Advanced Statistics
DS2003 (BDS-4A)
Lecture 05

Instructor: Dr. Syed Mohammad Irteza

Assistant Professor, Department of Computer Science, FAST

01 March, 2022

Previous Lecture

- Point Estimates and Sampling Variability
  - Parameter Estimation
  - Margin of Error
  - Sampling Distribution
- Central Limit Theorem
  - CLT conditions

Central Limit Theorem

Sample proportions will be nearly normally distributed with mean equal to the population proportion, *p*, and standard error equal to $\sqrt{\frac{p\,(1-p)}{n}}$

$$\hat{p} \sim N\left(mean = p, SE = \sqrt{\frac{p\,(1-p)}{n}}\right)$$

It wasn't a coincidence that the sampling distribution we saw earlier was symmetric, and centered at the true *population proportion*.

We won't go through a detailed proof of why $SE = \sqrt{\frac{p\,(1-p)}{n}}$

but note that as n increases SE decreases.

- As n increases samples will yield more consistent $\hat{p}$ *values*, i.e. variability among the different $\hat{p}$ will be lower.

CLT conditions

Certain conditions must be met for the CLT to apply:

- ***Independence***: Sampled observations must be independent. This is difficult to verify, but is more likely if
  - random sampling/assignment is used, and
  - if sampling without replacement, n < 10% of the population.
- ***Sample size***: There should be at least 10 expected successes and 10 expected failures in the observed sample.
  - This is difficult to verify if you don’t know the population proportion (or can’t assume a value for it). In those cases we look for the number of observed successes and failures to be at least 10.

When *p* is unknown

The CLT states

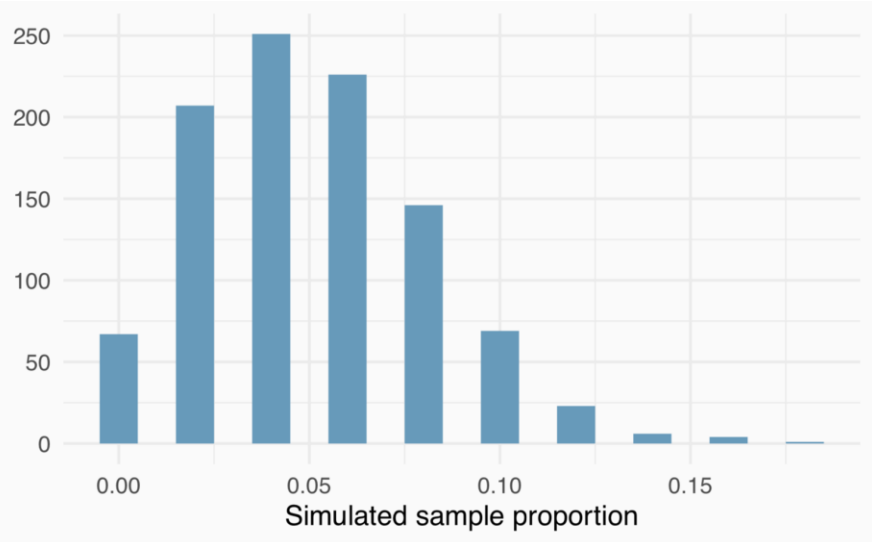
$$SE = \sqrt{\frac{p\,(1-p)}{n}}$$

- with the condition that *np* and *n(1 – p)* are *at least 10*.

- However, we often don’t know the value of *p*, the population proportion. In these cases we substitute $\hat{p}$ for *p*.

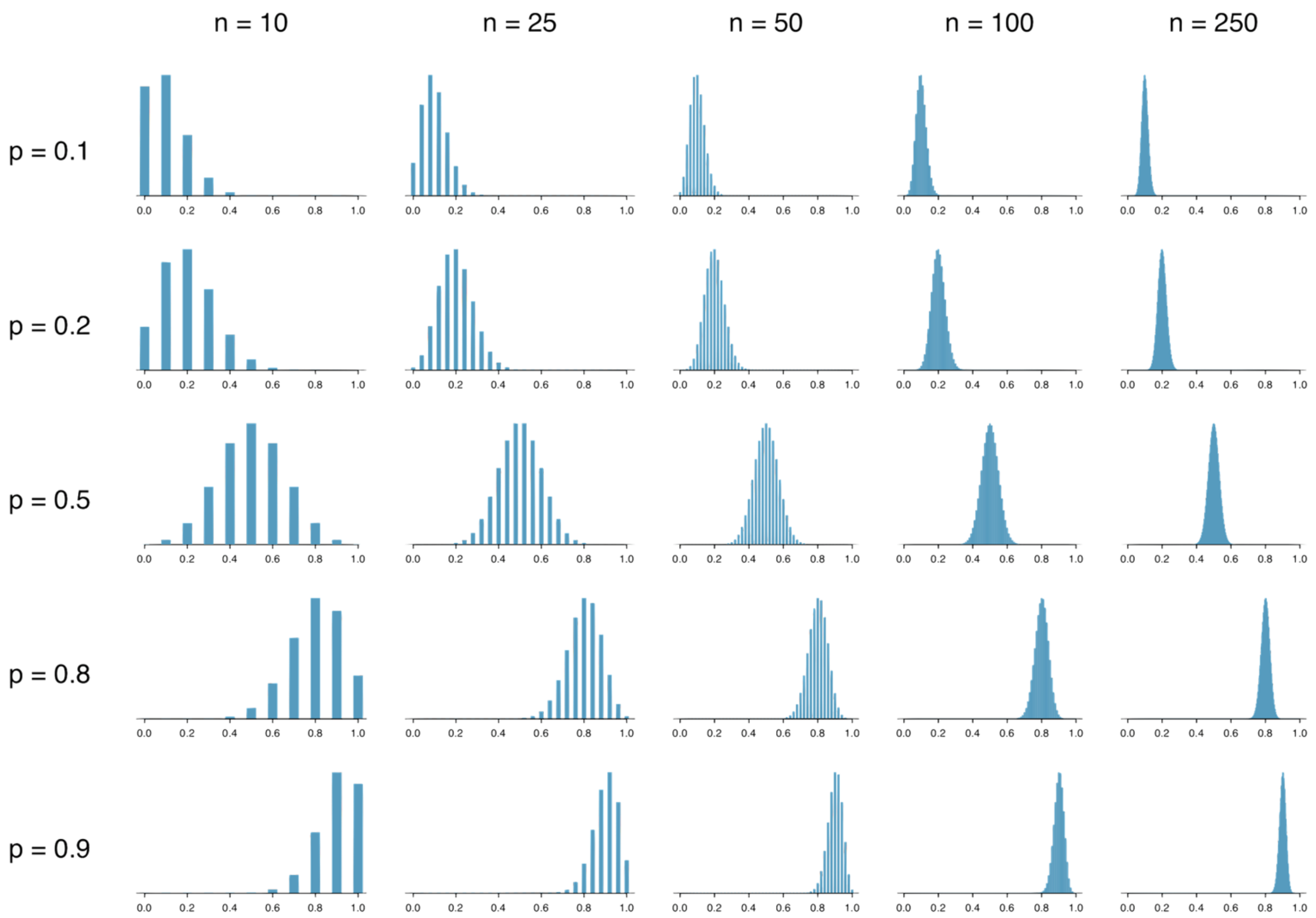
When *np* or *n(1 - p)* is small

- Suppose we have a population where the *true population proportion* is *p = 0.05*, and we take random samples of *size n = 50* from this population. We calculate the sample proportion in each sample and *plot these proportions*.
  - Would you expect this distribution to be *nearly normal*? Why, or why not?



- No, the *success-failure condition* is not met (50 x 0.05 = *2.5*), so we would *not expect* the sampling distribution to be *nearly normal*.

What happens when *np* and/or *n(1 – p) < 10*



When the conditions are not met....

- When either *np* or *n(1 − p)* is *small*, the *distribution is more discrete*.
- When *np* or *n(1 − p) < 10*, the *distribution is more skewed*.
- The *larger* both *np* and *n(1 − p)*, the *more normal* the distribution.
- When *np* and *n(1 − p)* are *both very large*, the *discreteness* of the *distribution* is *hardly evident*, and the *distribution looks* much more like a *normal distribution*.

Extending the framework for other statistics

- The strategy of using a *sample statistic* to *estimate a parameter* is quite common, and it's a strategy that we can apply to *other statistics* besides a *proportion*.
  - Take a random sample of students at a college and ask them how many extracurricular activities they are involved in to estimate the average number of extra curricular activities all students in this college are interested in.

- The principles and general ideas from this chapter apply to other parameters as well, even if the details change a little.

Confidence Intervals for a proportion

- Topic → 5.2 (OpenIntro website)

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 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.

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

Facebook’s categorization of user interests

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

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

- The approximate 95% confidence interval is defined as:

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

Facebook's categorization of user interests

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

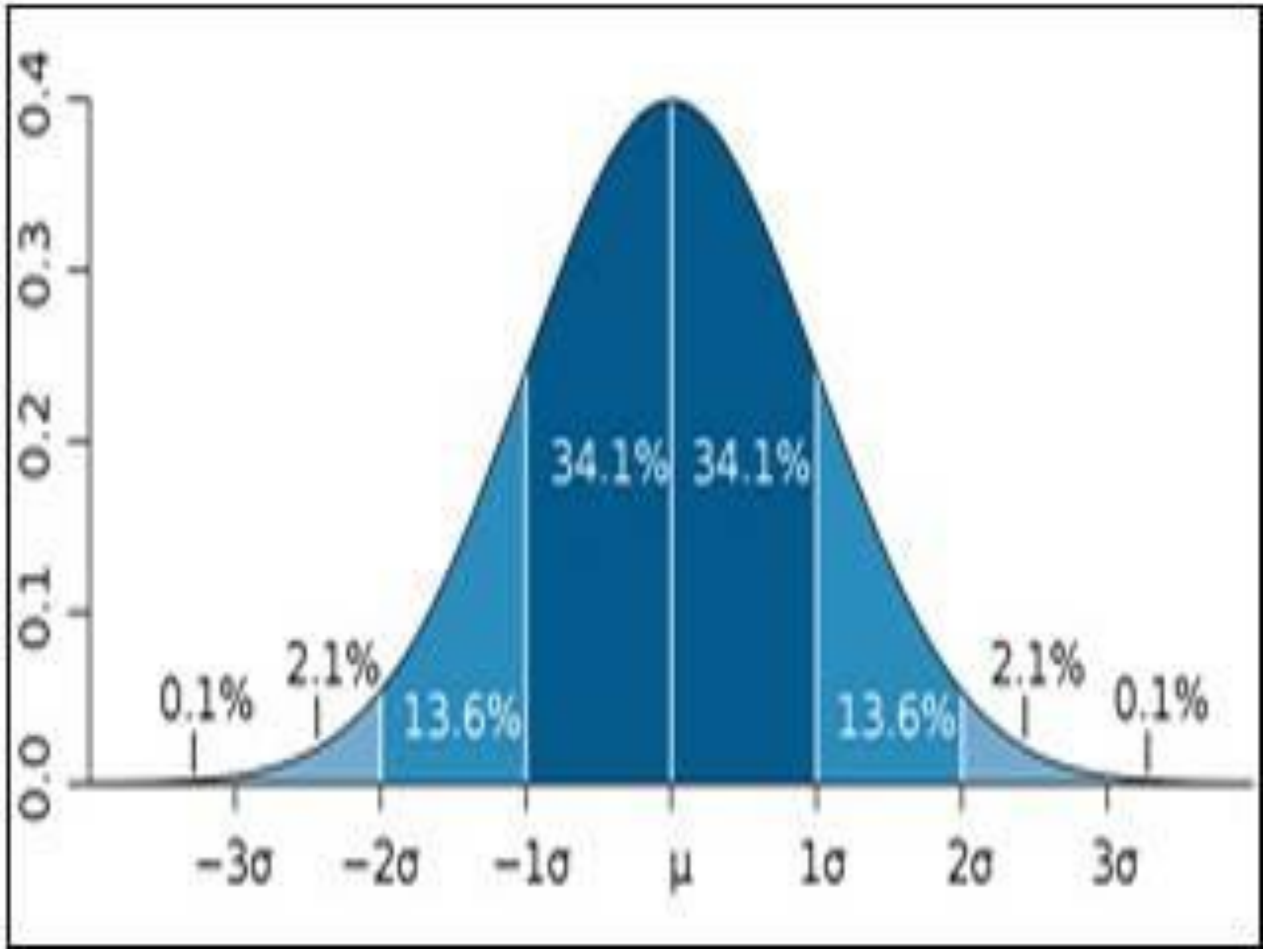
- The approximate 95% confidence interval is defined as:

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

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

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

Where did 1.96 come from?



In probability and statistics, ***1.96*** is the approximate value of the ***97.5*** percentile point of the standard normal distribution.

95% of the area under a normal curve lies within roughly ***1.96 standard deviations*** of the mean, and due to the ***central limit theorem***, this number is therefore used in the construction of ***approximate 95% confidence intervals***.

Its ubiquity is due to the arbitrary but common convention of using confidence intervals with ***95% coverage*** rather than other coverages (such as ***90% or 99%***).

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 *67%* 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.

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

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

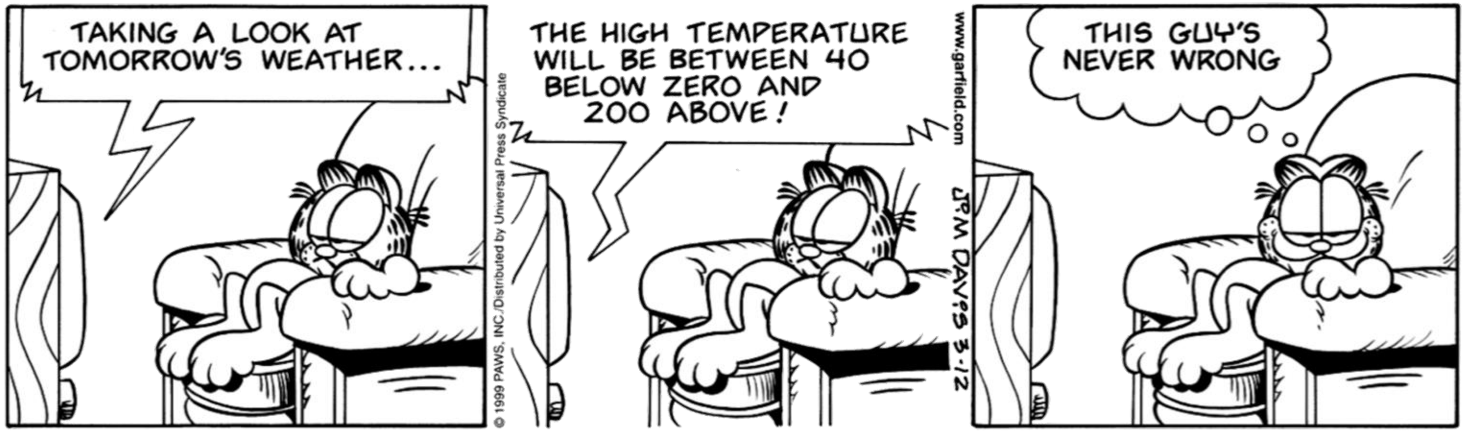
- Then *about 95%* of those *intervals would contain the true population proportion (p)*.

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*?

  A *wider interval*.

- Can you see any drawbacks to using a wider interval?



For us in Pakistan:
Fahrenheit → Celsius
-40F = -40C
200F = 93.334C

- *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.

Practice – Appropriate Z* value

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

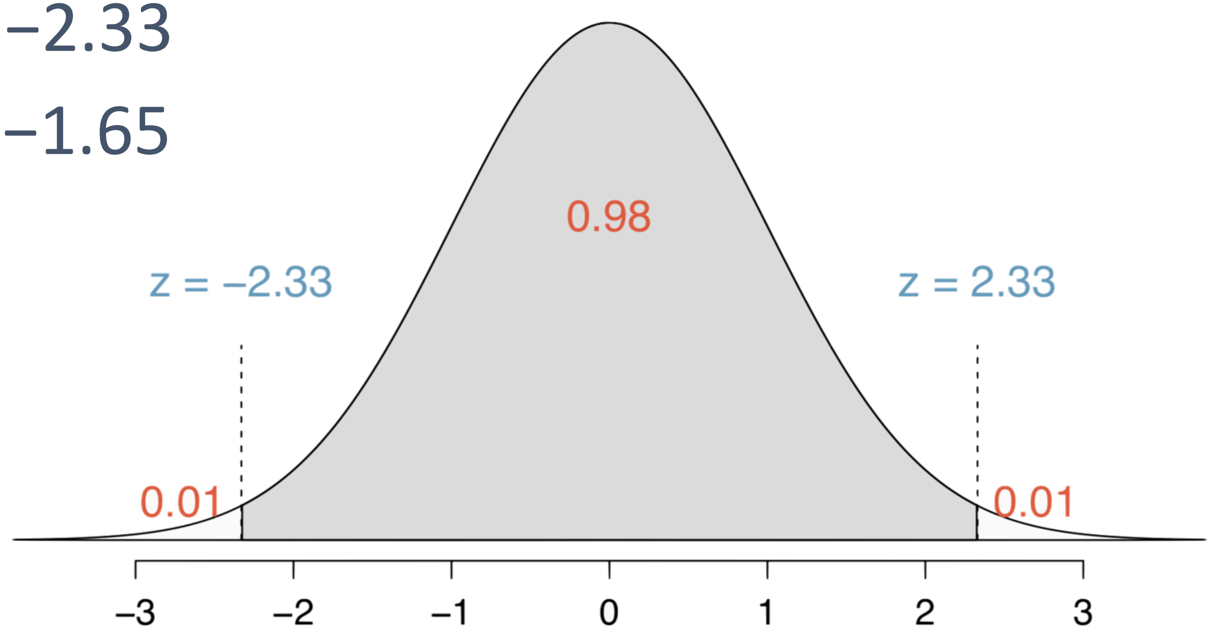
(a) Z = 2.05

(b) Z = 1.96

(c) Z = 2.33

(d) Z = −2.33

(e) Z = −1.65



Practice – Appropriate Z* value

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

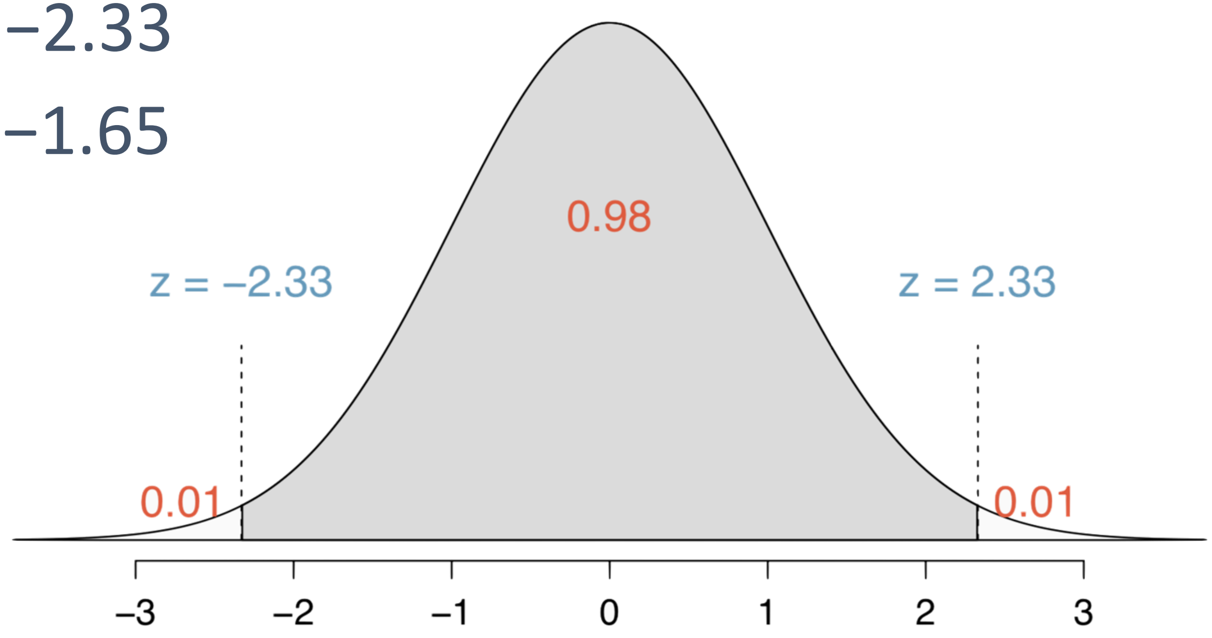
(a) Z = 2.05

(b) Z = 1.96

(c) Z = 2.33

(d) Z = −2.33

(e) Z = −1.65



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 are based on is an *unbiased estimator* of the population parameter

Average number of *close friendships*

- A random sample of 50 college students were asked how many *close friendships* they have formed in college so far. This sample yielded a mean of 3.2 and a standard deviation of 1.74. Estimate the true average number of *close friendships* using this sample.

Average number of *close friendships*

- A random sample of 50 college students were asked how many *close friendships* they have formed in college so far. This sample yielded a mean of 3.2 and a standard deviation of 1.74. Estimate the true average number of *close friendships* using this sample.

$\bar{x} = 3.2$ $s = 1.74$

The *approximate 95% confidence interval* is defined as

point estimate ± 2 x SE

$$SE = \frac{s}{\sqrt{n}} = \frac{1.74}{\sqrt{50}} \approx 0.25$$

Note, we are using 2 instead of 1.96, therefore it is now approx. 95% and not exactly 95%

Average number of *close friendships*

- A random sample of 50 college students were asked how many *close friendships* they have formed in college so far. This sample yielded a mean of 3.2 and a standard deviation of 1.74. Estimate the true average number of *close friendships* using this sample.

$\bar{x} = 3.2$ $s = 1.74$

The *approximate 95% confidence interval* is defined as

point estimate ± 2 x SE

$$SE = \frac{s}{\sqrt{n}} = \frac{1.74}{\sqrt{50}} \approx 0.25$$

$\bar{x} \pm 2 \times SE \rightarrow 3.2 \pm 2 \times 0.25$

$\rightarrow (3.2 - 0.5, 3.2 + 0.5)$

$\rightarrow (2.7, 3.7)$

Interpretation - Avg number of *close friendships*

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

We are 95% confident that:

(a) the average number of *close friendships* college students have in this sample is between 2.7 and 3.7.

(b) college students on average have between 2.7 and 3.7 *close friendships*.

(c) a randomly chosen college student has 2.7 to 3.7 *close friendships*.

(d) 95% of college students have 2.7 to 3.7 *close friendships*.

Interpretation - Avg number of *close friendships*

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

We are 95% confident that:

(a) the average number of *close friendships* college students have in this sample is between 2.7 and 3.7.

(b) ***college students on average have between 2.7 and 3.7 close friendships.***

(c) a randomly chosen college student has 2.7 to 3.7 *close friendships*.

(d) 95% of college students have 2.7 to 3.7 *close friendships*.

Difference between SD and SE

- Standard Deviation vs. Standard Error: What's the Difference? (statology.org)
- This is a good resource to understand the difference between the standard deviation and the standard error

Sources

- [openintro.org/os (Chapter 5)]