# Confidence Intervals for a Sample Proportion

August 20, 2019

# Midterm Scores

One of your observant peers caught a typo on my exam key! Exam grades have been updated in iLearn.

# Office Hours

Today's office hours are from 12-2 PM.

# A Note on Standard Error

Recall that standard error is closely related to both standard deviation and sample size. In fact,

$$SE = \frac{sd}{\sqrt{n}}$$

This is true regardless of the population parameter of interest.

# Confidence Intervals

- $\hat{p}$ is a single plausible value for the population proportion $p$.
- But there is always some standard error associated with $\hat{p}$.
- We want to be able to provide a plausible range of values instead.

# A Range of Values is Like a Net

- A point estimate is like spear fishing in murky waters.
- Chances are we'll miss our fish.
- A range of values is like casting a net.
- Now we have a much higher chance of catching our fish.

This range of values is called a **confidence interval**.

# Confidence Intervals

The idea behind a confidence interval is

- Building an interval related to $\hat{p}$
- This interval captures a range of plausible values.
- With more values come more opportunities to capture the true population parameter.

# Confidence Intervals

If we want to be very certain that we capture the population parameter, should we use a wider or a smaller interval?

# 95% Confidence Intervals

- Based on our sample, $\hat{p}$ is the most plausible value for $p$.
- Therefore will build our confidence interval around $\hat{p}$.
- The standard error will act as a guide for how large to make the interval.

# 95% Confidence Intervals

- When the Central Limit Theorem conditions are satisfied, the point estimate comes from a normal distribution.
- For a normal distribution, 95% of the data is within $|Z| = 1.96$ standard deviations of the mean.
- Our confidence interval will extend 1.96 standard errors from the sample proportion.

# 95% Confidence Intervals

Putting these together, we can be 95% confidence that the following interval captures the population proportion:

$$\text{point estimate} \pm 1.96 \times SE$$
$$\hat{p} \pm 1.96 \times \sqrt{\frac{p(1-p)}{n}}$$

# 95% Confidence Intervals

In this interval, the upper bound is

$$\hat{p} + 1.96 \times \sqrt{\frac{p(1-p)}{n}}$$

and the lower bound is

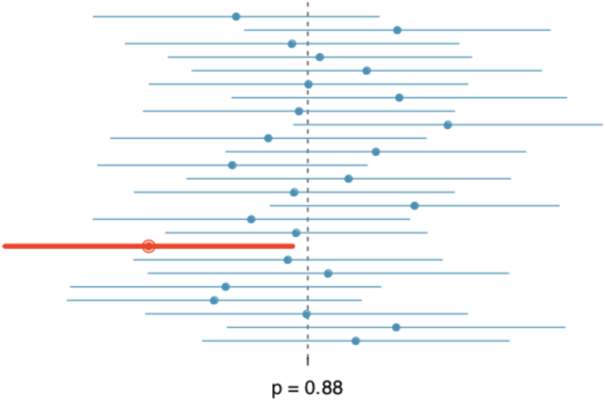
$$\hat{p} - 1.96 \times \sqrt{\frac{p(1-p)}{n}}$$

# 95% Confidence Intervals

What does 95% confident mean?

- Confidence is based on the concept of *repeated sampling*.
- Suppose we took 1000 samples and built a 95% confidence interval from each.
- Then about 95% of these would contain the true parameter $p$.

# 95% Confidence Intervals




25 confidence intervals built from 25 samples where the true proportion is $p = 0.88$. Only one of these did not capture the true proportion.

# Example

Last class we talked about a sample of 1000 Americans where 88.7% said that they supported expanding solar power.

Find a 95% confidence interval for $p$.

# Example

We decided during our last class that the Central Limit Theorem applies and that

$$\mu_{\hat{p}} = \hat{p} = 0.887$$

and

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.010$$

# Example

Plugging these into our confidence interval,

$$\begin{aligned} &\hat{p} \pm 1.96 \times SE_{\hat{P}} \\ &\to 0.887 \pm 1.96 \times 0.010 \\ &\to 0.887 \pm 0.0196 \\ &\to (0.8674, 0.9066) \end{aligned}$$

We can be 95% confident that the actual proportion of adults who support expanding solar power is between 86.7% and 90.7%.

# More General Confidence Intervals

- Suppose we want to cast a wider net and find a 99% confidence interval.
- To do so, we must widen our 95% confidence interval.
- If we wanted a 90% confidence interval, we would need to narrow our 95% interval.

# More General Confidence Intervals

We decided that the 95% confidence interval for a point estimate that follows the Central Limit Theorem is

$$\text{point estimate} \pm 1.96 \times SE$$

There are three components to this interval:

1. the point estimate
2. "1.96"
3. the standard error

# More General Confidence Intervals

- The point estimate and standard error won't change if we change our confidence level.
- 1.96 was based on capturing 95% of the data for our normal distribution.
- We will need to adjust this value for other confidence levels.

# Consider the Following

If $X$ is a normally distributed random variable, what is the probability of the value $x$ being within 2.58 standard deviations of the mean?

# Consider the Following

We want to know how often the Z-score will be between -2.58 and 2.58:

$$\begin{aligned} P(-2.58 < Z < 2.58) &= P(Z < 2.58) - P(Z < -2.58) \\ &= 0.9951 - 0.0049 \\ &\approx 0.99 \end{aligned}$$

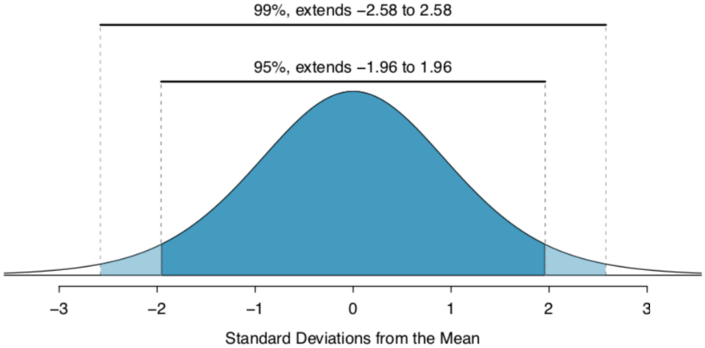
So there is a 99% probability that $X$ will be within 2.58 standard deviations of $\mu$

# 99% Confidence Intervals

With this in mind, we can create a 99% confidence interval:

$$\text{point estimate} \pm 2.58 \times SE$$

All we needed to do was change 1.96 in the 95% confidence interval formula to 2.58.




Crucially, the area between $-z_{\alpha/2}$ and $z_{\alpha/2}$ increases as $z_{\alpha/2}$ becomes larger.

# What is $\alpha$?

For now, we will think of $\alpha$ (Greek letter alpha) as the chance that $p$ is *not* in our interval.

$$\alpha = 1 - \text{confidence level}$$

We call $\alpha$ the **level of significance**.

# What is $\alpha$?

We can rework our formula for $\alpha$ to say that our confidence level is

$$1 - \alpha$$

as a proportion, or

$$(1 - \alpha) \times 100\%$$

as a percent.

Over the next few slides, we will consider why we use the notation $z_{\alpha/2}$.

# General Confidence Intervals

- Using Z-scores and the normal model is appropriate when our point estimate is associated with a normal model.
- This is true when
  1. our point estimate is the mean of a variable that is itself normally distributed
  2. the Central Limit Theorem holds for our point estimate

When a normal model is not a good fit, we will use alternative distributions. These will come up in later chapters.

# General Confidence Intervals

If a point estimate closely follows a normal model with standard error $SE$, then a confidence interval for the population parameter is

$$\text{point estimate} \pm z_{\alpha/2} \times SE$$

where $z_{\alpha/2}$ corresponds to the desired confidence level.

# General Confidence Intervals

In this general setting, the upper bound for the interval is

$$\text{point estimate} + z_{\alpha/2} \times SE$$

and the lower bound is

$$\text{point estimate} - z_{\alpha/2} \times SE$$

# Margin of Error

In a confidence interval,

$$\text{point estimate} \pm z_{\alpha/2} \times SE,$$

we refer to $z_{\alpha/2} \times SE$ as the **margin of error**.

# Margin of Error

- The margin of error is the maximum amount of error that we allow from the point estimate.
- That is, this is the furthest distance from the point estimate that we consider to be plausible.
- We expect the true parameter to be within this error, limited by the confidence level.

# Margin of Error

Margin of error will *decrease* when

- $n$ increases.
- $1 - \alpha$ decreases.
- $\alpha/2$ increases.
- $z_{\alpha/2}$ decreases.

Margin of error will increase under opposite conditions.

# Critical Value

In a confidence interval,

$$\text{point estimate} \pm z_{\alpha/2} \times SE,$$

we refer to $z_{\alpha/2}$ as the **critical value**.

# Finding $z_{\alpha/2}$

We want to select $z_{\alpha/2}$ so that the area between $-z_{\alpha/2}$ and $z_{\alpha/2}$ in the standard normal distribution, $N(0,1)$, corresponds to the confidence level.

Let $c$ be the desired confidence level. We want to find $z_{\alpha/2}$ such that

$$c = P(-z_{\alpha/2} < Z < z_{\alpha/2})$$

# Finding $z_{\alpha/2}$

Rewriting this,

$$\begin{aligned} c &= P(-z_{\alpha/2} < Z < z_{\alpha/2}) \\ &= 1 - P(Z > z_{\alpha/2}) - P(Z < -z_{\alpha/2}) \end{aligned}$$

Since $Z \sim N(0,1)$ is symmetric,

$$P(Z > z_{\alpha/2}) = P(Z < -z_{\alpha/2})$$

# Finding $z_{\alpha/2}$

So

$$\begin{aligned}
c &= P(-z_{\alpha/2} < Z < z_{\alpha/2}) \\
&= 1 - P(Z > z_{\alpha/2}) - P(Z < -z_{\alpha/2}) \\
&= 1 - P(Z < -z_{\alpha/2}) - P(Z < -z_{\alpha/2}) \\
&= 1 - 2P(Z < -z_{\alpha/2})
\end{aligned}$$

# Finding $z_{\alpha/2}$

Solving for $P(Z < -z_{\alpha/2})$, we find

$$\frac{1-c}{2} = \frac{\alpha}{2} = P(Z < -z_{\alpha/2})$$

Hence $z_{\alpha/2}$!

Since $c$ is some number, say 0.90 (a 90% confidence level), we now have an easy way to find $z_{\alpha/2}$!

# Example: Finding $z_{\alpha/2}$

Suppose you want to find a 99% confidence interval. Find $z_{\alpha/2}$.

We know that

$$\frac{1-c}{2} = P(Z < -z_{\alpha/2})$$

and that a 99% confidence level translates to $c = 0.99$.

# Example: Finding $z_{\alpha/2}$

So

$$P(Z < -z_{\alpha/2}) = \frac{1-c}{2}$$
$$= \frac{1-0.99}{2}$$
$$= 0.005$$

Using software to find this percentile, $-z_{\alpha/2} = -2.58$ (so $z_{\alpha/2} = 2.58$). This is what the textbook told us earlier!

# Example

Recall our sample of 1000 adults, 88.7% of whom were found to support the expansion of solar energy. Find a 90% confidence interval for the proportion. Note that we have already verified conditions for normality.

First, our point estimate is $\hat{p} = 0.887$.

# Example

Now we need to find $z_{\alpha/2}$. Our confidence level is $c = 0.90$.

$$\begin{aligned} P(Z < -z_{\alpha/2}) &= \frac{1-c}{2} \\ &= \frac{1-0.9}{2} \\ &= 0.05 \end{aligned}$$

Using `R`, we find $-z_{\alpha/2} = -1.65$ (so $z_{\alpha/2} = 1.65$).

# Example

Then the 90% confidence interval can be computed as

$$\hat{p} \pm 1.65 \times SE \quad \longrightarrow \quad 0.887 \pm 1.65 \times 0.010$$

which is the interval $(0.8705, 0.9035)$.

Thus we are 90% confident that 87.1% to 90.4% of American adults support the expansion of solar power.

# Confidence Interval for a Single Proportion

There are four steps to constructing these confidence intervals:

1. Identify $\hat{p}$, $n$, and the desired confidence level.
2. Verify that $\hat{p}$ is approximately normal
   - Use the success-failure condition with $\hat{p}$ to verify the Central Limit Theorem.
3. Compute $SE$ using $\hat{p}$ and find $z_{\alpha/2}$, using these values to construct your interval.
4. Interpret your confidence interval *in the context of the problem.*

# Example: Ebola

After a doctor contracted Ebola in New York City, a poll of 1042 New Yorkers found that 82% were in favor of a mandatory quarantine for anyone who'd come in contact with with an Ebola patient.

We will walk through developing and interpreting a 95% confidence interval for the proportion of New Yorkers who favor mandatory quarantine.

# Example: Ebola

First, we need to find the point estimate and confirm that a normal model is appropriate.

$$\hat{p} = 0.82$$

This is the given proportion of polled New Yorkers who favored mandatory quarantine.

# Example: Ebola

To confirm that a normal model is appropriate, we check our success-failure condition using the plug-in approach:

$$n\hat{p} = 1042 \times 0.82 = 853.62 \geq 10$$

and

$$n(1-\hat{p}) = 1042 \times (1-0.82) = 187.38 \geq 10$$

# Example: Ebola

Since the normal model is appropriate, we can move on to calculating the standard error for $\hat{p}$ based on the Central Limit Theorem. We will again use the plug-in approach.

$$SE_{\hat{p}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.82(1-0.82)}{1041}} = 0.012$$

# Example: Ebola

Now we want to find our critical value $z_{\alpha/2}$ for our 95% confidence interval. In this case,

$$\alpha = 1 - \text{confidence level} = 0.05$$

# Example: Ebola

Then, using software, $z_{\alpha/2} = z_{0.025} = 1.96$ and our confidence interval is

$$\begin{aligned}\hat{p} \pm z_{\alpha/2} \times SE &= 0.82 \pm 1.96 \times 0.012 \\ &= 0.82 \pm 0.0235\end{aligned}$$

or $(0.796, 0.844)$.

# Example: Ebola

Finally, to interpret the interval $(0.796, 0.844)$:

We can be 95% confident that the proportion of New York adults in October 2014 who supported a quarantine for anyone who had come into contact with an Ebola patients was between 0.796 and 0.844.

# Example: Ebola

When we say that we are 95% confident, we mean:

If we took many such samples and computed a 95% confidence interval for each

- About 95% of those intervals would contain the actual proportion.
- This proportion is of New York adults who supported a quarantine for anyone who has come into contact with an Ebola patient.

# Interpreting Confidence Intervals

Whenever we interpret a confidence interval,

1. The statement should be about the population parameter of interest.
2. We do *not* want to talk about the probability that that interval captures the population parameter.
    - This is an important technical detail that has to do with our definition of "95% confident".

# Interpreting Confidence Intervals

Whenever we interpret a confidence interval,

3. The confidence interval says nothing about individual observations or point estimates.
4. These methods apply to sampling error and ignore bias entirely!
    - If we are systematically over- or under-estimating, confidence intervals will not address this problem.

# Example: Interpreting Confidence Intervals

Consider the 90% confidence interval for the solar energy survey: 87.1% to 90.4%. If we ran the survey again, can we say that we're 90% confident that the new survey's proportion will be between 87.1% and 90.4%?

# Example: Interpreting Confidence Intervals

No! Confidence intervals don't tell us anything about future point estimates.

Our point estimate will change so our confidence interval will change.

# Sample Size Calculation

Exactly how many observations do we need to get an accurate estimate?

# Example: Sample Size Calculation

Suppose a manufacturer claims that he is 95% confident that the proportion of defective units coming from his factory is 2%. We want to examine this claim at a margin of error no greater than 0.5%. How many samples do we need?

# Example: Sample Size Calculation

For our proportion, we will consider a Bernoulli distribution with $p = 0.02$. We will calculate the $n$ for this distribution. Then

$$\mu = p = 0.02$$

and

$$sd = \sqrt{p(1-p)} = \sqrt{0.02 \times 0.98} = 0.14$$

# Example: Sample Size Calculation

The margin of error (MoE) is

$$\begin{aligned}
\text{MoE} &= z_{\alpha/2} \times SE \\
&= z_{0.05/2} \times \frac{sd}{\sqrt{n}} \\
&= 1.96 \times \frac{0.14}{\sqrt{n}}
\end{aligned}$$

# Example: Sample Size Calculation

Note that this is a 95% confidence claim and we want the margin of error (MoE) to be $\leq 0.005$. So

$$0.005 \geq MoE$$
$$0.005 \geq 1.96 \times \frac{0.0196}{\sqrt{n}}$$

# Example: Sample Size Calculation

Solving for $n$,

$$n \geq \left(1.96 \times \frac{0.0196}{0.005}\right)^2 = 59.032$$

- Since $n \geq 59.032$ and we need a whole number of samples, we will always round up!
- We will need at least 60 samples to achieve a margin of error of no more than 0.5%.

# Sample Size Calculations

In general, for a confidence interval,

$$n \geq \left( z_{\alpha/2} \times \frac{sd}{MoE} \right)^2$$

where $MoE$ is the desired maximum margin of error. We will always round $n$ up to the nearest integer.