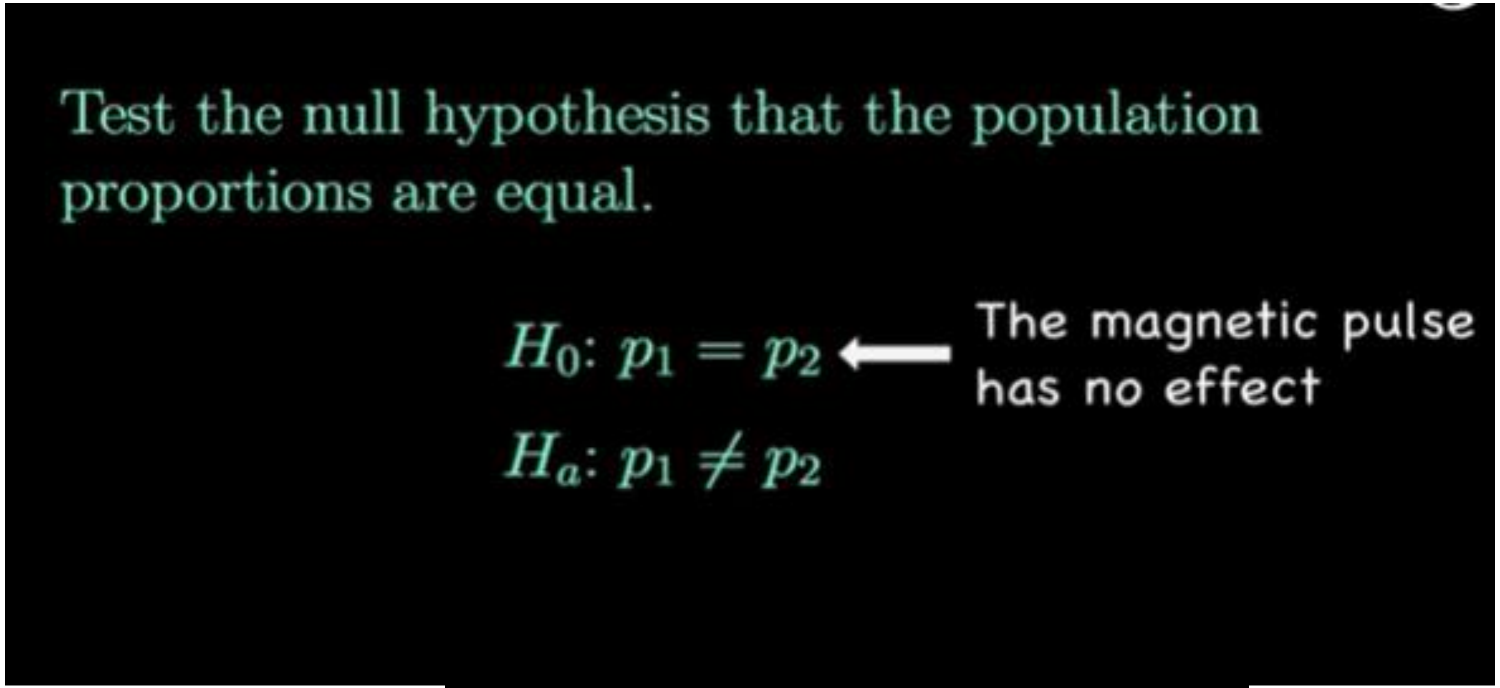
Hypothesis testing



Test the null hypothesis that the population proportions are equal.

$$H_0\colon p_1 = p_2 \longleftarrow$$ The magnetic pulse has no effect

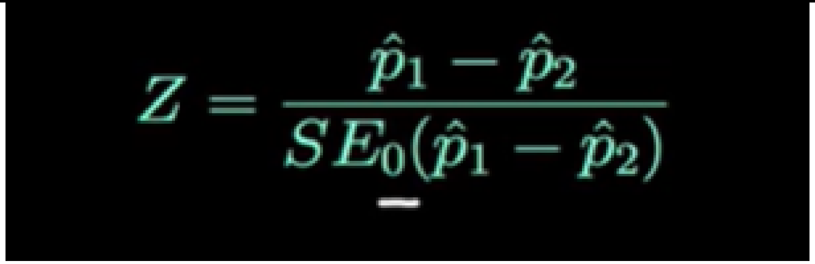
One could make an argument for the alternative: $H_a\colon p_1 < p_2$



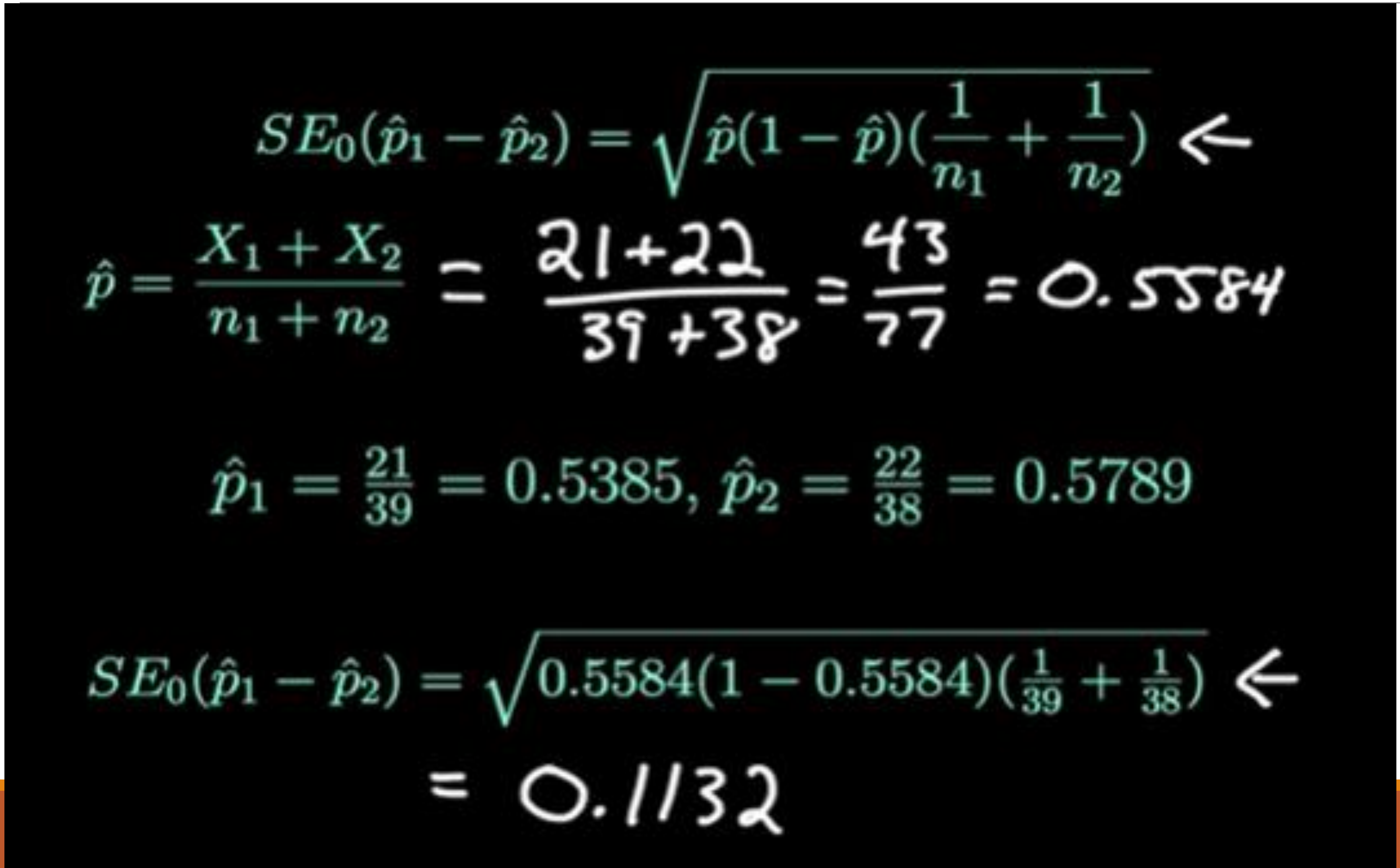
Test the null hypothesis that the population proportions are equal.

$H_0\colon p_1 = p_2$ ← The magnetic pulse has no effect

$H_a\colon p_1 \neq p_2$



$$Z = \frac{\hat{p}_1 - \hat{p}_2}{SE_{\underline{0}}(\hat{p}_1 - \hat{p}_2)}$$



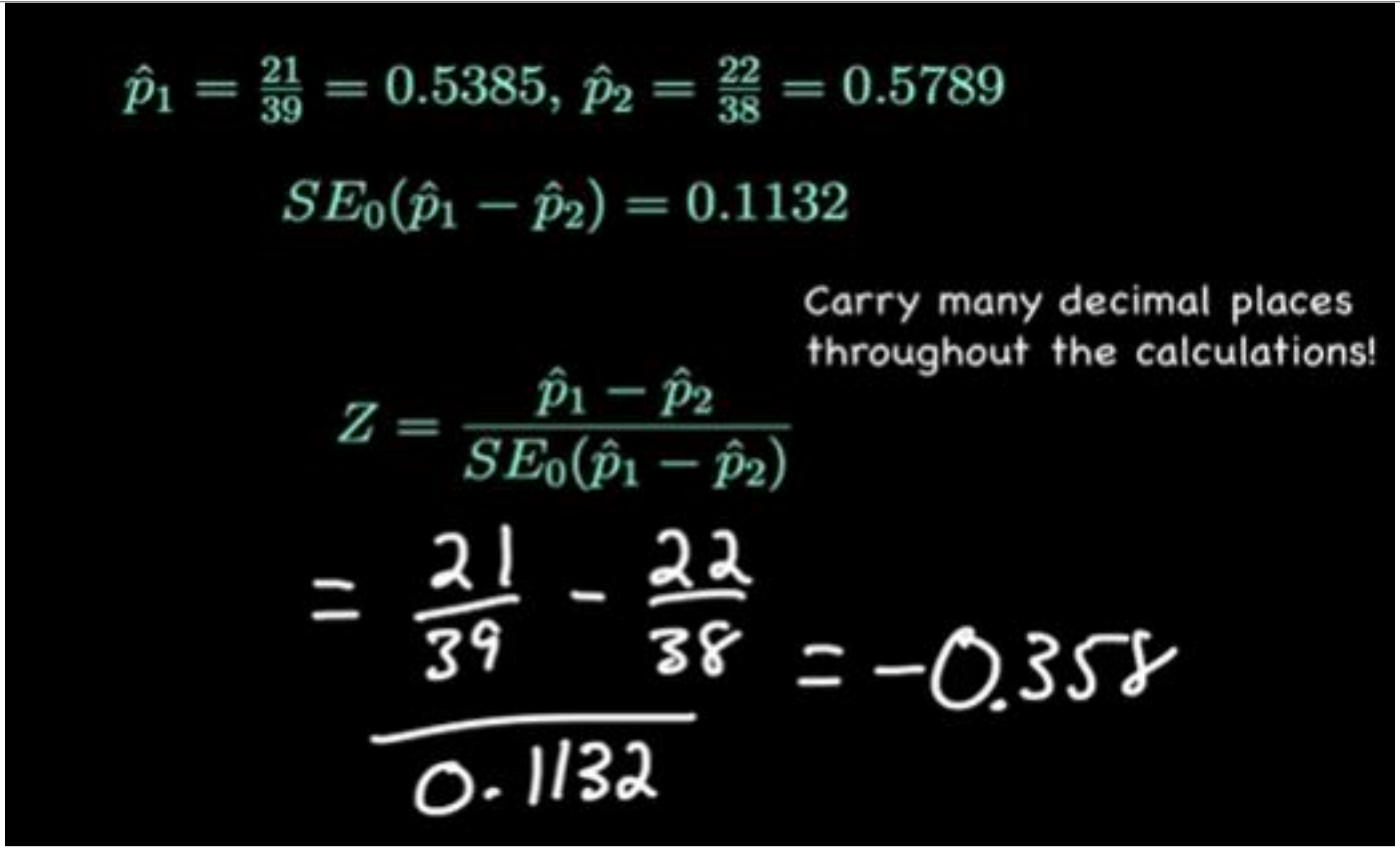
$$SE_0(\hat{p}_1 - \hat{p}_2) = \sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})} \leftarrow$$

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{21+22}{39+38} = \frac{43}{77} = 0.5584$$

$$\hat{p}_1 = \tfrac{21}{39} = 0.5385,\ \hat{p}_2 = \tfrac{22}{38} = 0.5789$$

$$SE_0(\hat{p}_1 - \hat{p}_2) = \sqrt{0.5584(1-0.5584)(\tfrac{1}{39} + \tfrac{1}{38})} \leftarrow$$

$$= 0.1132$$



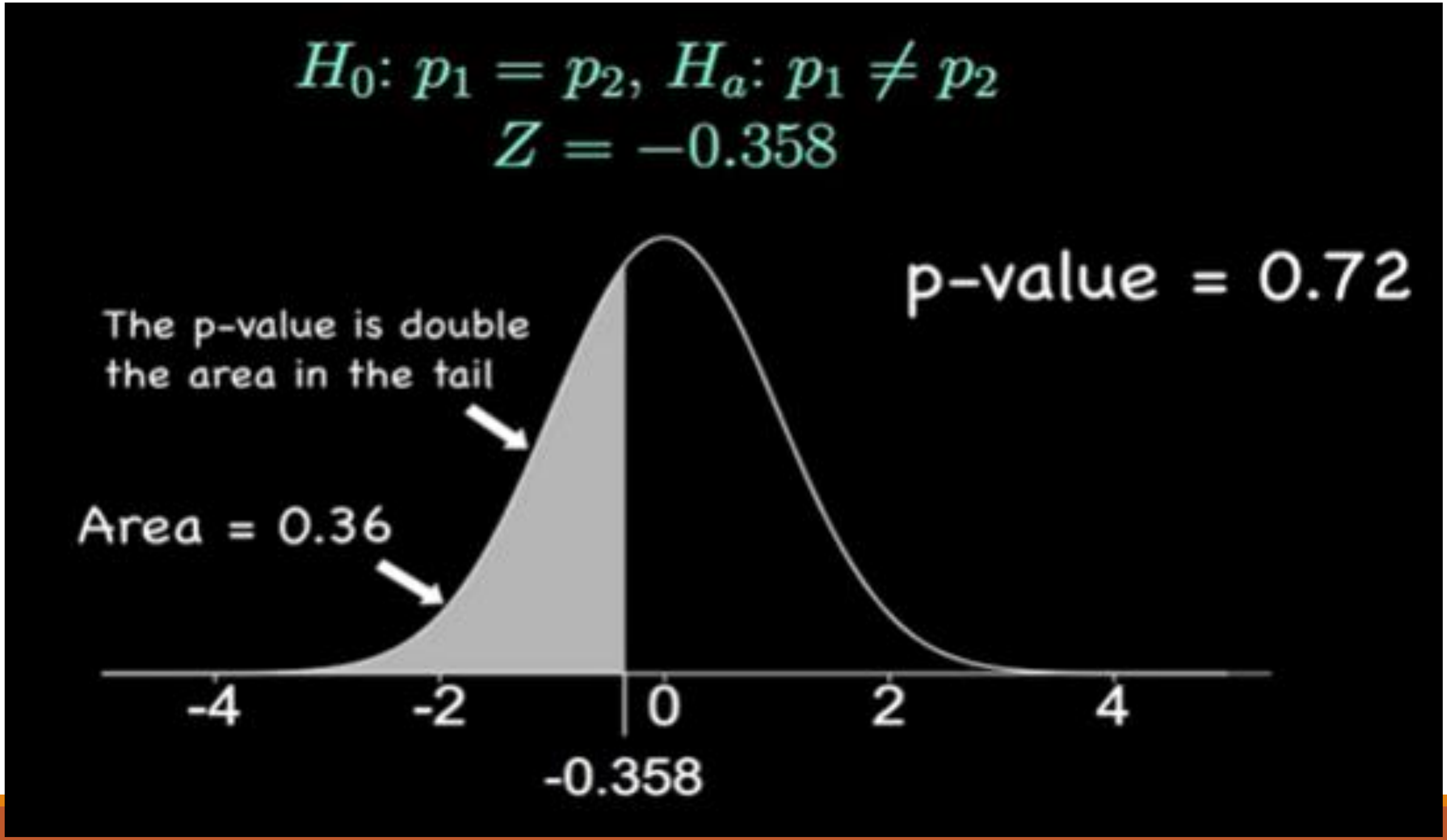
$$\hat{p}_1 = \frac{21}{39} = 0.5385,\ \hat{p}_2 = \frac{22}{38} = 0.5789$$

$$SE_0(\hat{p}_1 - \hat{p}_2) = 0.1132$$

Carry many decimal places throughout the calculations!

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{SE_0(\hat{p}_1 - \hat{p}_2)}$$

$$= \frac{\frac{21}{39} - \frac{22}{38}}{0.1132} = -0.358$$



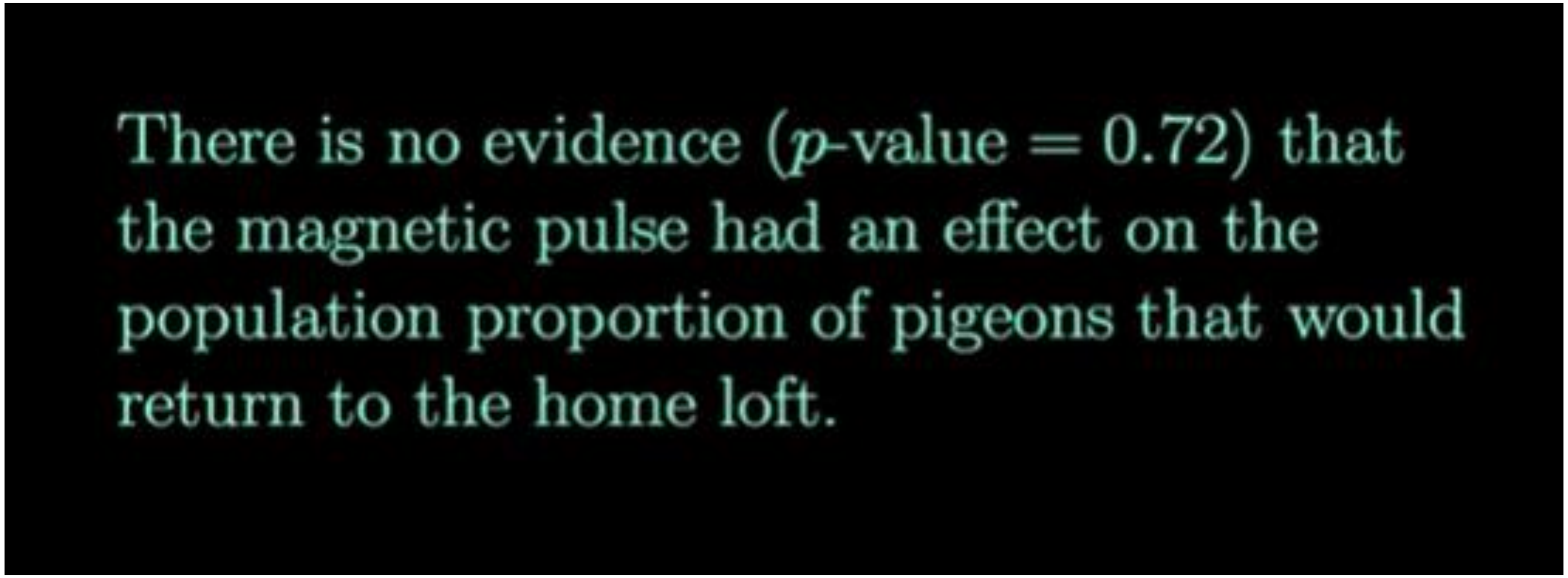
H_0: $p_1 = p_2$, H_a: $p_1 \neq p_2$

$Z = -0.358$

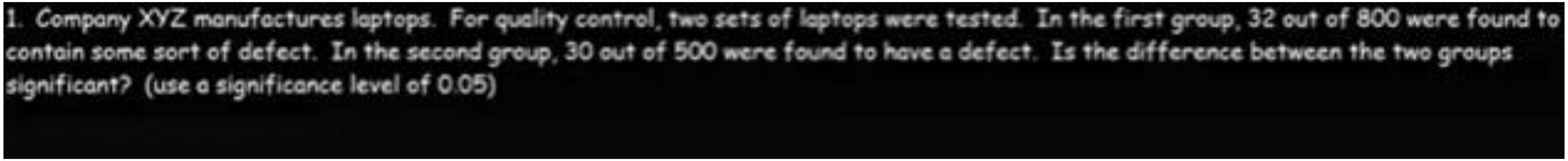
p-value = 0.72

The p-value is double the area in the tail

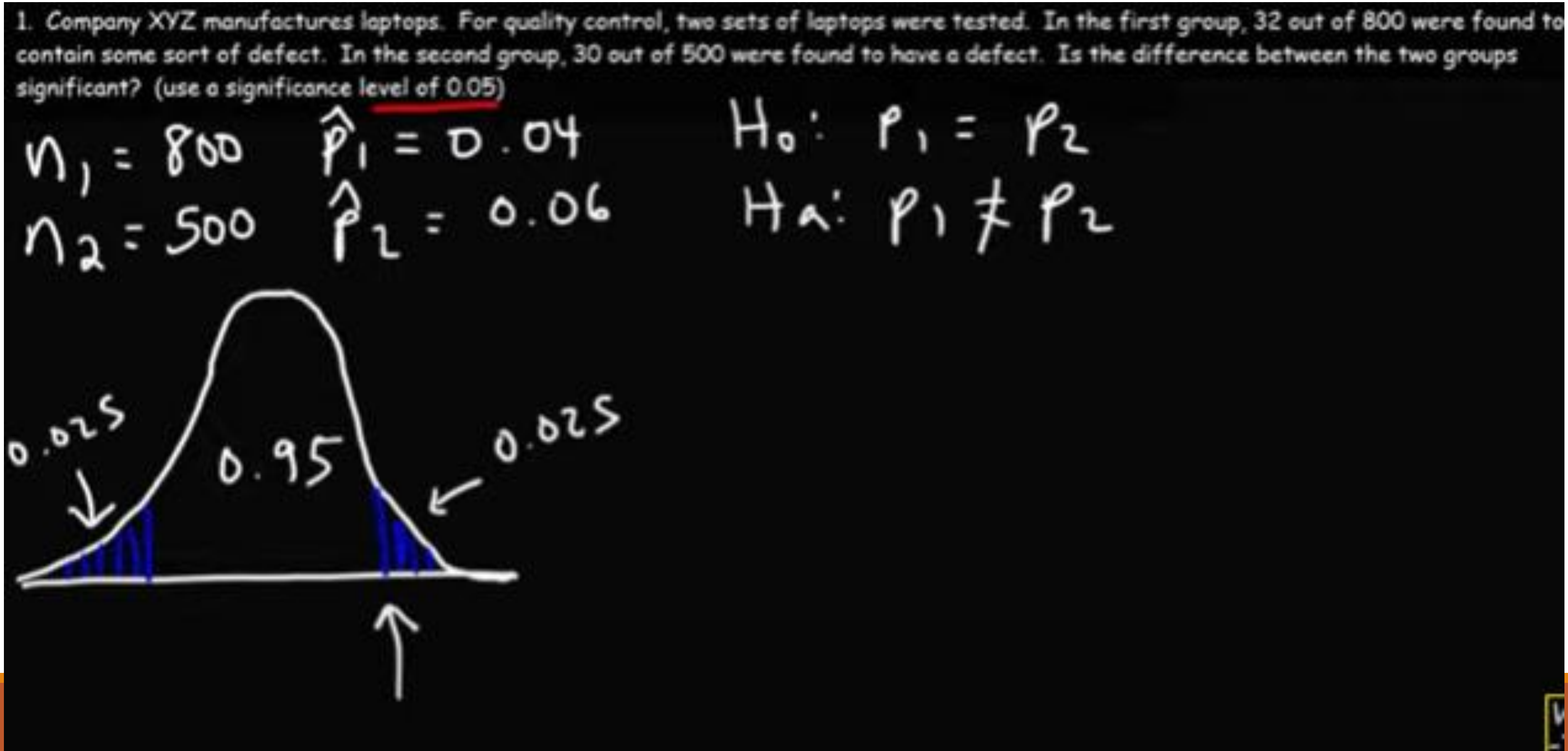
Area = 0.36



There is no evidence (p-value $= 0.72$) that the magnetic pulse had an effect on the population proportion of pigeons that would return to the home loft.



1. Company XYZ manufactures laptops. For quality control, two sets of laptops were tested. In the first group, 32 out of 800 were found to contain some sort of defect. In the second group, 30 out of 500 were found to have a defect. Is the difference between the two groups significant? (use a significance level of 0.05)



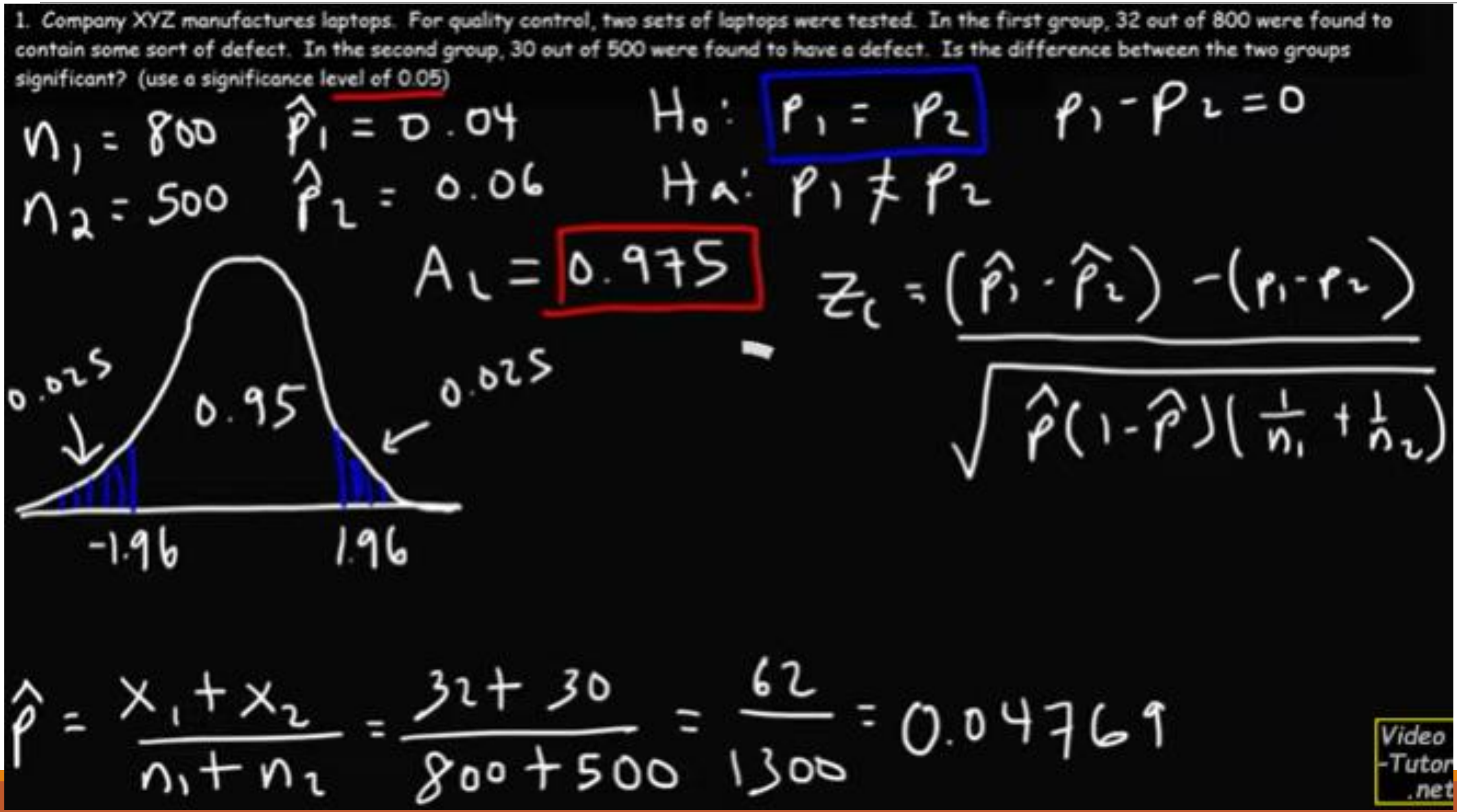
1. Company XYZ manufactures laptops. For quality control, two sets of laptops were tested. In the first group, 32 out of 800 were found to contain some sort of defect. In the second group, 30 out of 500 were found to have a defect. Is the difference between the two groups significant? (use a significance level of 0.05)

$n_1 = 800 \quad \hat{p}_1 = 0.04$

$n_2 = 500 \quad \hat{p}_2 = 0.06$

$H_0: p_1 = p_2$

$H_a: p_1 \neq p_2$



1. Company XYZ manufactures laptops. For quality control, two sets of laptops were tested. In the first group, 32 out of 800 were found to contain some sort of defect. In the second group, 30 out of 500 were found to have a defect. Is the difference between the two groups significant? (use a significance level of 0.05)

$n_1 = 800 \quad \hat{p}_1 = 0.04$

$n_2 = 500 \quad \hat{p}_2 = 0.06$

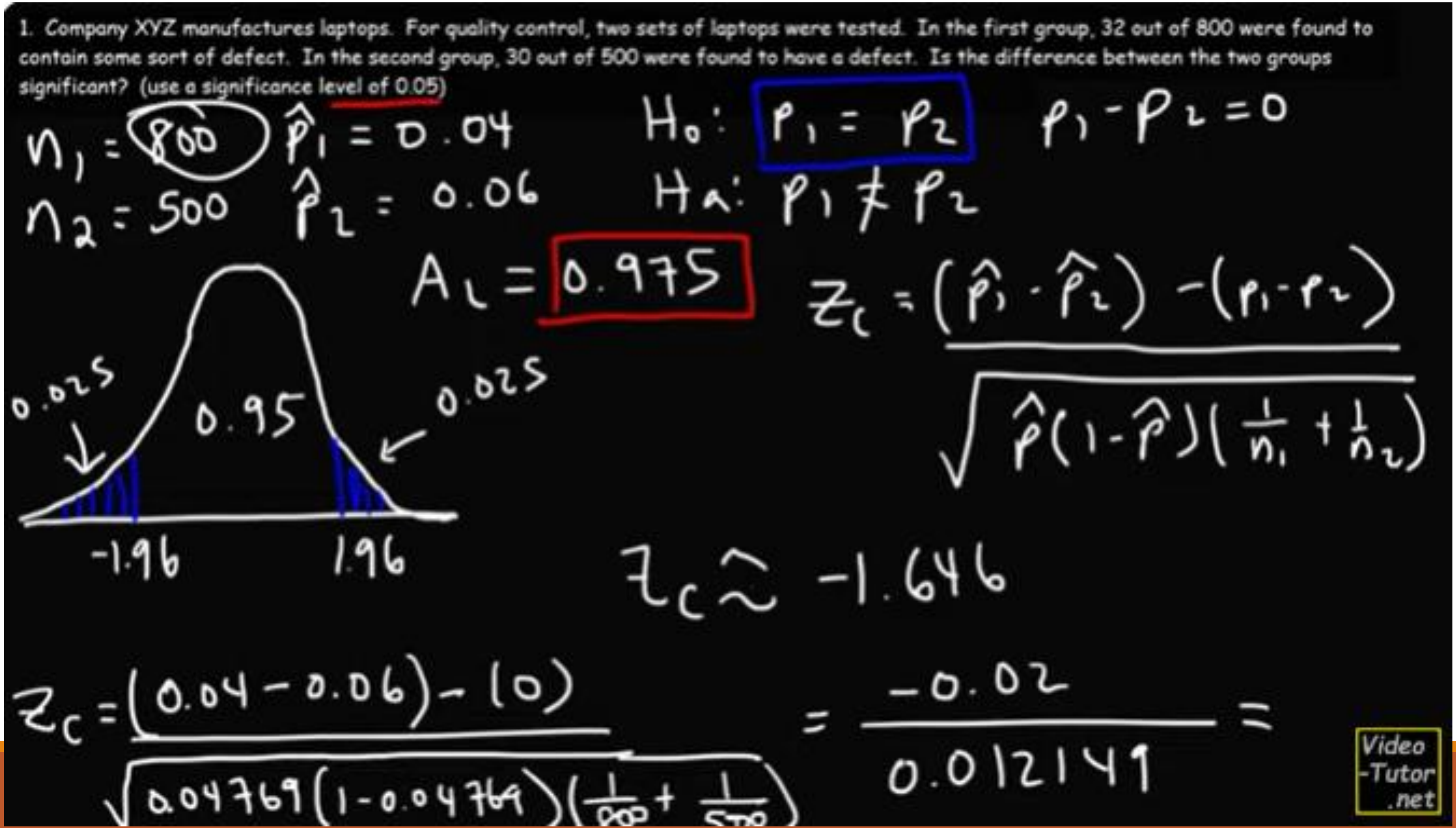
$H_0: \boxed{p_1 = p_2} \quad p_1 - p_2 = 0$

$H_a: p_1 \neq p_2$

$A_L = \boxed{0.975}$

$$Z_c = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{32 + 30}{800 + 500} = \frac{62}{1300} = 0.04769$$



1. Company XYZ manufactures laptops. For quality control, two sets of laptops were tested. In the first group, 32 out of 800 were found to contain some sort of defect. In the second group, 30 out of 500 were found to have a defect. Is the difference between the two groups significant? (use a significance level of 0.05)

$n_1 = 800 \quad \hat{p}_1 = 0.04$

$n_2 = 500 \quad \hat{p}_2 = 0.06$

$H_0: \boxed{p_1 = p_2} \quad p_1 - p_2 = 0$

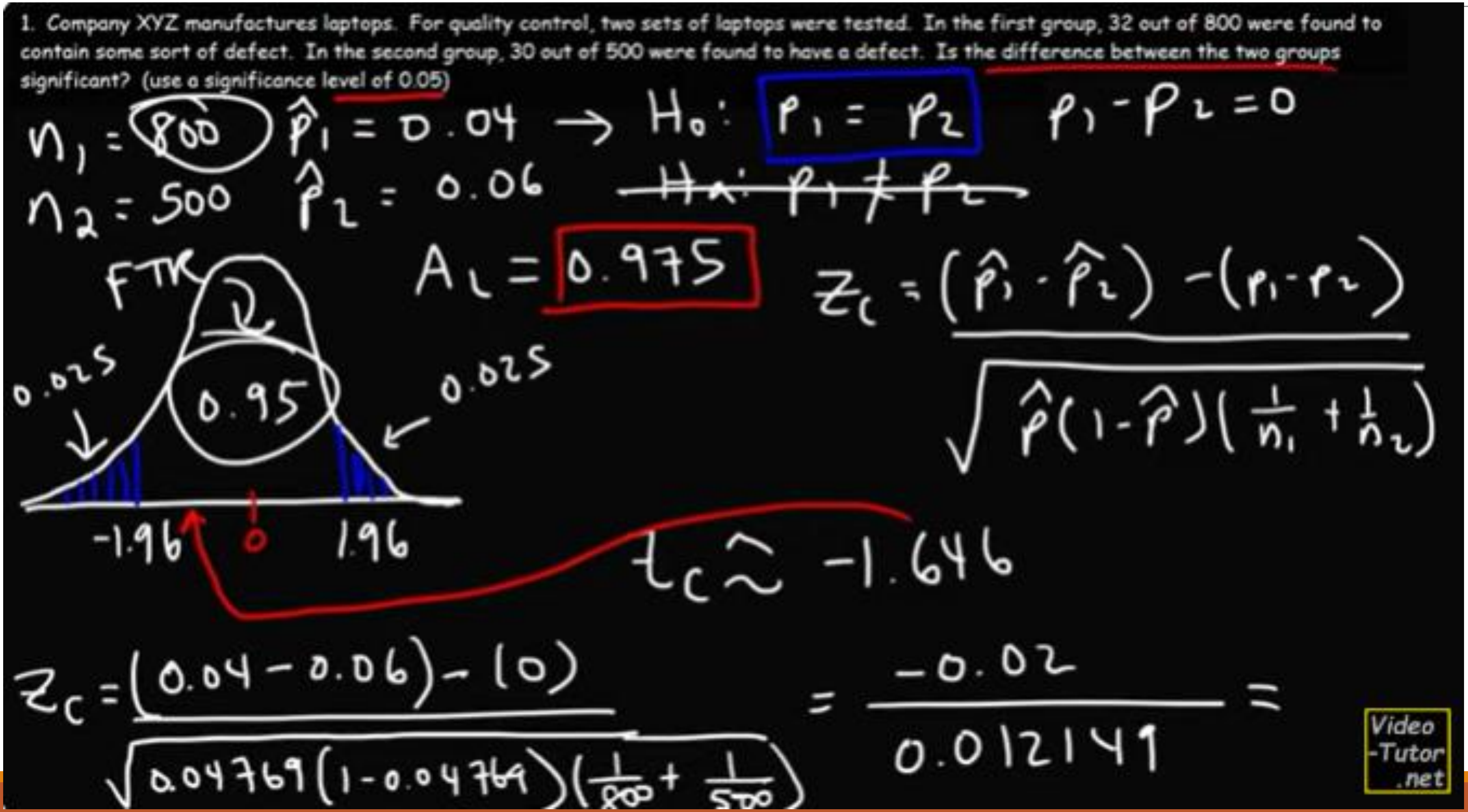
$H_a: p_1 \neq p_2$

$A_L = \boxed{0.975}$

$$Z_c = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$Z_c \approx -1.646$

$$Z_c = \frac{(0.04 - 0.06) - (0)}{\sqrt{0.04769(1-0.04769)\left(\frac{1}{800} + \frac{1}{500}\right)}} = \frac{-0.02}{0.012149} =$$



1. Company XYZ manufactures laptops. For quality control, two sets of laptops were tested. In the first group, 32 out of 800 were found to contain some sort of defect. In the second group, 30 out of 500 were found to have a defect. Is the difference between the two groups significant? (use a significance level of 0.05)

$n_1 = 800$ $\hat{p}_1 = 0.04 \rightarrow H_0: \boxed{p_1 = p_2}$ $p_1 - p_2 = 0$

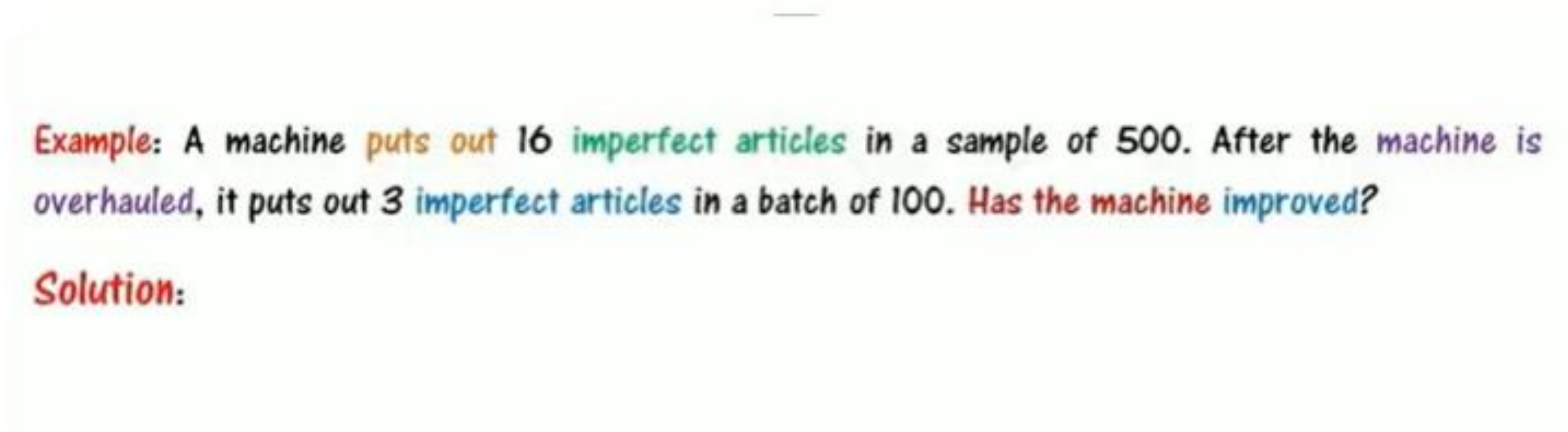
$n_2 = 500$ $\hat{p}_2 = 0.06$ ~~$H_a: p_1 \neq p_2$~~

$A_L = \boxed{0.975}$

$$Z_c = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

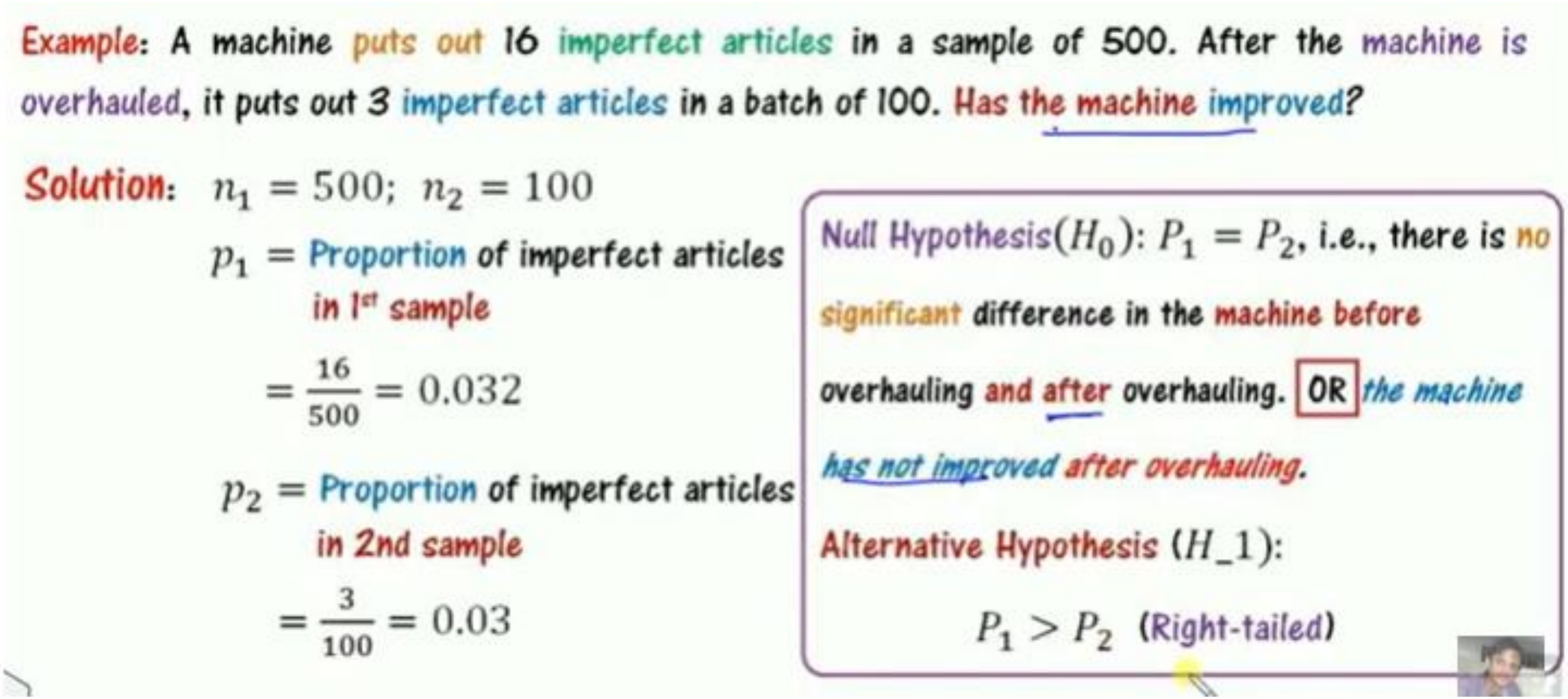
$t_c \approx -1.646$

$$Z_c = \frac{(0.04 - 0.06) - (0)}{\sqrt{0.04769(1-0.04769)\left(\frac{1}{800} + \frac{1}{500}\right)}} = \frac{-0.02}{0.012149} =$$



Example: A machine puts out 16 imperfect articles in a sample of 500. After the machine is overhauled, it puts out 3 imperfect articles in a batch of 100. Has the machine improved?

Solution:



Example: A machine puts out 16 imperfect articles in a sample of 500. After the machine is overhauled, it puts out 3 imperfect articles in a batch of 100. Has the machine improved?

Solution: $n_1 = 500;\ n_2 = 100$

$p_1 =$ Proportion of imperfect articles in 1st sample

$$= \frac{16}{500} = 0.032$$

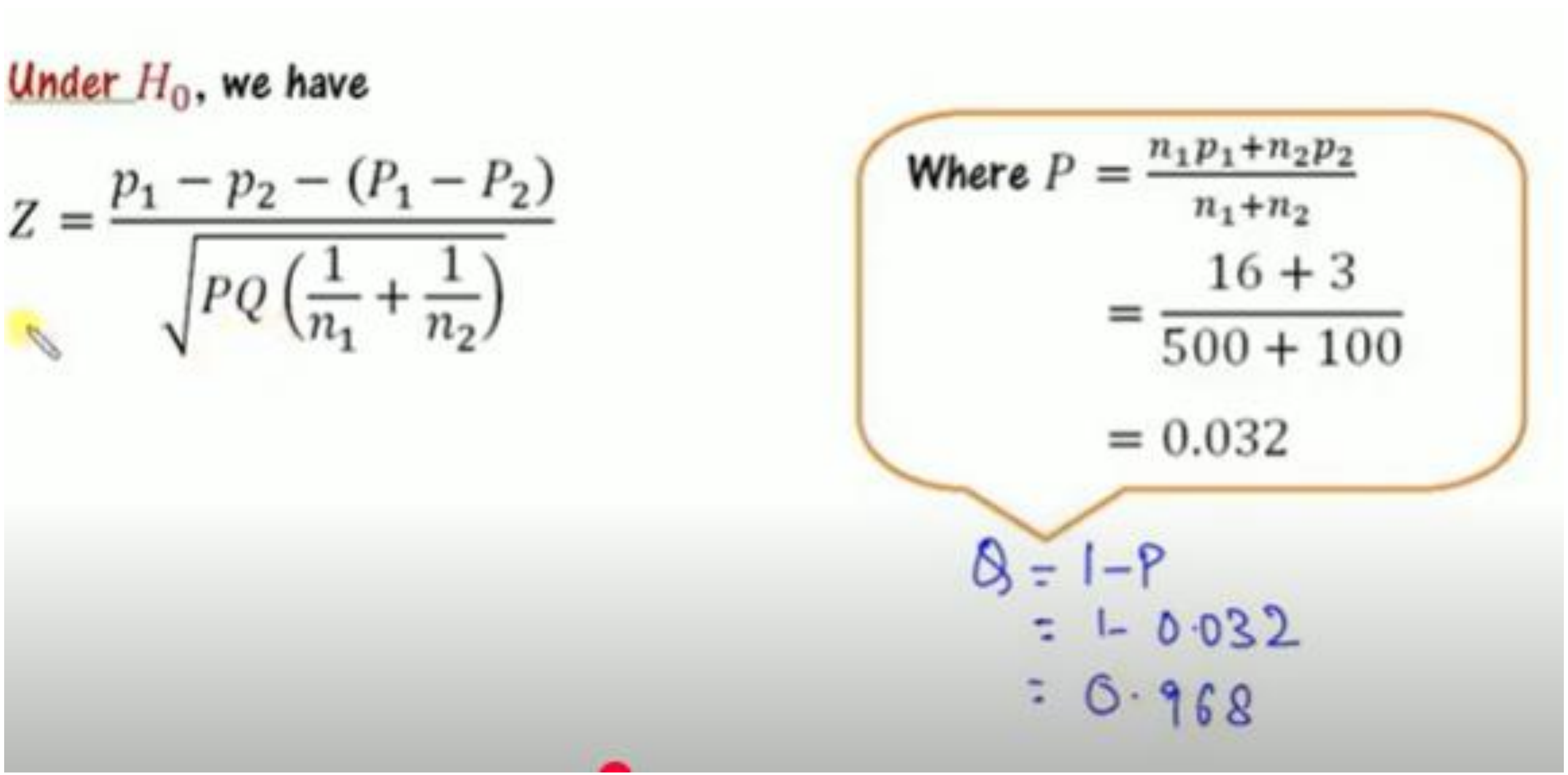
$p_2 =$ Proportion of imperfect articles in 2nd sample

$$= \frac{3}{100} = 0.03$$

Null Hypothesis(H_0): $P_1 = P_2$, i.e., there is no significant difference in the machine before overhauling and after overhauling. OR *the machine has not improved after overhauling*.

Alternative Hypothesis (H_1):

$P_1 > P_2$ (Right-tailed)



Under H_0, we have

$$Z = \frac{p_1 - p_2 - (P_1 - P_2)}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$\text{Where } P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

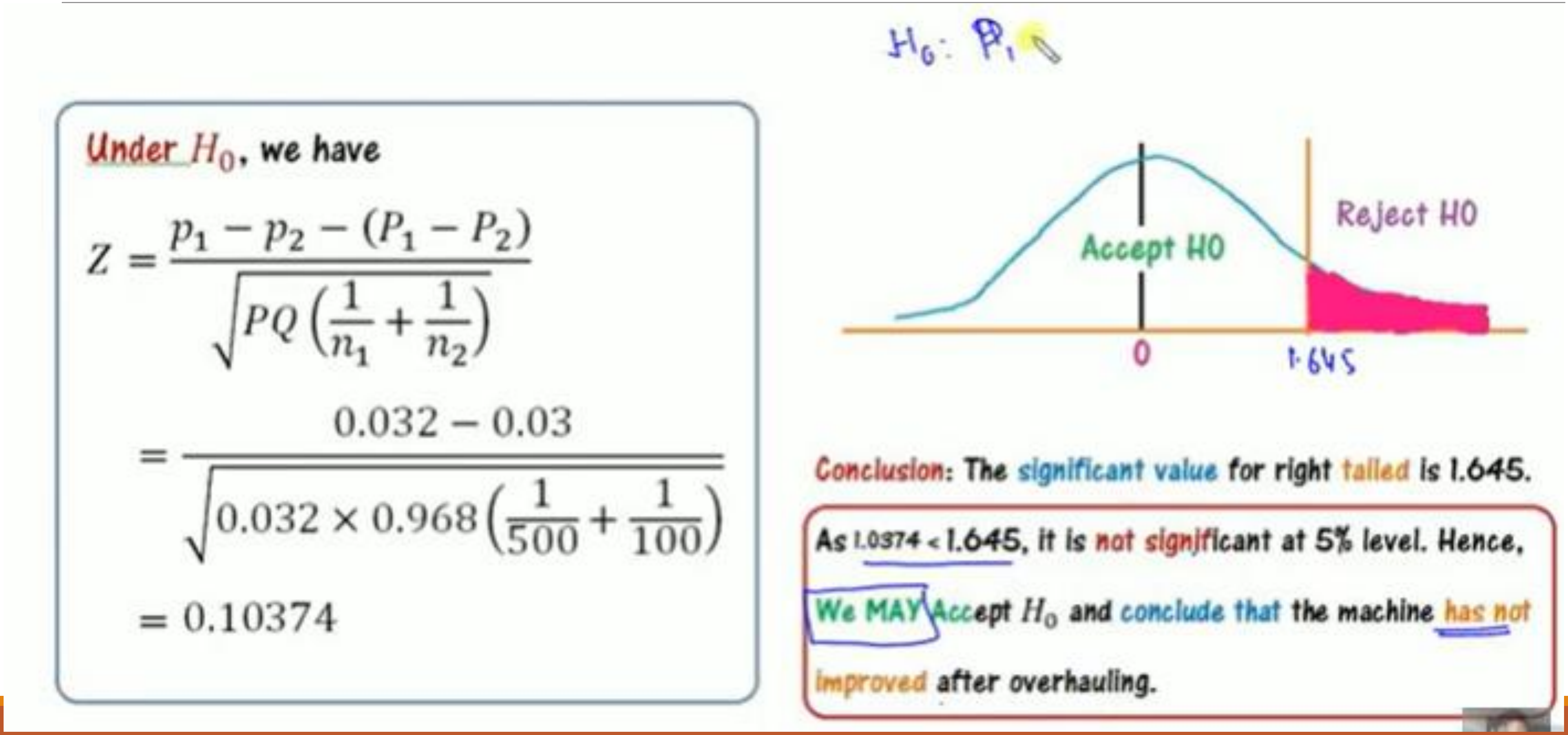
$$= \frac{16 + 3}{500 + 100}$$

$$= 0.032$$

$$Q = 1 - P$$

$$= 1 - 0.032$$

$$= 0.968$$



H_0: P_1

Under H_0, we have

$$Z = \frac{p_1 - p_2 - (P_1 - P_2)}{\sqrt{PQ\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

$$= \frac{0.032 - 0.03}{\sqrt{0.032 \times 0.968\left(\frac{1}{500}+\frac{1}{100}\right)}}$$

$$= 0.10374$$

Conclusion: The significant value for right tailed is 1.645.

As 1.0374 < 1.645, it is not significant at 5% level. Hence, We MAY Accept H_0 and conclude that the machine has not improved after overhauling.