

Chapter 5

Foundations for inference¹

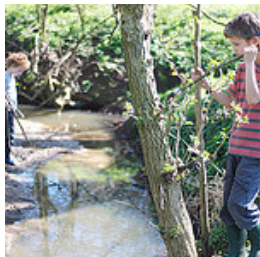
Department of Mathematics & Statistics
North Carolina A&T State University

¹These notes use content from OpenIntro Statistics Slides by Mine Cetinkaya-Rundel.

Confidence intervals for a proportion

Confidence intervals

- ▶ A plausible range of values for the population parameter is called a **confidence interval**.
- ▶ Using only a sample statistic to estimate a parameter is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net.



We can throw a spear where we saw a fish but we will probably miss. If we toss a net in that area, we have a good chance of catching the fish.



- ▶ If we report a point estimate, we probably won't hit the exact population parameter. If we report a range of plausible values we have a good shot at capturing the parameter.

Facebook's categorization of user interests

Most commercial websites (e.g. social media platforms, news outlets, online retailers) collect a data about their users' behaviors and use these data to deliver targeted content, recommendations, and ads. To understand whether Americans think their lives line up with how the algorithm-driven classification systems categorizes them, Pew Research asked a representative sample of 850 American Facebook users how accurately they feel the list of categories Facebook has listed for them on the page of their supposed interests actually represents them and their interests. 67% of the respondents said that the listed categories were accurate. Estimate the true proportion of American Facebook users who think the Facebook categorizes their interests accurately.

Facebook's categorization of user interests

$$\hat{p} = 0.67 \quad n = 850$$

Facebook's categorization of user interests

$$\hat{p} = 0.67 \quad n = 850$$

The approximate 95% confidence interval is defined as

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

Facebook's categorization of user interests

$$\hat{p} = 0.67 \quad n = 850$$

The approximate 95% confidence interval is defined as

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

$$SE = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.67 \times 0.33}{850}} \approx 0.016$$

Facebook's categorization of user interests

$$\hat{p} = 0.67 \quad n = 850$$

The approximate 95% confidence interval is defined as

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

$$SE = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.67 \times 0.33}{850}} \approx 0.016$$

$$\begin{aligned}\hat{p} \pm 1.96 \times SE &= 0.67 \pm 1.96 \times 0.016 \\ &= (0.67 - 0.03, 0.67 + 0.03) \\ &= (0.64, 0.70)\end{aligned}$$

Practice

Which of the following is the correct interpretation of this confidence interval?

We are 95% confident that

- A) 64% to 70% of American Facebook users in this sample think Facebook categorizes their interests accurately.
- B) 64% to 70% of all American Facebook users think Facebook categorizes their interests accurately.
- C) There is a 64% to 70% chance that a randomly chosen American Facebook user's interests are categorized accurately.
- D) There is a 64% to 70% chance that 95% of American Facebook users' interests are categorized accurately.

Practice

Which of the following is the correct interpretation of this confidence interval?

We are 95% confident that

- A) 64% to 70% of American Facebook users in this sample think Facebook categorizes their interests accurately.
- B) 64% to 70% of all American Facebook users think Facebook categorizes their interests accurately.
- C) There is a 64% to 70% chance that a randomly chosen American Facebook user's interests are categorized accurately.
- D) There is a 64% to 70% chance that 95% of American Facebook users' interests are categorized accurately.

What does 95% confident mean?

- ▶ Suppose we took many sample and built a confidence interval from each sampling using the equation

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

- ▶ Then about 95% of those intervals would contain the true population proportion (p).

Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

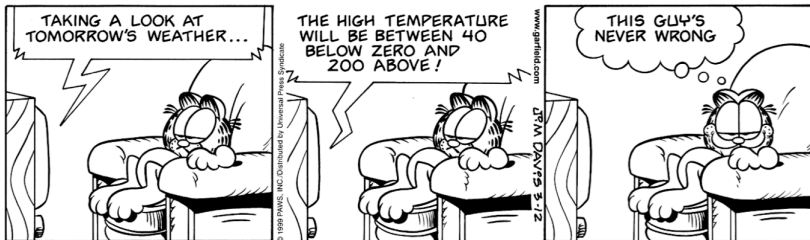
A wider interval.

Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

A wider interval.

Can you see any drawbacks to using a wider interval?

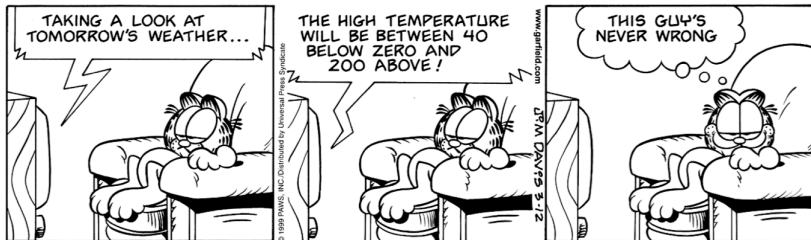


Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

A wider interval.

Can you see any drawbacks to using a wider interval?



If the interval is too wide it may not be very informative.

Changing the confidence level

$$\text{point estimate} \pm z^* \times SE$$

- ▶ In a confidence interval, $z^* \times SE$ is called the **margin of error**, and for a given sample, the margin of error changes as the confidence level changes.
- ▶ In order to change the confidence level we need to adjust z^* in the above formula.
- ▶ Commonly used confidence levels in practice are 90%, 95%, 98% and 99%.
- ▶ For a 95% confidence interval, $z^* = 1.96$.
- ▶ However, using the standard normal (z) distribution, it is possible to find the appropriate z^* for any confidence level.

Practice

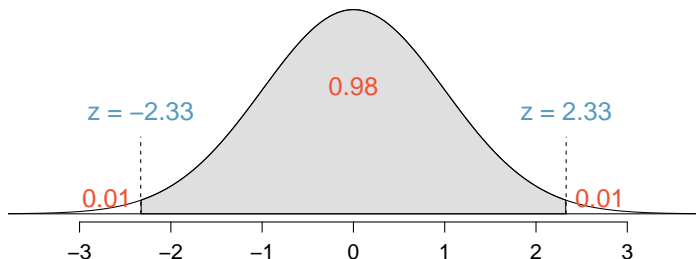
Which of the below Z scores is the appropriate z^* when calculating a 98% confidence interval?

A) $Z = 2.05$ B) $Z = 1.96$ C) $Z = 2.33$ D) $Z = -2.33$ E) $Z = -1.65$

Practice

Which of the below Z scores is the appropriate z^* when calculating a 98% confidence interval?

- A) $Z = 2.05$ B) $Z = 1.96$ C) $Z = 2.33$ D) $Z = -2.33$ E) $Z = -1.65$



where z is obtained from the standard normal table or RStudio

Interpreting confidence intervals

Confidence intervals are ...

- ▶ Always about the population.
- ▶ Not probability statements.
- ▶ Only about population parameters, not individual observation.
- ▶ Only reliable if the sample statistic they're based on is an unbiased estimator of the population parameter.