Section 5.2
Confidence Intervals for a Proportion

Stats 7 Summer Session II 2022

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.

Photos by Mark Fischer (http://www.flickr.com/photos/fischerfotos/7439791462)

and Chris Penny (http://www.flickr.com/photos/clearlydived/7029109617) on Flickr.

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.

https://www.pewinternet.org/2019/01/16/facebook-algorithms-and-personal-data/

Facebook's categorization of user interests

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

The approximate 95% confidence interval is defined as

$$point\ estimate \pm 1.96 \times SE$$

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

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

Interpreting confidence intervals

Confidence intervals are ...

- always about the population
- are not probability statements
- only about population parameters, not individual observations
- only reliable if the sample statistic they're based on is an unbiased estimator of the population parameter

Facebook's categorization of user interests

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.

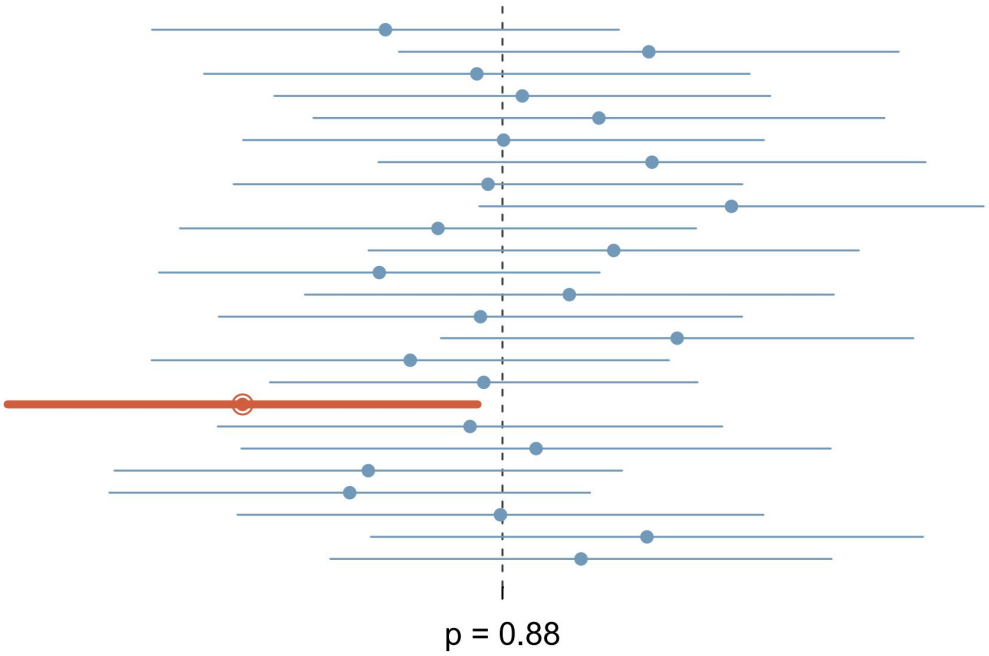
(d) there is a 64% to 70% chance that 95% of American Facebook users' interests are categorized accurately.

What does 95% confident mean?

Suppose we took many samples and built a confidence interval from each sample using the equation

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

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



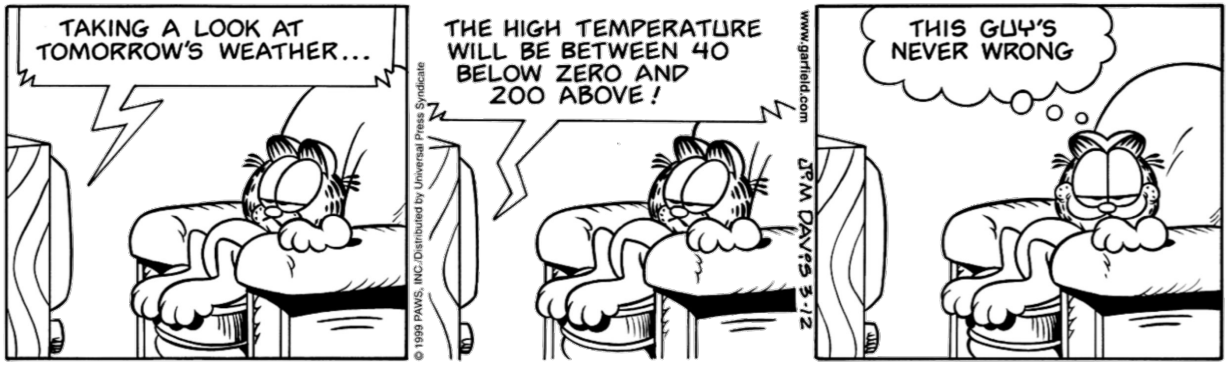
Twenty-five point estimates and confidence intervals from repeated sampling. These intervals are shown relative to the population proportion p = 0.88. Only 1 of these 25 intervals did not capture the population proportion, and this interval has been bolded.

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?



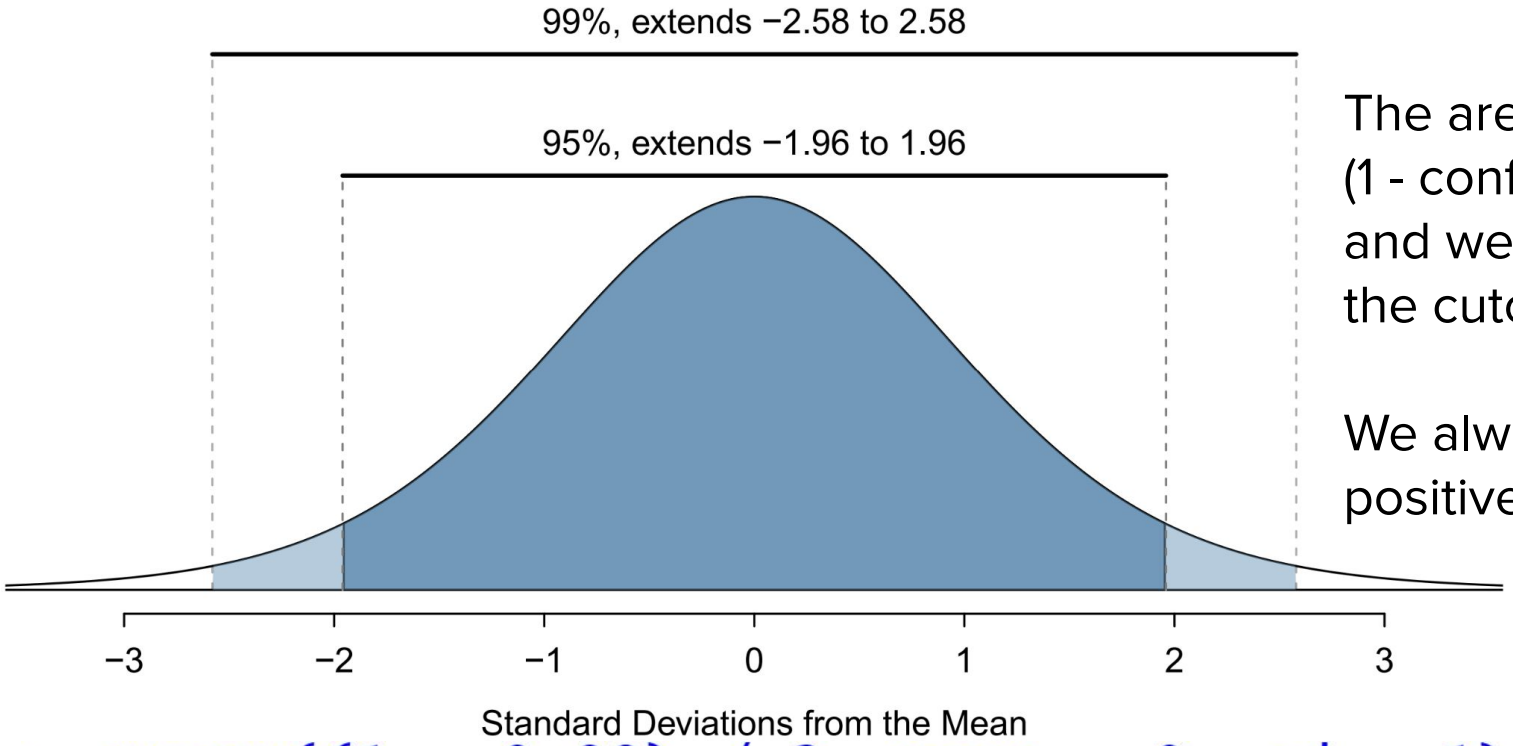
If the interval is too wide it may not be very informative.

Changing the confidence level

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

- In a confidence interval, $z^\star \times SE$ 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 $z^\star$ in the above formula.
- Commonly used confidence levels in practice are 90%, 95%, 98%, and 99%.
- For a 95% confidence interval, $z^\star = 1.96$.
- However, using the standard normal (z) distribution, it is possible to find the appropriate $z^\star$ for any confidence level.

The area between -$z^\star$ and $z^\star$ increases as $z^\star$ becomes larger. If the confidence level is 99%, we choose $z^\star$ such that 99% of a normal normal distribution is between -$z^\star$ and $z^\star$, which corresponds to 0.5% in the lower tail and 0.5% in the upper tail: $z^\star = 2.58$.



The area in a tail is (1 - confidence) / 2 and we want to find the cutoff value

We always use the positive cutoff

```
> qnorm((1 - 0.99) / 2, mean = 0, sd = 1)
[1] -2.575829
> qnorm((1 - 0.95) / 2, mean = 0, sd = 1)
[1] -1.959964
```

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

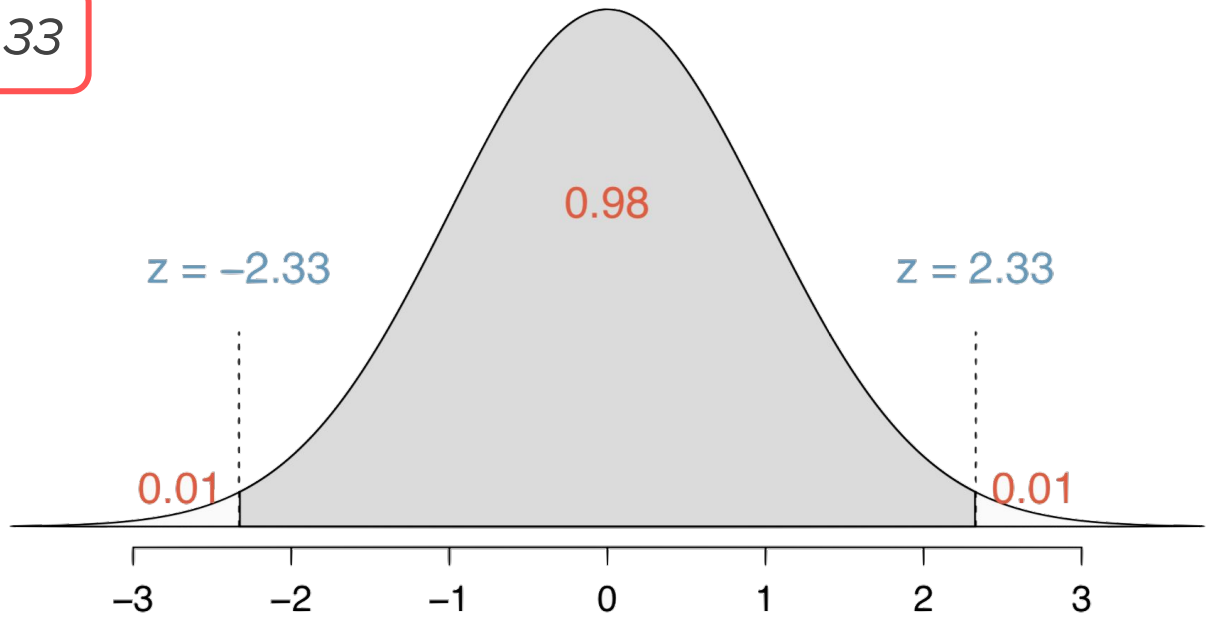
(a) Z = 2.05

(b) Z = 1.96

(c) Z = 2.33

(d) Z = −2.33

(e) Z = −1.65




```
> qnorm((1 - 0.98) / 2, mean = 0, sd = 1)
[1] -2.326348
```

Confidence interval for a single proportion

Once you've determined a one-proportion confidence interval would be helpful for an application, there are four steps to constructing the interval:

Prepare. Identify $\hat{p}$ and n, and determine what confidence level you wish to use.

Check. Verify the conditions to ensure $\hat{p}$ is nearly normal. For one-proportion confidence intervals, use $\hat{p}$ in place of p to check the success-failure conditions ($np \geq 10$ and $n(1 - p) \geq 10$).

Calculate. If the conditions hold, compute SE using $\hat{p}$, find $z^\star$, and construct the interval.

Conclude. Interpret the confidence interval in the context of the problem.