Lecture 15
Confidence interval for proportions

CHOONG-WAN WOO | COCOAN lab | http://cocoanlab.github.io

Review: Key Points

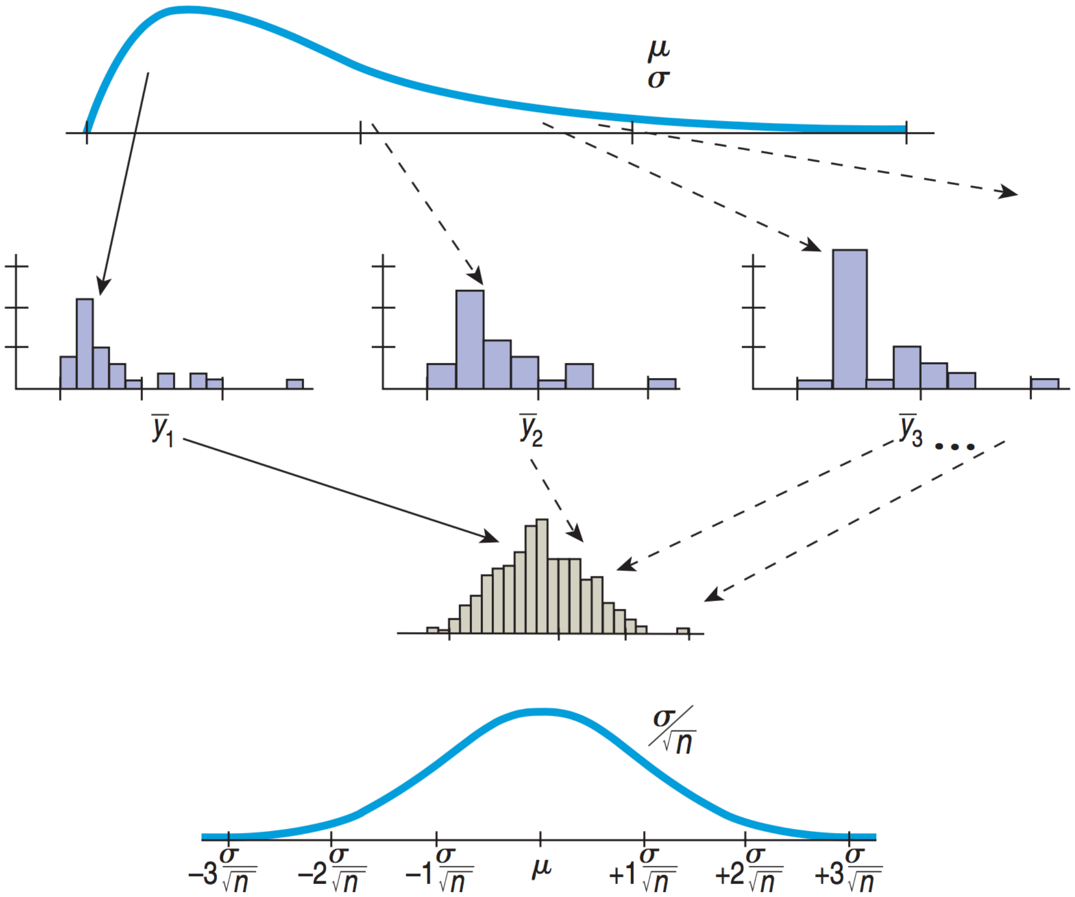
Chapter 18: Sampling Distribution Models

- A parameter is a number that describes the population.
- A statistic is a number that describes a sample.
- The sampling distribution of a statistic is the distribution of its value in all possible samples of the same size from the same population.
- **Central Limit Theorem:** The mean of a random sample is a random variable whose sampling distribution can be approximated by a Normal model. The larger the sample, the better the approximation will be.
- This works no matter what the original data's distribution is.
- The sampling distribution of a proportion can be modeled with a Normal model, $N(p, \sqrt{\frac{pq}{n}})$
- The sampling distribution of a mean can be modeled with a Normal model, $N(\mu, \frac{\sigma}{\sqrt{n}})$

Let's revisit what we learned in the last class

- Population model, that we cannot observe.
- We draw one real sample (solid line) of size n and show its histogram and summary statistics.
- We imagine (or simulate) drawing many other samples (dotted lines).
- We (imagine) gathering all the means into a histogram.
- The CLT tells us we can model the shape of this histogram with a Normal model, $N(\mu, \frac{\sigma}{\sqrt{n}})$

But... can we actually know the mean and SD of the sampling distribution?



- Population model, that we cannot observe.
- We draw one real sample (solid line) of size n and show its histogram and summary statistics.
- We imagine (or simulate) drawing many other samples (dotted lines).

No.. Then, what should we do?

To be continued...

- We (imagine) gathering all the means into a histogram.
- The CLT tells us we can model the shape of this histogram with a Normal model, $N(\mu, \frac{\sigma}{\sqrt{n}})$

An example data on "Facebook use":

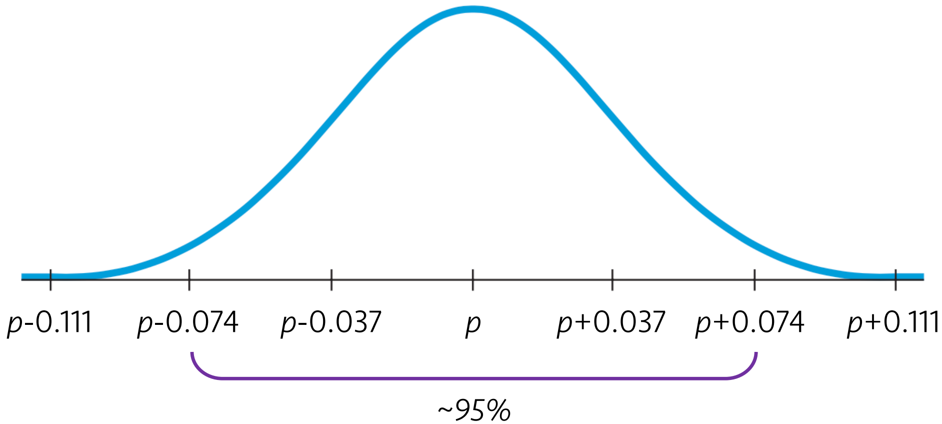
- Late in 2010, Pew Research surveyed US residents about their use of social networking sites.
- Among 156 respondents aged 18-22, 48 said that they update their status at least daily.

- $\hat{p}$ = *48/156 = 30.8%*
- We don't know about *p*. What can we say about the population, *p, with* $\hat{p}$*?*

- What we know: $N(p, \sqrt{\frac{pq}{n}})$
  - The sampling distribution model of $\hat{p}$ is centered at *p*.
  - The standard deviation of the sampling distribution is $\sqrt{\frac{pq}{n}}$

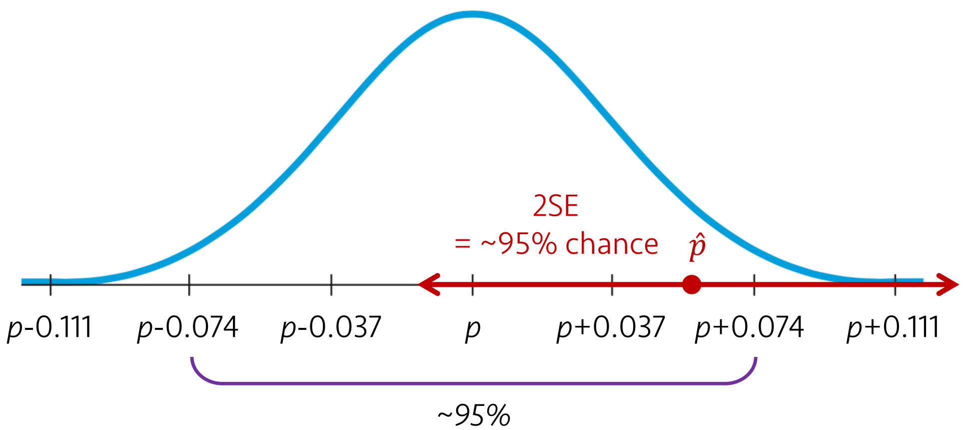
Standard error

- When we estimate the standard deviation of a sampling distribution, we call it a standard error.
- For a sample proportion, $\hat{p}$, the standard error is $SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}$
- For the Facebook users, $SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{0.308 \times 0.692}{156}} = 0.037 = 3.7\%$



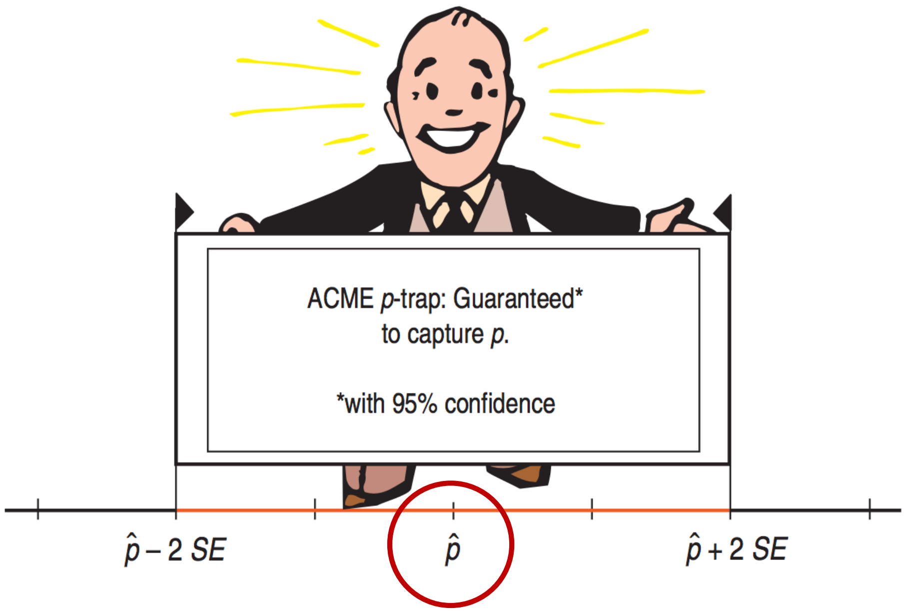
Confidence Interval

- From $\hat{p}$'s point of view:
  - "95% chance that p is no more than 2 SEs away from me." (from 68-95-99.7 rule)
  - "I'm 95% sure that p will be within my 2SE."
  - "Even if my interval catch p, I still don't know its true value, but the best I can do is an interval."



Confidence Interval

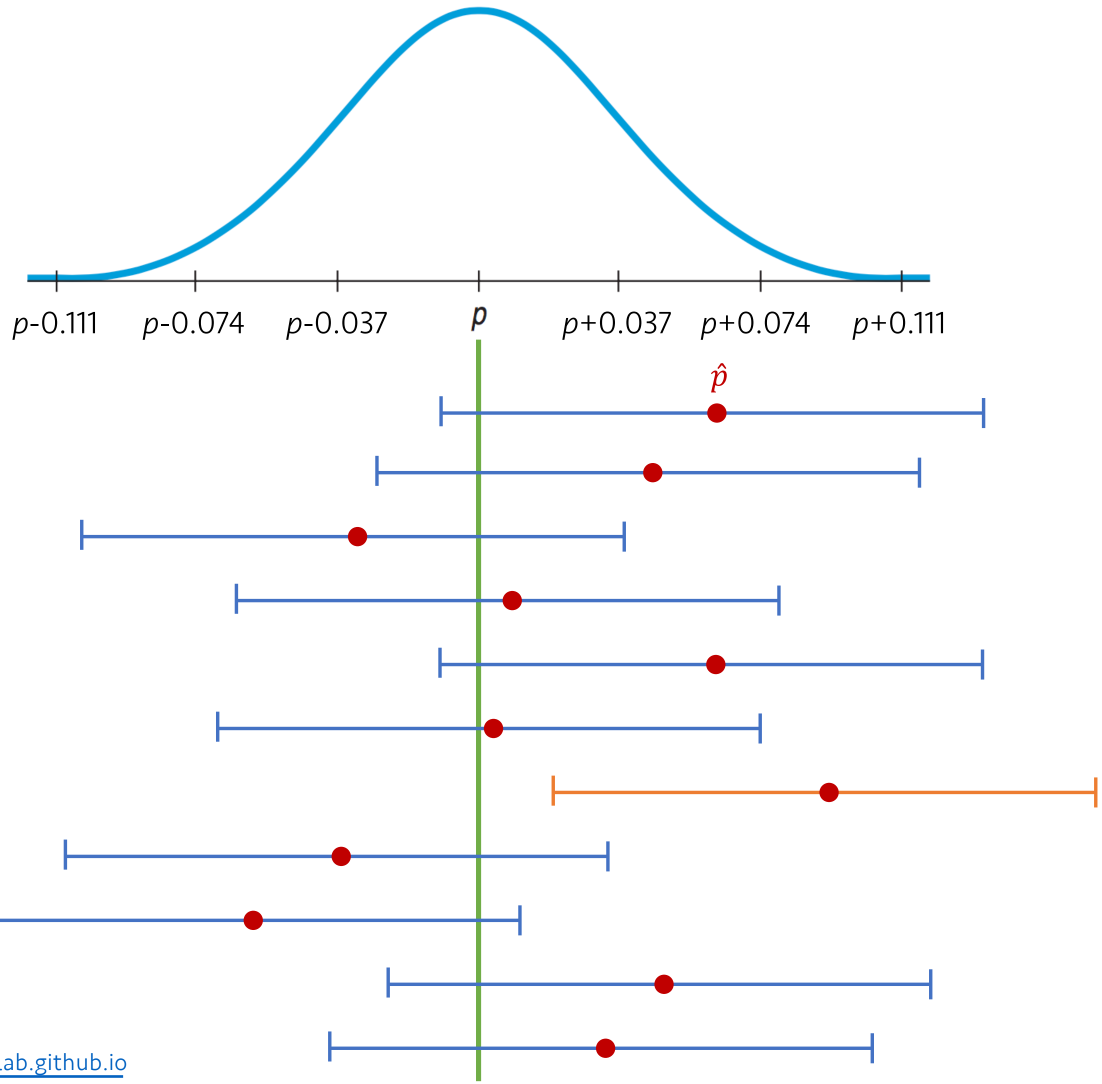
- From $\hat{p}$'s point of view:
  - "95% chance that p is no more than 2 SEs away from me." (from 68-95-99.7 rule)
  - "I'm 95% sure that p will be within my 2SE."
  - "Even if my interval catch p, I still don't know its true value, but the best I can do is an interval."



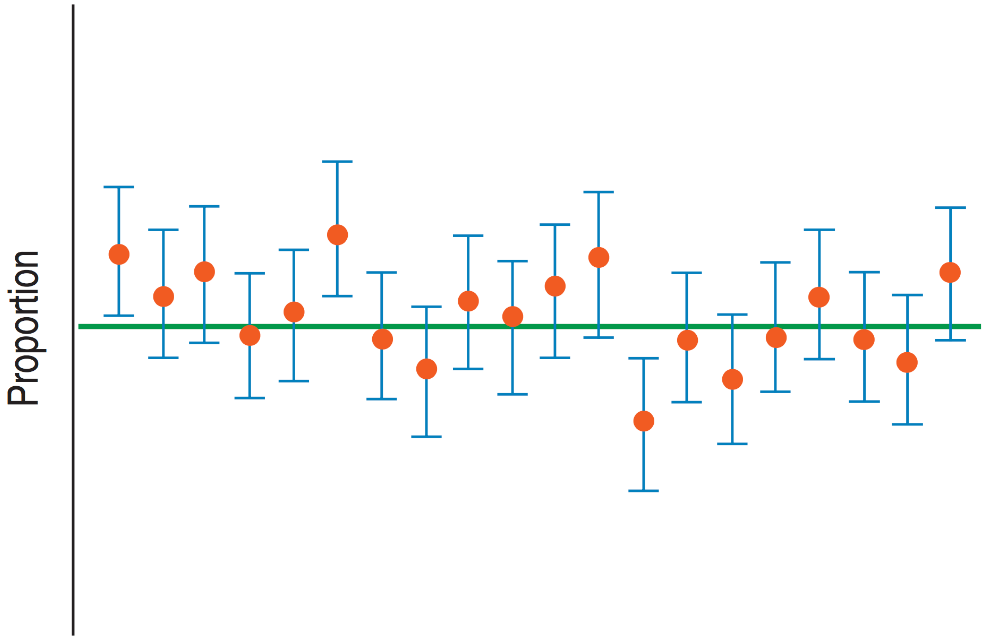
Confidence Interval

- What can we really say about p for our example data?
  - "30.8% of *all* Facebook users between the ages of 18 and 22 update their status daily." NO
  - "It is probably truly that 30.8% of *all* Facebook users between the ages of 18 and 22 update their status daily." NO, we don't know true p.
  - "We don't know exactly what proportion of Facebook users between the ages of 18 and 22 update their status daily, but *we know* that it's within the interval 30.8% $\pm$ 2 $\times$ 3.7% (23.4% to 38.2%)." We're getting closer… but do we know really?
  - "We don't know exactly what proportion of Facebook users between the ages of 18 and 22 update their status daily, but the interval from 23.4% to 38.2% *probably* contains the true proportion." We're getting closer… but can we more specific?
  - "We are 95% confident that between 23.4% and 38.2% of Facebook users between the ages of 18 and 22 update their status at least daily." OKAY, finally!

Confidence Interval



Confidence Interval



Confidence Interval

- Each sample proportion can be used to make a confidence interval.
- The **Central Limit Theorem** assures us that (in the long run) 95% of the intervals cover the true value, and only 5% are wrong.

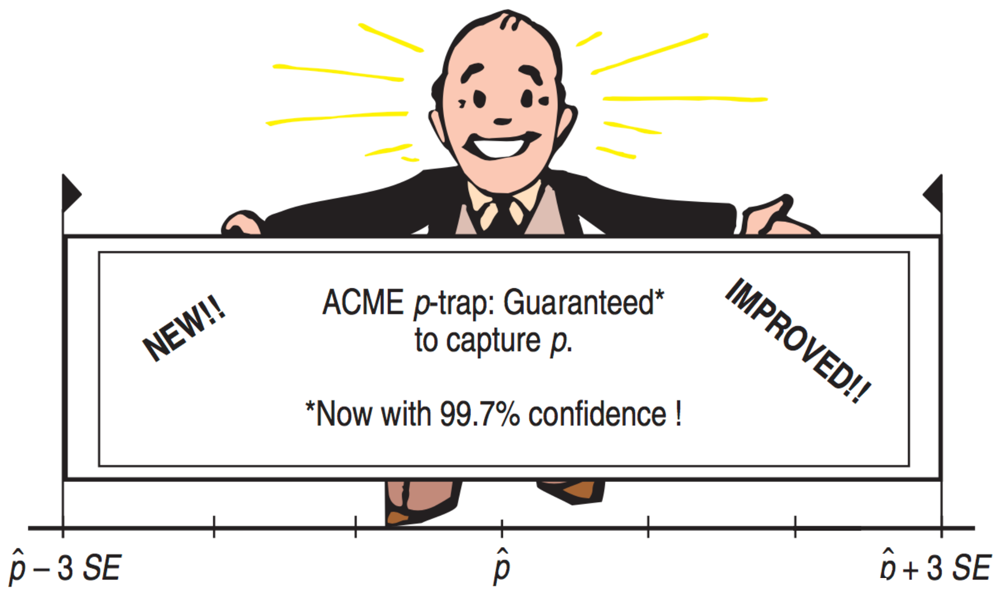
- The confidence intervals are random because they are based on random samples.
- Our confidence (and our uncertainty) is about the interval, not the true proportion.

Certainty vs. Precision

- Our confidence interval has this form: $\hat{p} \pm 2SE(\hat{p})$
  - Here, $2SE(\hat{p})$ is called the margin of error (ME).
  - Any population parameter (proportion, mean, regression slope, etc.) can be estimated with some margin of error.

General form: Estimate $\pm$ *ME*

- ME can be
  - 95% confidence interval = 2SE
  - 99.7% confidence interval = 3SE



Certainty vs. Precision

- Our confidence interval has this form: $\hat{p} \pm 2SE(\hat{p})$
  - Here, $2SE(\hat{p})$ is called the margin of error (ME).
  - Any population parameter (proportion, mean, regression slope, etc.) can be estimated with some margin of error.

$$\text{General form: Estimate} \pm ME$$

- ME can be
  - 95% confidence interval = 2SE
  - 99.7% confidence interval = 3SE
  - The more confident we want to be, the larger the margin of error must be.
  - E.g., 100% confidence: the proportion of Facebook users who update daily is between 0 and 100%
    - *Is this useful?* NO
  - Or we can give a very narrow interval (e.g., 30.7%-30.9%) with very low confidence.
    - *is this useful?* Maybe not.
- **Every confidence interval is a balance between certainty and precision.**
  - 90%, 95%, 99% are commonly used.

Critical values

- Critical value = the *number* of SEs (e.g., 2 in 2SEs)
- Denoted as z^*
- For 95% confidence interval, the precise critical value is z^* = 1.96 (though we used 2 to make it simple).

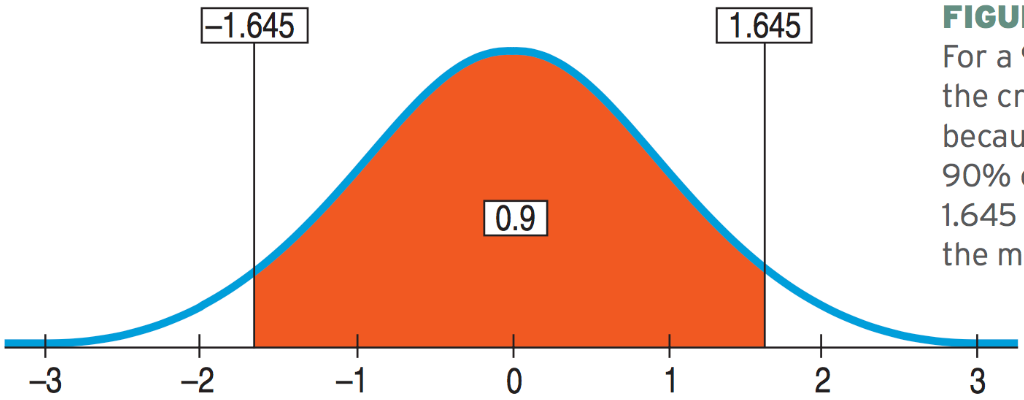


FIGURE 19.4
For a 90% confidence interval, the critical value is 1.645, because, for a Normal model, 90% of the values are within 1.645 standard deviations from the mean.

Assumptions and Conditions

Assumptions and Conditions

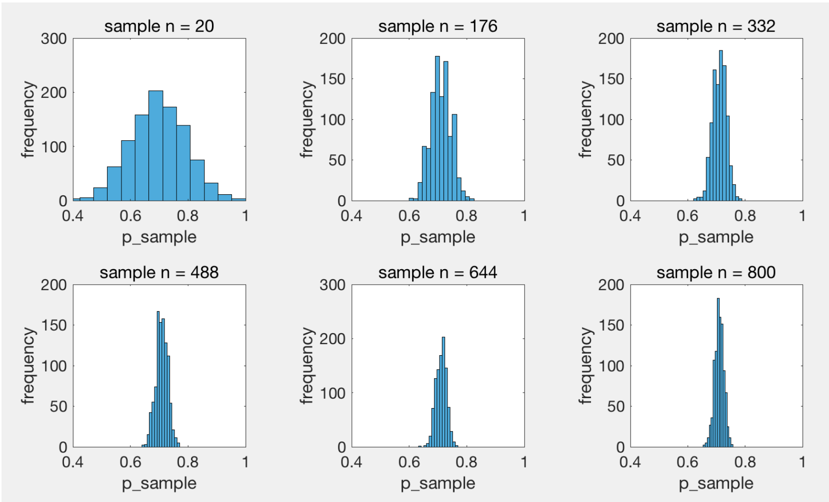
- **Independence Assumption:** The individuals in the samples must be independent of each other.

- We can't know if this assumption is true or not for sure, but we can check the following *conditions* that provide information about the assumption.

- Conditions:
  - **Randomization Condition:** random assignment (experiment), random sampling (survey)
  - **10% Condition:** The sample size, n, must be no larger than 10% of the population. If you sample more than about 10% of the population, the remaining individuals are no longer truly independent of each other.
  - **Success/Failure Condition:** The sample size has to be big enough so that we expect at least 10 successes and at least 10 failures (np and nq should be > 10)

Choosing your sample size

- $ME = z^* \sqrt{\frac{\hat{p}\hat{q}}{n}}$
- If we want 3% Margin of error (ME) for 95% confidence ($z^* = 1.96$)
- $0.03 = 1.96 \times \sqrt{\frac{\hat{p}\hat{q}}{n}}$
- If you don't know $\hat{p}$, just take its maximum: $\hat{p} = 0.5$, then $\hat{p}\hat{q} = 0.25$
- $\sqrt{n} = \frac{1.96\sqrt{0.5 \times 0.5}}{0.03} \approx 32.67$
- $n \approx (32.67)^2 \approx 1067.1$
- Then, our sample size should be 1068 (rounding up).



Key Points

Chapter 19: Confidence Interval for Proportions

- **Standard error:** standard deviation of a sampling distribution
- For a sample proportion, $\hat{p}$, the standard error is $SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}$
- 95% confidence interval for a proportion, $\hat{p} \pm 1.96\ SE(\hat{p})$
- General form: *Estimate* $\pm$ *Margin of Error (ME)*
- Critical value, z^* = the *number* of SEs (e.g., 2 in 2SEs)
- Every confidence interval is a balance between certainty and precision.
- You can choose your sample size based on confidence interval.