Review of confidence intervals and introduction to the bootstrap

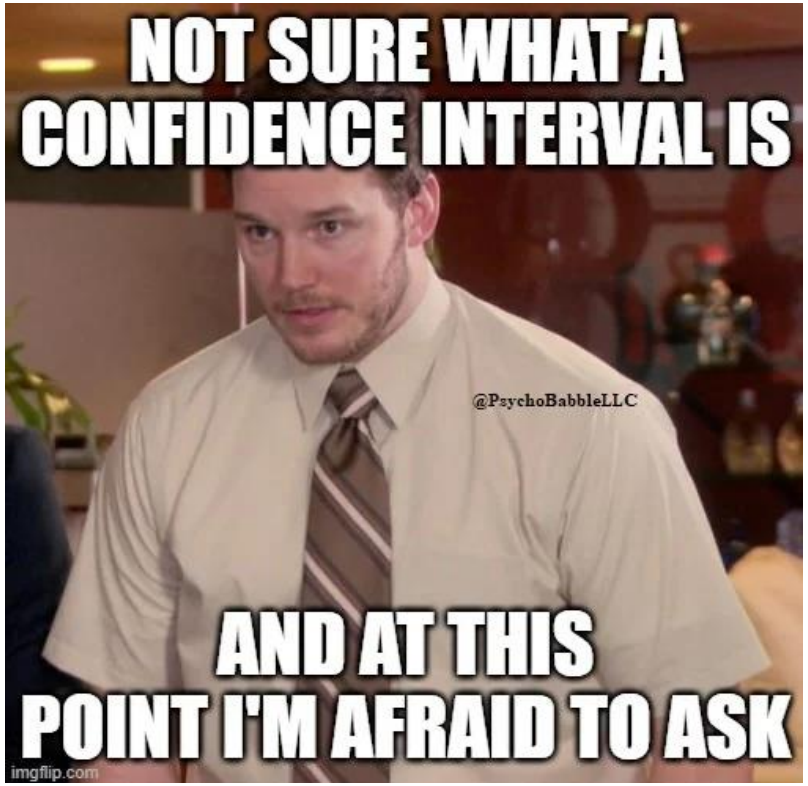
Overview

Extensive review and continuation of sampling distributions and confidence intervals

The bootstrap

If there is time:

- Using the bootstrap to create confidence intervals in R



Announcement

Homework 4 has been posted!

It is due on Gradescope on Sunday September 28th at 11pm

- **Be sure to mark each question on Gradescope!**

The material this week is going to be a bit more conceptually challenging.

Please attend the practice sessions and office hours to reinforce your understanding!

Review of sampling distributions, standard errors and confidence intervals

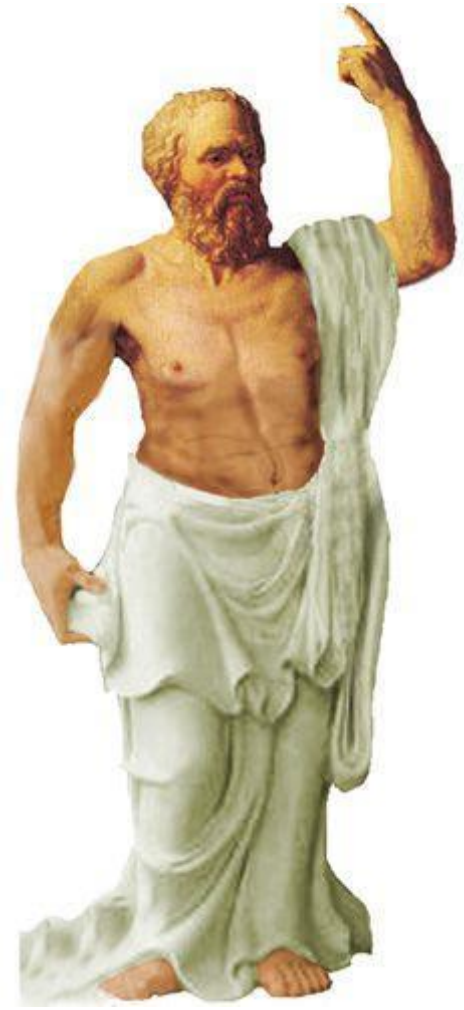
Review of confidence intervals and sampling distributions

Question$_0$: Who is this?

- Socrates!

Question$_1$: Why do I have an image of Socrates?

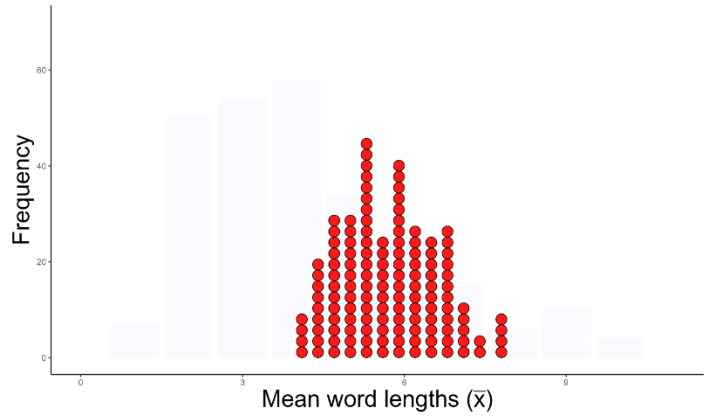
- Because I'm going to explain concepts using the Socratic method
  - i.e., I'm going to ask you a series of questions
  - We will pause so you can come up with answers own your own
  - Then we will discuss the answer as a class



Sampling distributions

Q_2: What is a sampling distribution?

- A **sampling distribution** is the distribution of sample statistics computed for different samples of the same size (n) from the same population



Q_3: What does a sampling distribution show us?

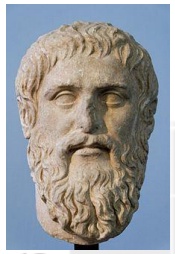
- A sampling distribution shows us how the sample statistics vary from sample to sample

Art time

Please draw:

1. Population
2. Plato
3. Population parameter for the mean with appropriate symbol
4. One sample that has 10 points
5. Sample statistic with the appropriate symbol
6. Nine more samples that have 10 points
7. Nine more sample statistics with appropriate symbol
8. A sampling distribution

Gettysburg address word length sampling distribution



10, 3, 3, 3, 4,
3, 2, 6, 10, 5

$\bar{x} = 5$

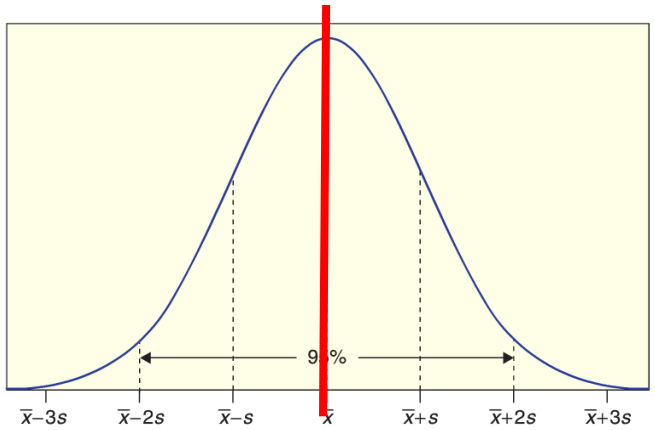
2, 6, 2, 6, 6,
2, 5, 3, 2, 9

$\bar{x} = 4.3$

3, 9, 3, 4, 4,
3, 6, 6, 2, 2

$\bar{x} = 4.2$

$E[\bar{x}] = \mu$



Sampling distribution!

Unbiased: $E[\bar{x}] = \mu$

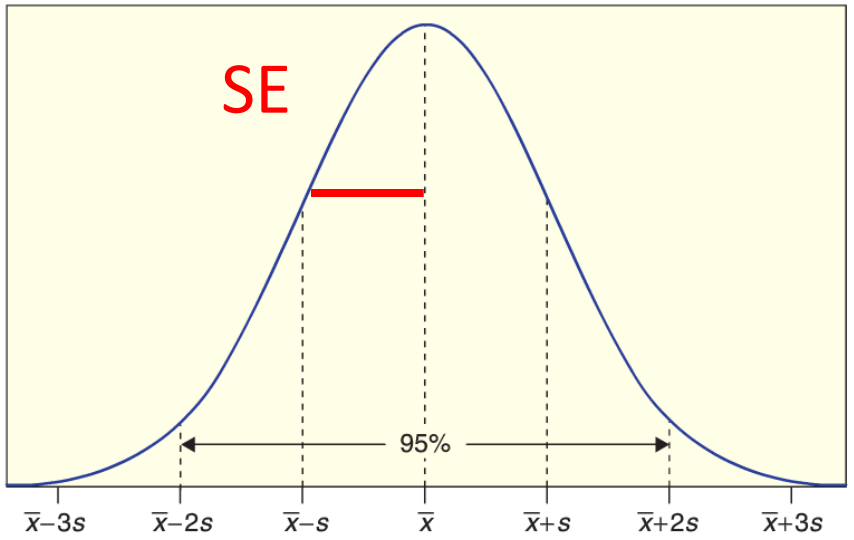
The standard error

Q_4: What is the **standard error**?

- The **standard error** of the standard deviation of the sampling distribution

Q_5: What symbol do we use to denote the standard error?

- SE



Sampling distribution in R

Q_6: Suppose we have a function called get_sample() that can generate samples from a population

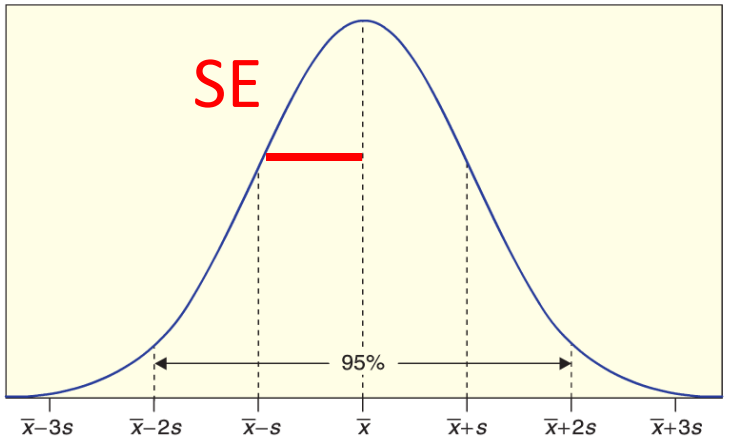
How could we estimate the SE of the mean using R?

```
sampling_dist <- do_it (10000) *  {
        curr_sample <- get_sample()
        mean(curr_sample)
}

SE_mean <- sd(sampling_dist)
```

What symbol should we use for this quantity? $\bar{x}_i$

The standard error



Q_7: What does the size of the standard error tell us?

- It tell us how much statistic values vary from sample to sample

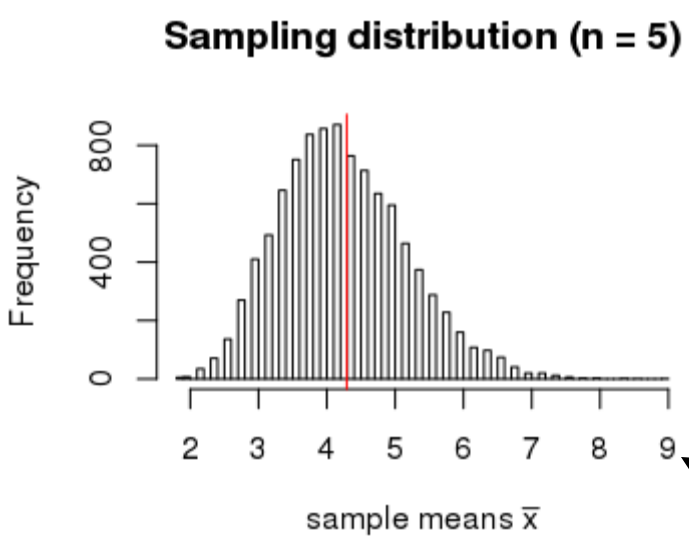
Q_8: What would it mean if there is a large SE?

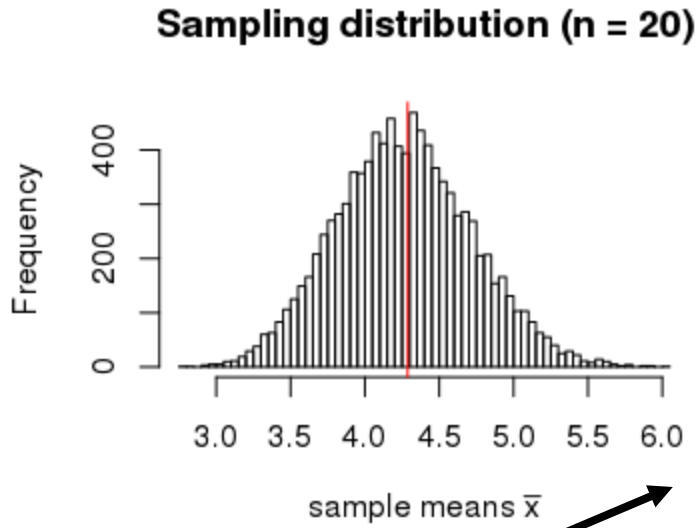
- A large SE means our statistic (point estimate) could be far from the parameter
  - E.g., $\bar{x}$ could be far from μ

Q_9: What are two ways that sampling distribution changes with a larger sample size n?

As the sample size n increases:

1. The sampling distribution becomes more like a normal distribution
2. The sampling distribution statistics become more concentrated around population parameter
   - i.e., the SE becomes smaller





x-axis range 9 vs. 6

Shapes of sampling distributions

Q_{10}: What is a commonly seen shape for sampling distributions?

Normal!

Confidence Intervals

Q_{11}: What is a **confidence interval**?

- A **confidence interval** is an interval computed by a method that will contain the *parameter* a specified percent of times

Q_{12}: What is the **confidence level**?

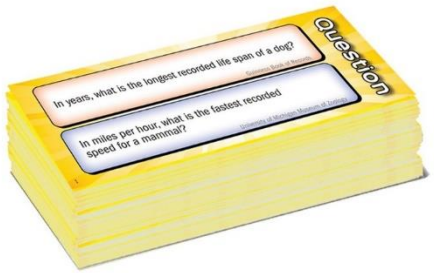
- A: The **confidence level** is the percent of all intervals that contain the parameter

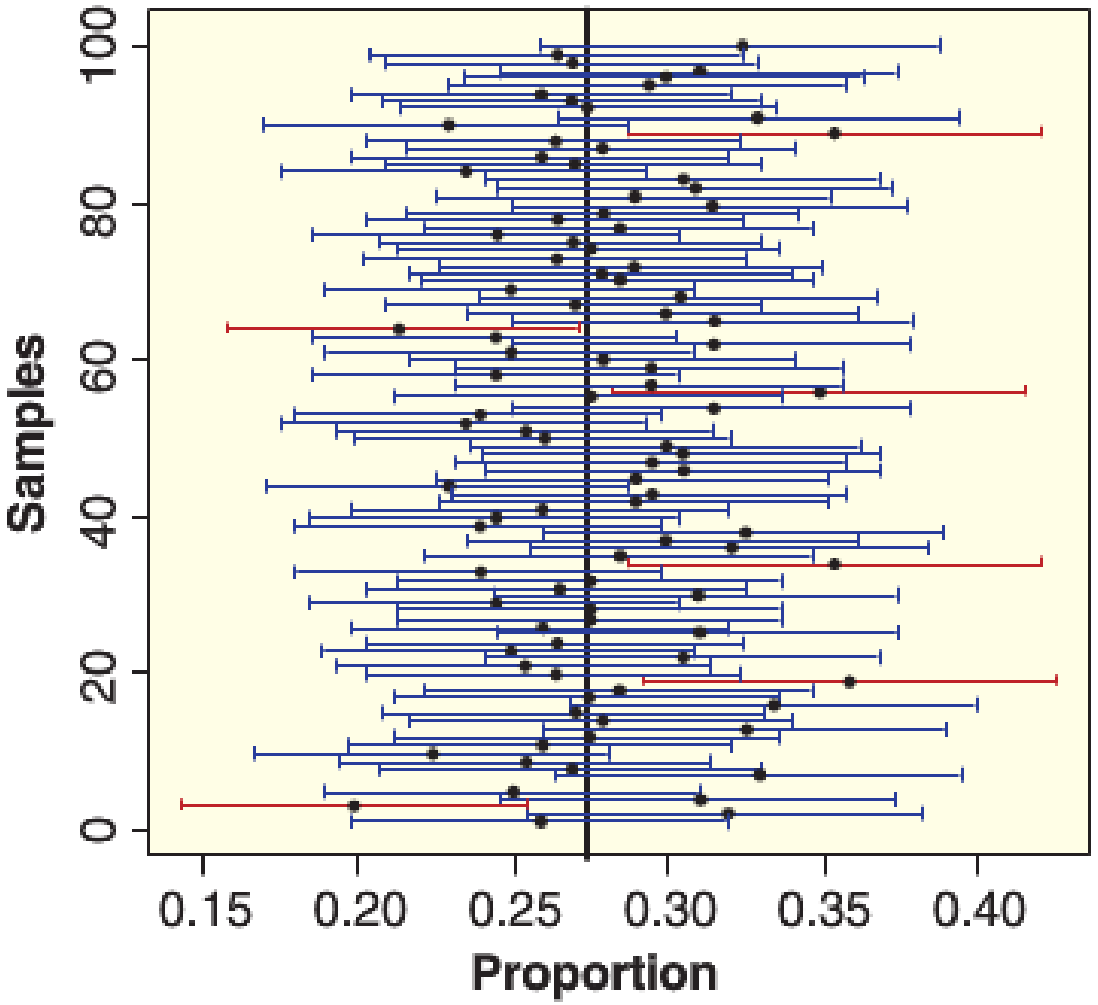
Confidence Intervals

Q_{13}: For a **confidence level** of 90%, what percent of the intervals will contain the population parameter?

90% of the **confidence intervals** will have the parameter in them!

Right???

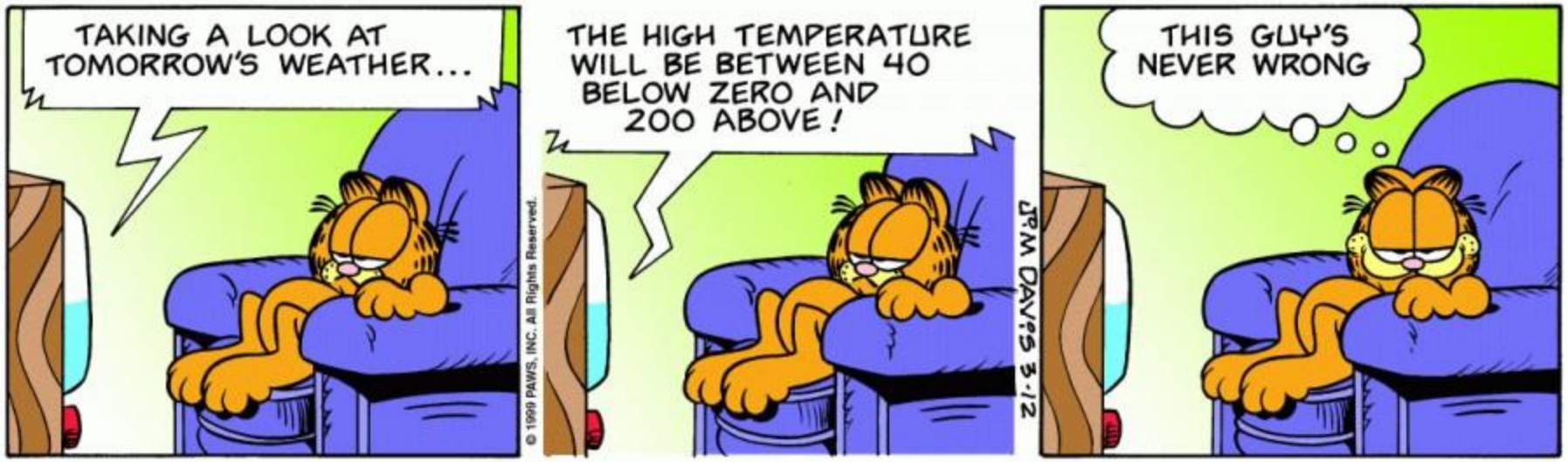




Confidence Intervals

Q_{14}: Is there a tradeoff between the **confidence level** (percent of times we capture the parameter) and the **confidence interval size?**

- Yes!



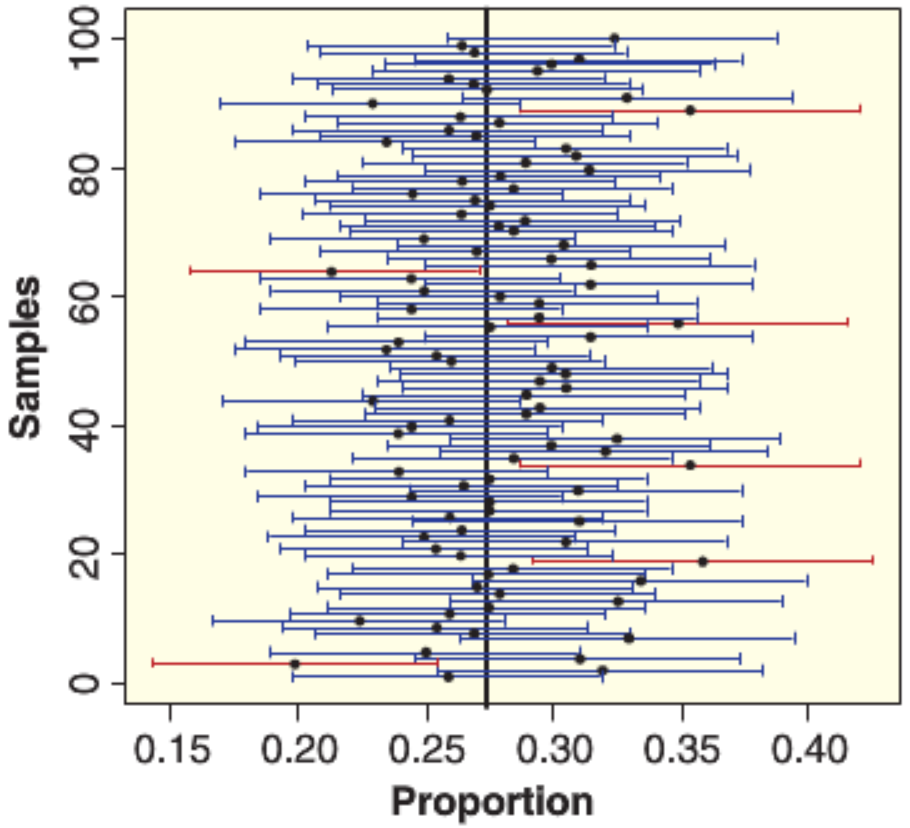
Confidence Intervals

Q_{15}: For any given confidence interval, do we know whether it has really captured the parameter?

- No ☹

But we do know that if we create 100 intervals, ~95 of these intervals will have the parameter in it

(for a 95% confidence interval)



Shapes of sampling distributions

Q_{10}: What is a commonly seen shape for sampling distributions?

- Normal!

Normal distributions

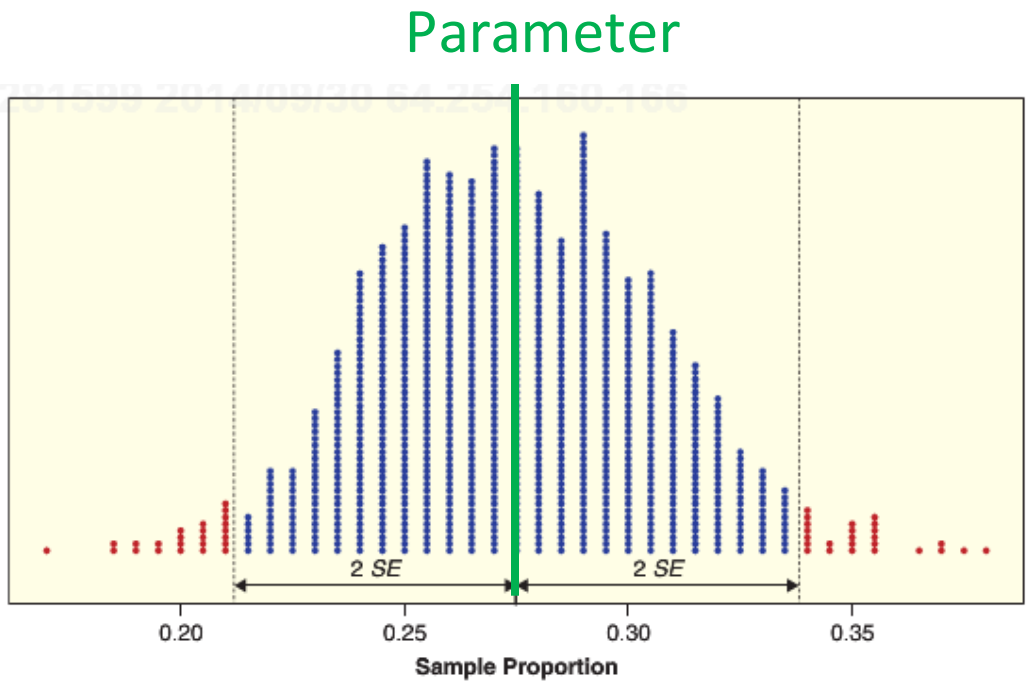
Q_{16}: For a normal distribution, what percentage of points lie within 2 standard deviations for the population mean?

- 95%

Sampling distributions

Q_{17}: For a sampling distribution that is a normal distribution, what percentage of ***statistics*** lie within 2 standard deviations (SE) of the parameter value?

- 95%



Q_{18}: If we had a *statistic value* and the value of the *SE*, could we compute a 95% confidence interval?

A: Yes, assuming the sampling distribution is normal (which it usually is)

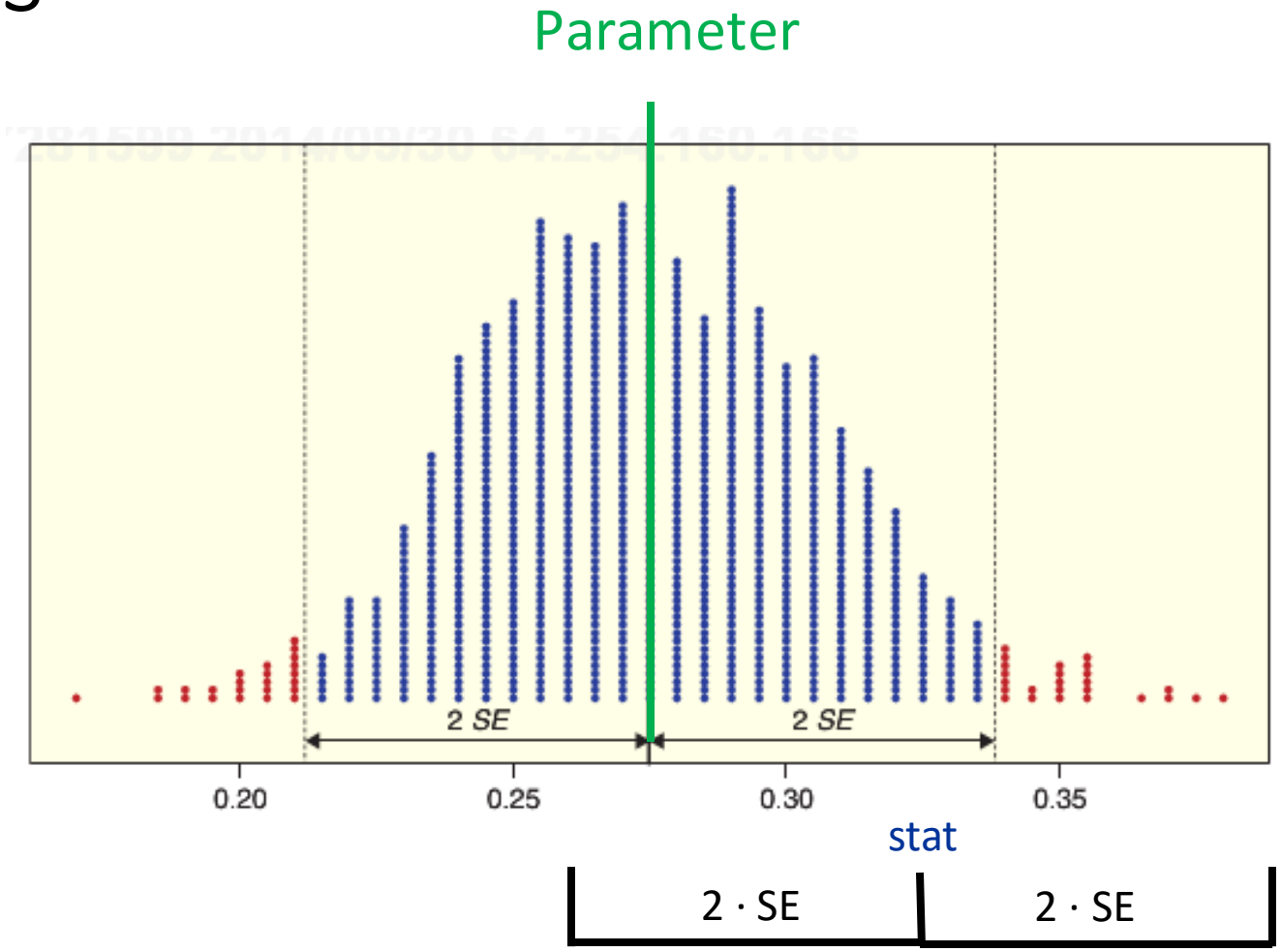
Confidence intervals

Q_{19}: Suppose we are going to randomly chosen statistic value

And we are going create an interval centered at the statistic value

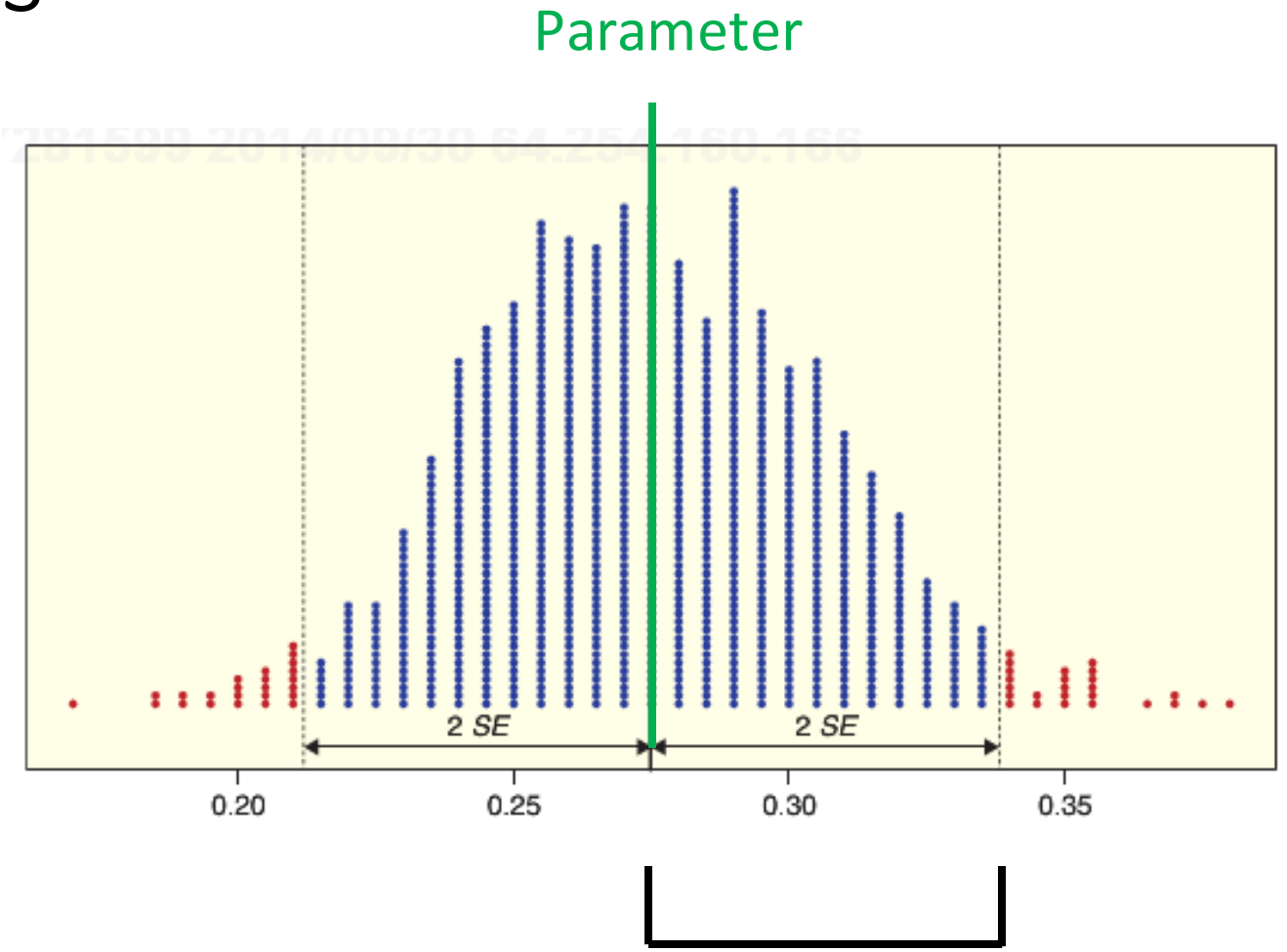
How large would the interval need to be to overlap with the parameter 95% of the time?

- A: stat ± 2 · SE



Confidence interval

Confidence intervals



Q_{20}: What is a formula can we use to calculate 95% confidence intervals?

95% confidence interval: stat $\pm$ 2 $\cdot$ SE

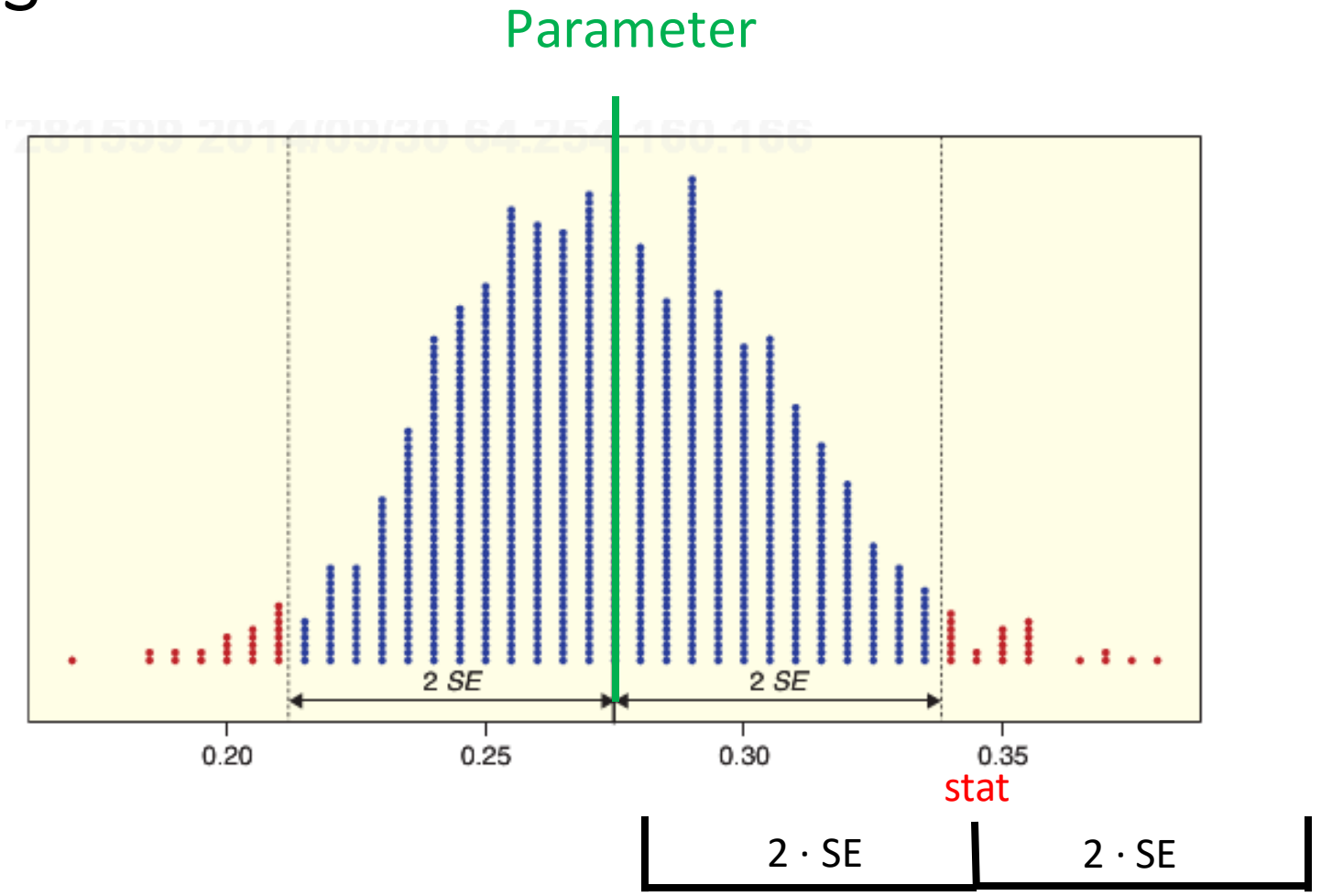
Q_{21}:What is this quantity called?

A: Margin of error

Confidence intervals

Q_{22}: How frequently do 95% confidence intervals **fail** to capture the parameter of interest?

- 5% of the time

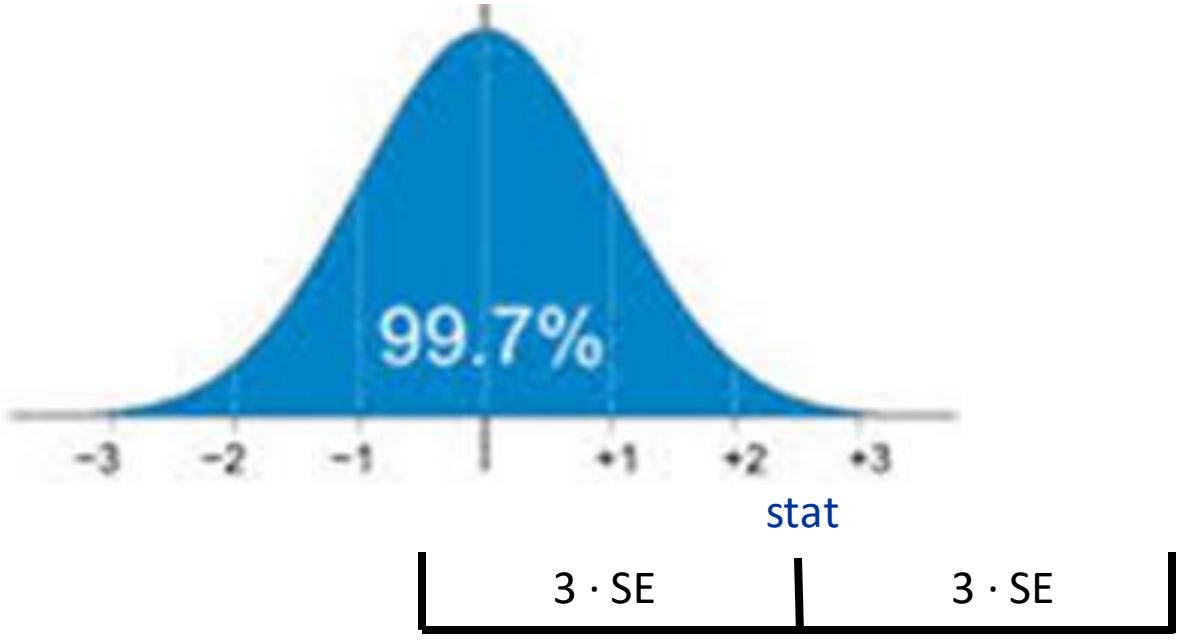


95% confidence interval: stat ± 2 · SE

Confidence intervals for other confidence levels

Q_{23}: How could we get a 99.7% confidence interval confidence level?

For normally distributed statistics, 99.7% of our statistics lie within 3 standard deviations of the mean

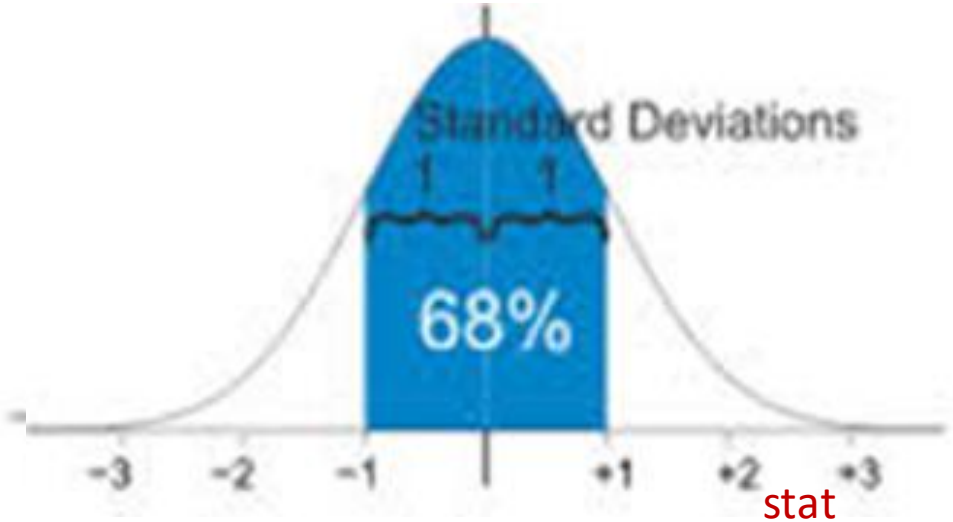


$$CI_{99.7} = \text{stat} \pm 3 \cdot SE$$

Confidence intervals for other confidence levels

Q_{24}: How could we get a 68% confidence interval confidence level?

A: For normally distributed statistics, 68% of our statistics lie within 1 standard deviations of the mean



Confidence interval

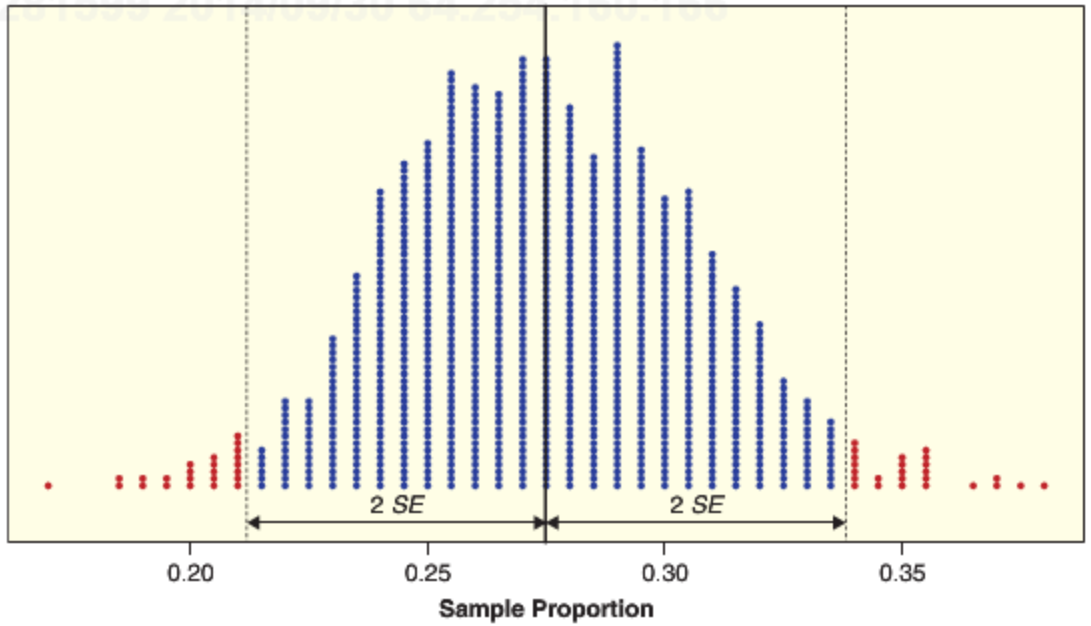
$$CI_{99.7} = \text{stat} \pm 3 \cdot SE$$

$$CI_{68} = \text{stat} \pm 1 \cdot SE$$

Confidence intervals for other confidence levels

Q_{25}: How could we get a confidence interval for the q^{th} confidence level?

A: We need to find the critical value q* such that q% of our statistics are within ± q* · SE for a normal distribution



$CI_{\cdot} = stat \pm q^* \cdot SE$

In R: > qnorm(0.975)

[1] 1.96

Sampling distributions

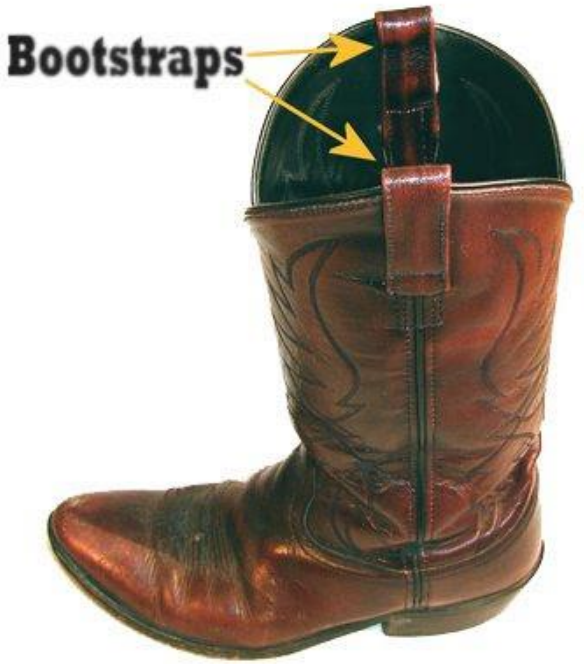
Q_{26}: Could we calculate the SE by repeatedly sampling from a population to create sampling distribution, and then take the sd of this sampling distribution?

- Not in the real world because it would require running our experiment over and over again…

Sampling distributions

Q_{27}: If we can't calculate the sampling distribution, what else can we do?

- We can pick ourselves up from the bootstraps



1. Estimate SE with $\hat{SE}$ ***from a single sample of data***
2. Then use $\bar{x} \pm 2 \cdot \hat{SE}$ to get the 95% CI

Confident intervals

Q_{28}: Do you feel confident about confidence intervals?

- I hope so!

Q_{29}: If you don't feel confident about confidence intervals, how can you become more confident?

- Go to the practice sessions
- Review material
- Ask questions on Ed Discussion
- Come to office hours!