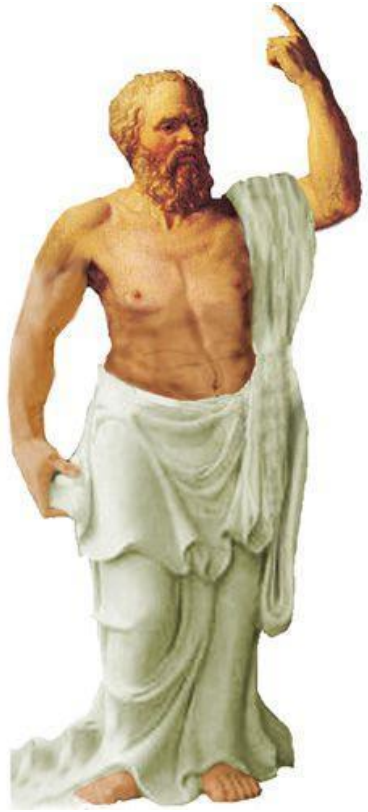
Questions about sampling distributions, standard errors and confidence intervals



Overview

Questions about homework 3?

Review: sampling distributions

Confidence intervals continued

The bootstrap

Any questions about homework 3?

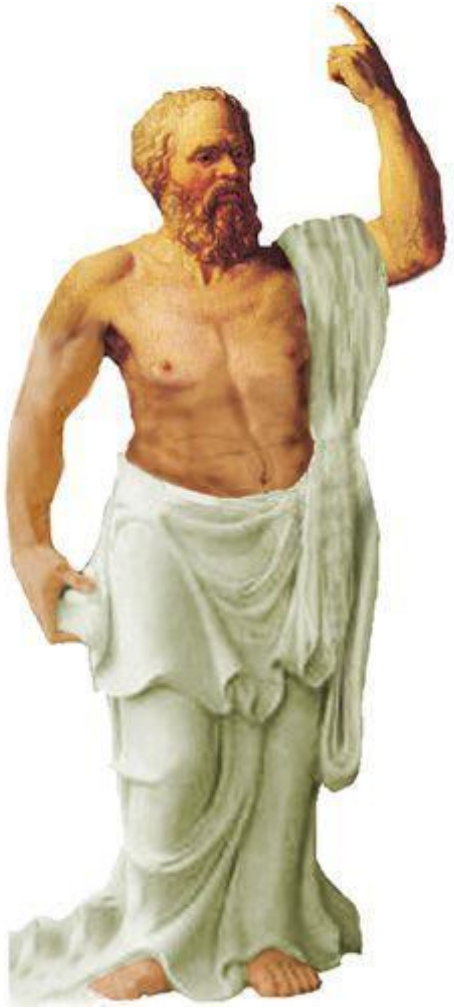
Homework 4 has been posted

- Use the link on Canvas to access homework 4 on R Studio Cloud
- Due on Gradescope at 11:30pm on Sunday February 16th

Review of confidence intervals and sampling distributions

Question$_0$: Who is this?

- Socrates!



Sampling distributions

Q_2: What is a sampling distribution?

- A: 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?

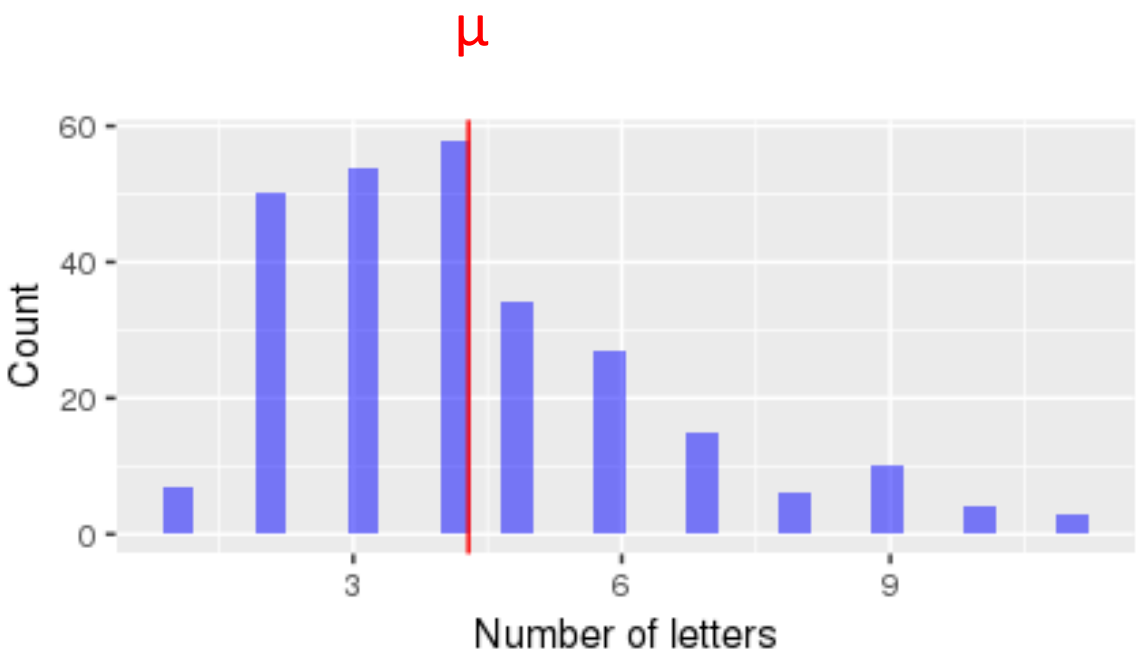
- A: A sampling distribution shows us how the sample statistic varies from sample to sample

Art time

Draw:

- Population
- 1 sample that has 10 points
- Sample statistic with appropriate symbol
- 9 more samples that have 10 points
- 9 more sample statistics with appropriate symbol
- A sampling distribution
- Plato
- Population parameter with appropriate symbol

Gettysburg address word length sampling distribution



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

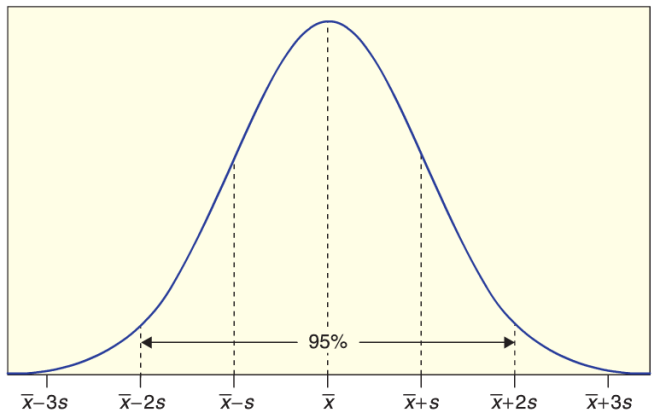
$\bar{x} = 5$

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

$\bar{x} = 4.3$

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

$\bar{x} = 4.2$



Sampling distribution!

[Gettysburg sampling distribution app]

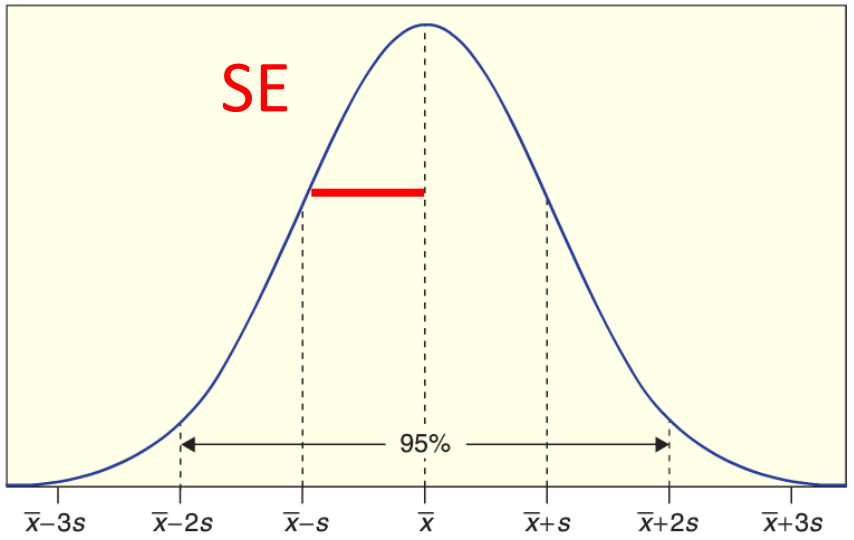
The standard error

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

- The **standard error** of a statistic is the standard deviation of the sampling distribution

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

- SE



Sampling distribution in R

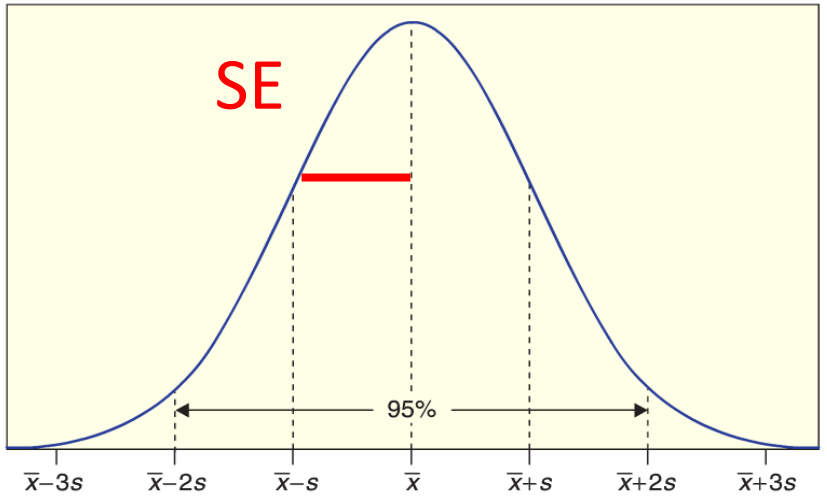
Q_6: If we had a function called "get_sample()" that could generate samples from a population, how could we estimate the SE of the mean using R?

```
sampling_dist <- do_it (100000) *  {

        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?

- A: It tell us how much statistics vary from each other

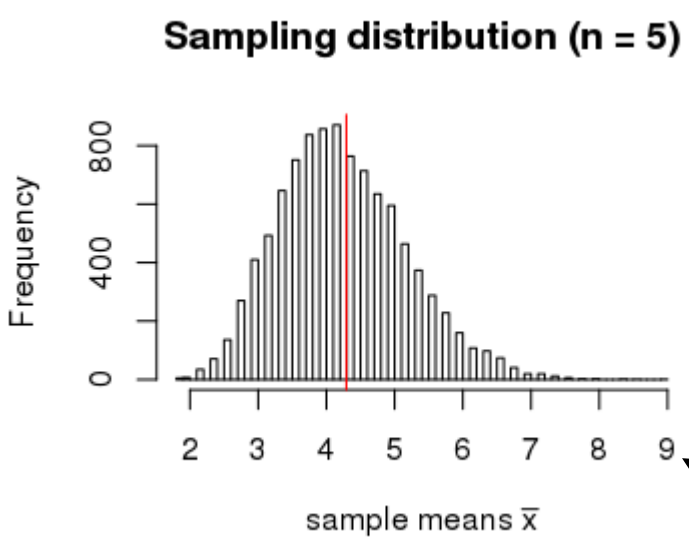
Q_8: What would it mean if there is a large SE?

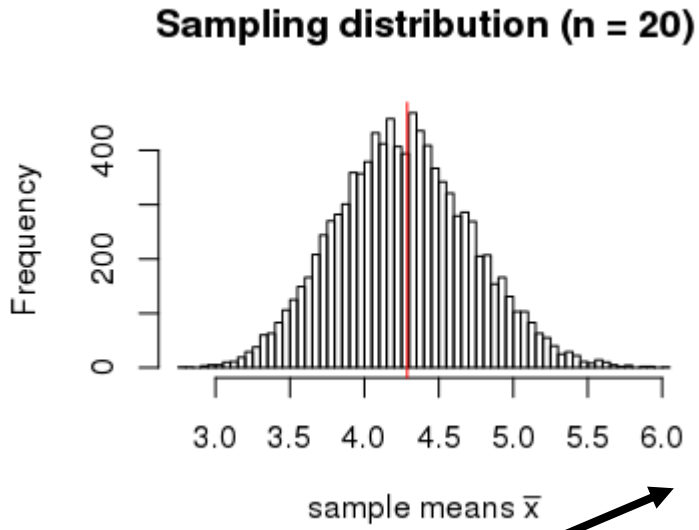
- A large SE means our statistic (point estimate) could be far from the parameter
- E.g., $\bar{x}$ could be far from μ

Q_9: How does the sampling distribution change with larger sample size n?

A: 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





x-axis range 9 vs. 6

Shapes of sampling distributions

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

A: Normal!

Confidence Intervals

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

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

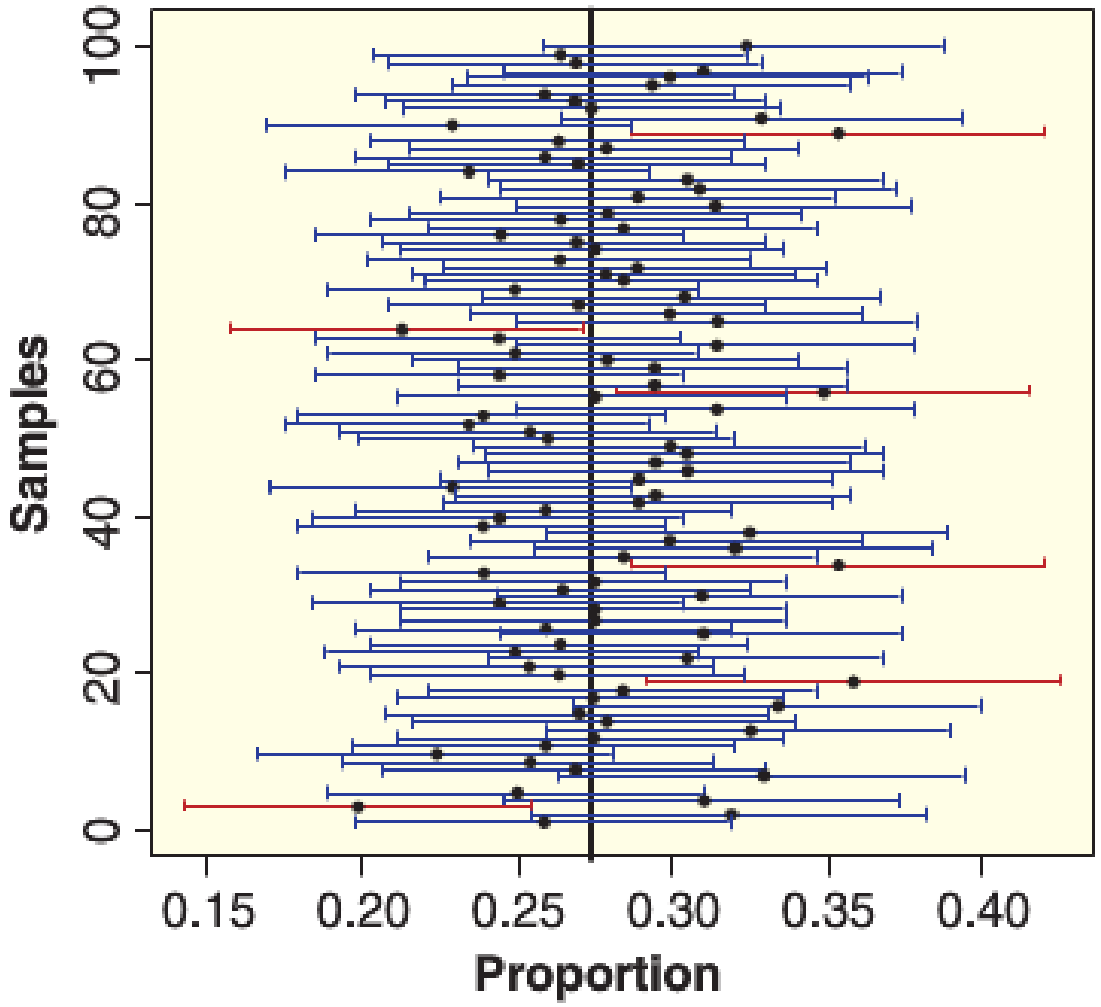
Q_{12}: What is the **confidence level**?

- A_2: The **confidence level** is the percent of all intervals that contain the parameter

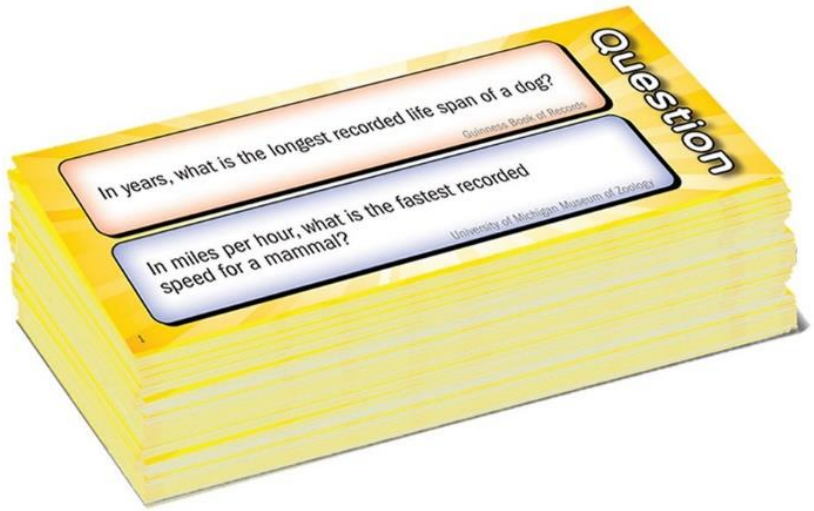
Confidence Intervals

Q_{13}: For a **confidence level** of 95%, what percent of the intervals will contain the population parameter?

A: 95% of the **confidence intervals** will have the parameter in them!



Let’s try to be 90% confidence intervals estimators



Wits and Wagers…

Wits and Wagers…

Question 1: In feet and inches, how tall was the tallest human in recorded history?

Question 2: How many floors does the leaning tower of Pisa have?

Question 3: What year was the parking meter invented?

Wits and Wagers…

Question 4: How many time zones does Russia have?

Question 5: In miles, how far does the average American drive each year?

Question 6: What percent of the world's population lives in the U.S.?

Question 7: On average, what percent of a watermelon's weight comes from water?

Wits and Wagers…

Question 8: What percentage of Americans say that reading is their favorite leisure-time activity?

Question 9: What percent of the world's surface is water?

Question 10: In what year was an ATM machine first installed in the U.S.?

Answers…

Wits and Wagers…

Question 1: In feet and inches, how tall was the tallest human in recorded history?

- 8’ 11”

Question 2: How many floors does the leaning tower of Pisa have?

- 8

Question 3: What year was the parking meter invented?

- 1935

Wits and Wagers…

Question 4: How many time zones does Russia have?

- 11

Question 5: In miles, how far does the average American drive each year?

- 13,476

Question 6: What percent of the world's population lives in the U.S.?

- 4.27%

Question 7: On average, what percent of a watermelon's weight comes from water?

- 92%

Wits and Wagers…

Question 8: What percentage of Americans say that reading is their favorite leisure-time activity?

- 35%

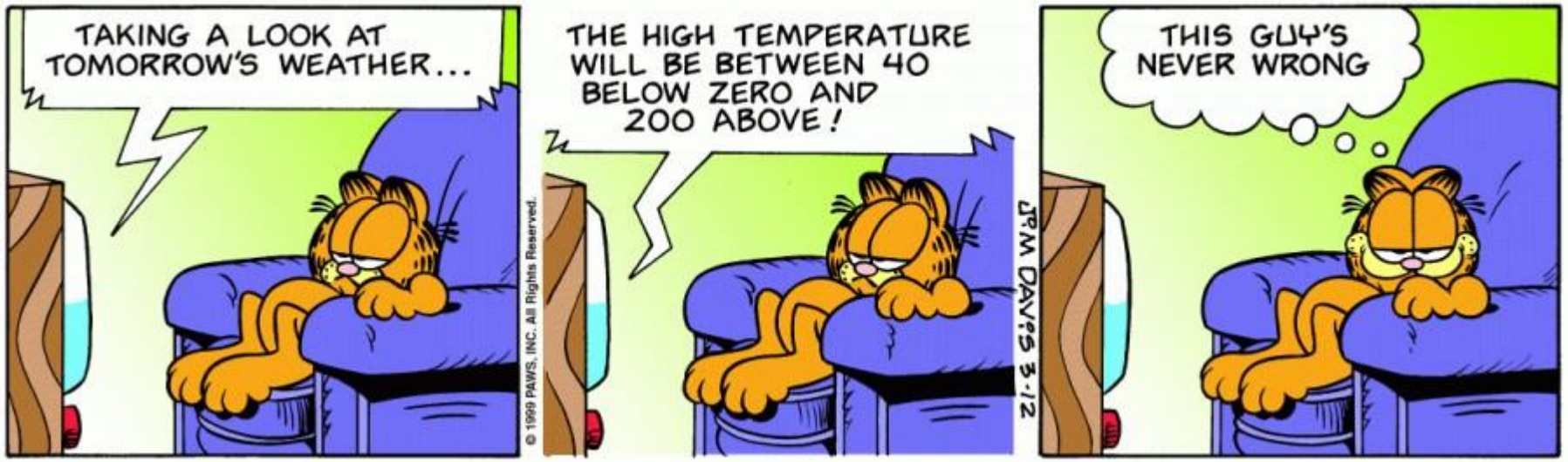
Question 9: What percent of the world's surface is water?

- 71%

Question 10: In what year was an ATM machine first installed in the U.S.?

- 1969

100% confidence intervals



There is a tradeoff between the **confidence level** (percent of times we capture the parameter) and the **confidence interval size**

Note

For any given confidence interval we compute, we don’t know whether it has really captured the parameter

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

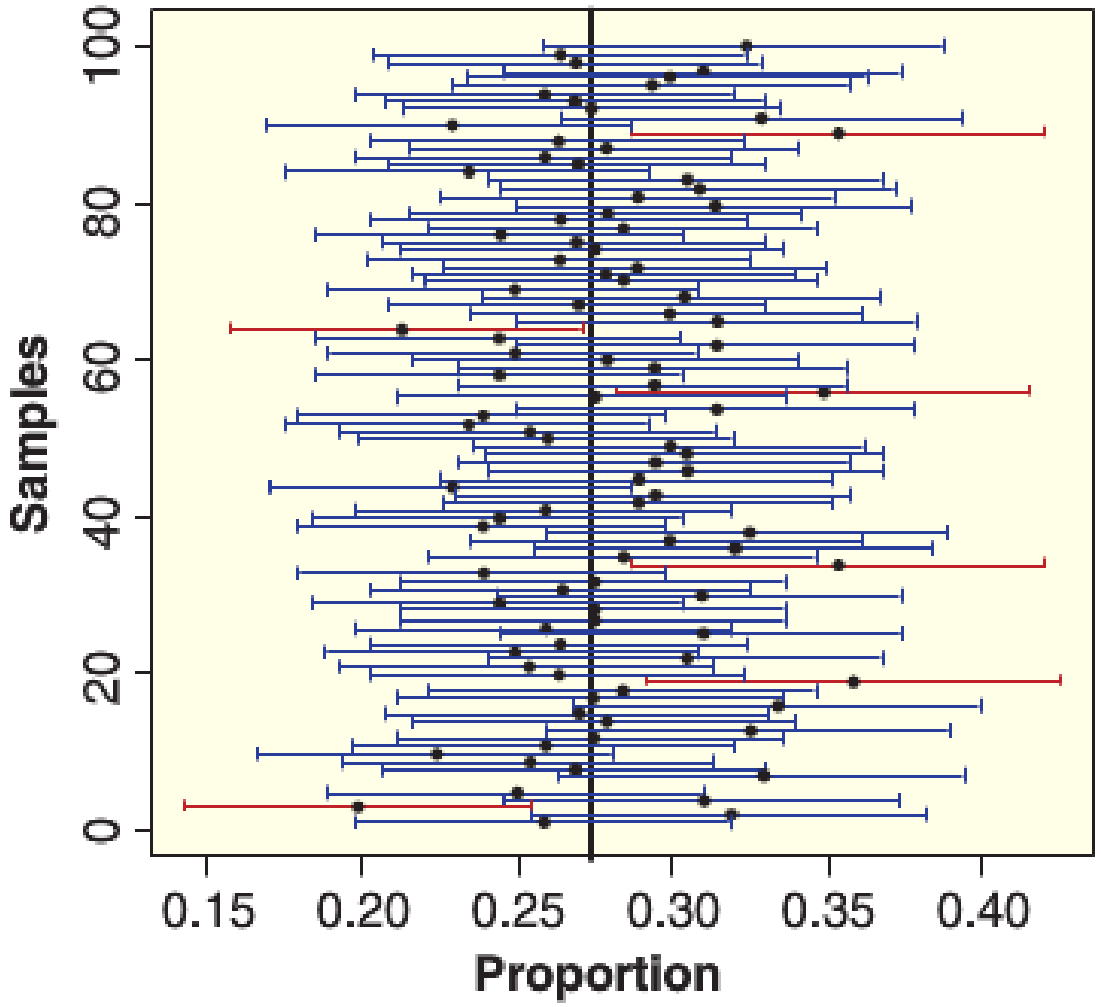
(for a 95% confidence interval)

Confidence Intervals

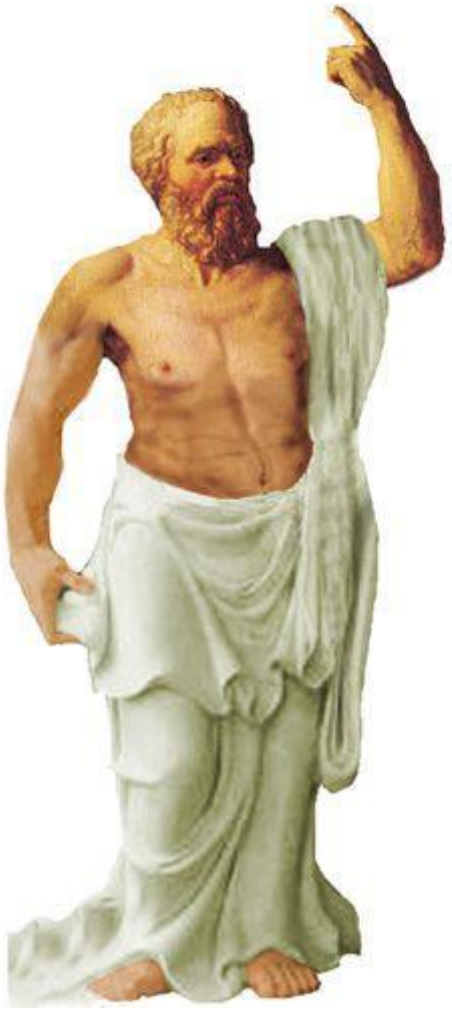
For a **confidence level** of 90%...

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

Right???



Back to Socrates…



Shapes of sampling distributions

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

A: Normal!

Normal distributions

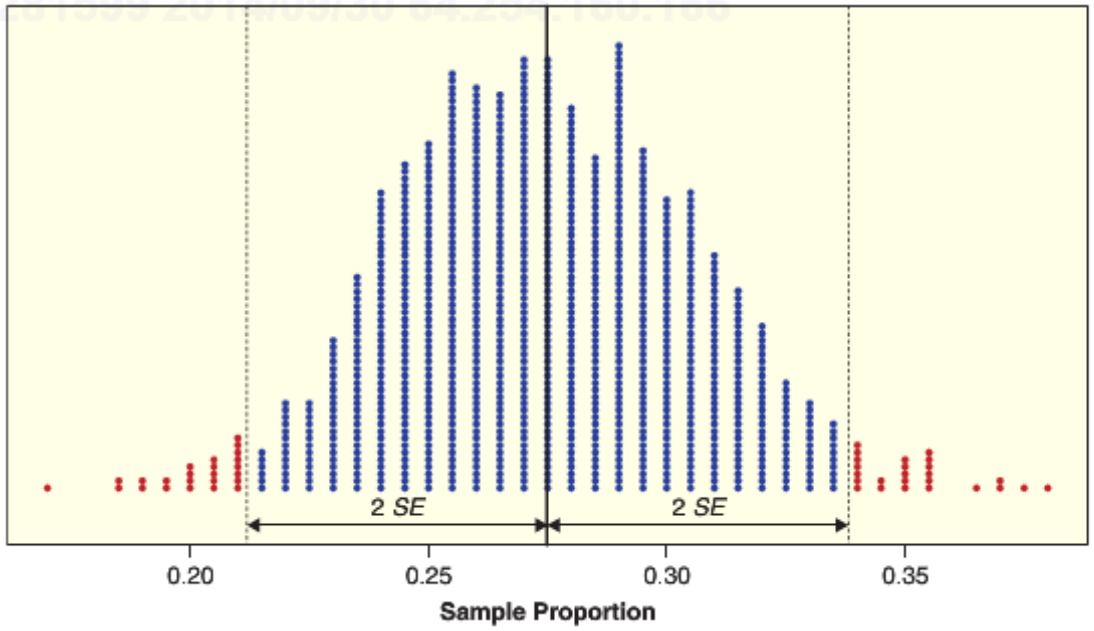
Q_{14}: For a normal distribution, what percentage of points lie within 2 standard deviations for the population mean?

A: 95%

Sampling distributions

Q_{15}: For a sampling distribution that is a normal distribution, what percentage of ***statistics*** lie within 2 standard deviations (SE) for the population mean?

A: 95%



Q_{16}: 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 often is)

Sampling distributions

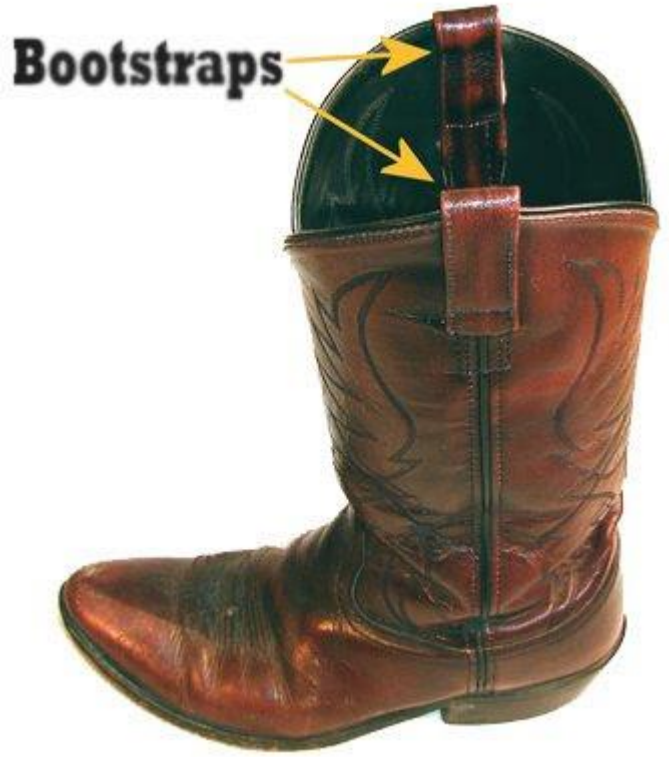
Q_{17}: Could we repeat the sampling process many times to create a sampling distribution and then calculate the SE?

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

Sampling distributions

Q_{18}: If we can't calculate the sampling distribution, what's else could we do?

- A: We could pick ourselves up from the bootstraps



1. Estimate SE with $\hat{SE}$
2. Then use $\bar{x} \pm 2 \cdot \hat{SE}$ to get the 95% CI

Plug-in principle

Suppose we get a sample from a population of size *n*

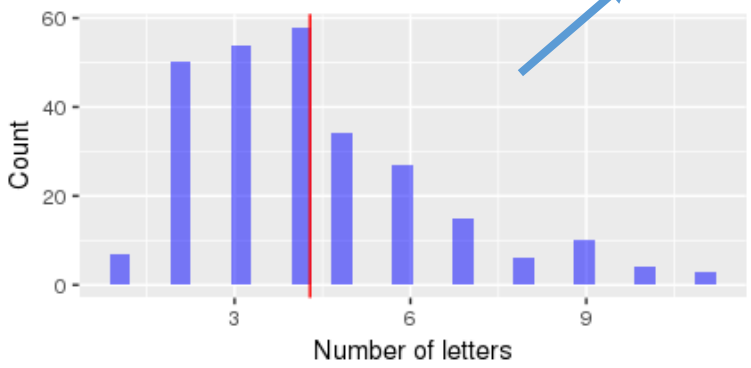
We pretend that this sample is the population (plug-in principle)

1. We then sample *n* points with replacement from *our sample*, and compute our statistic of interest

2. We repeat this process 1000's of times and get a *bootstrap* sample distribution

3. The standard deviation of this bootstrap distribution (SE* bootstrap) is a good approximate for standard error SE from the real sampling distribution

Gettysburg address word length bootstrap distribution



The sample (n = 10)

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

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

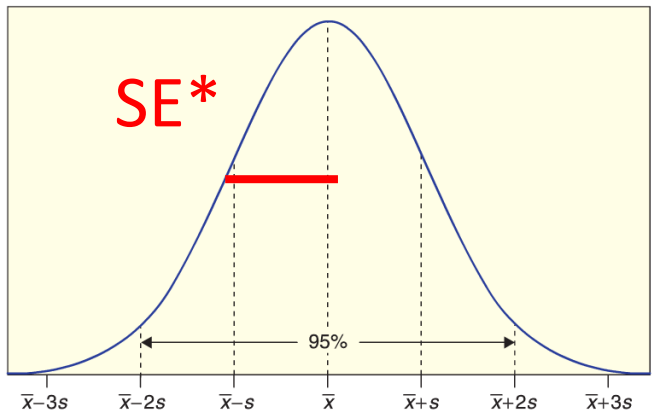
$\bar{x}^* = 4$

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

$\bar{x}^* = 4.1$

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

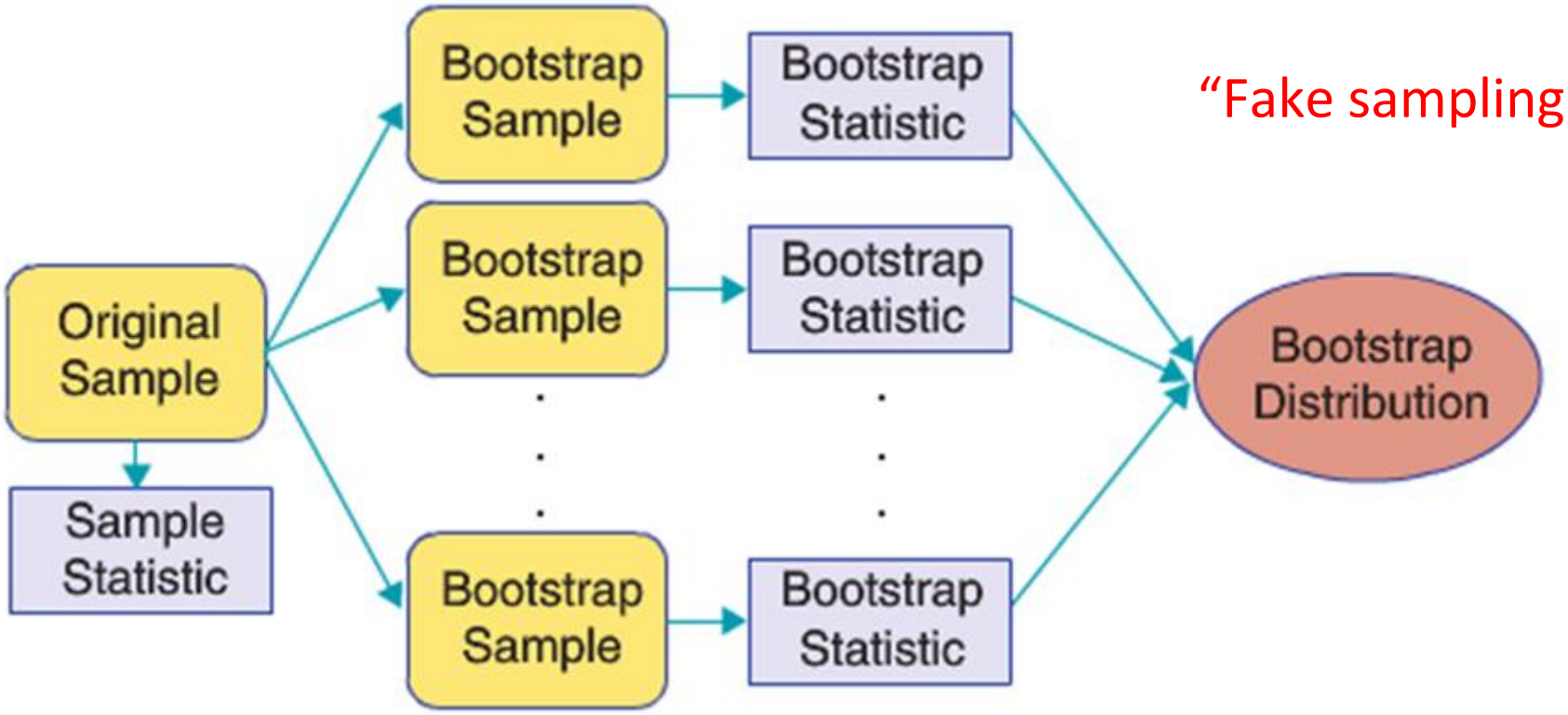
$\bar{x}^* = 3.9$



Bootstrap distribution!

Notice there is no 9's in the bootstrap samples

Bootstrap process



"Fake sampling distribution"

95% Confidence Intervals

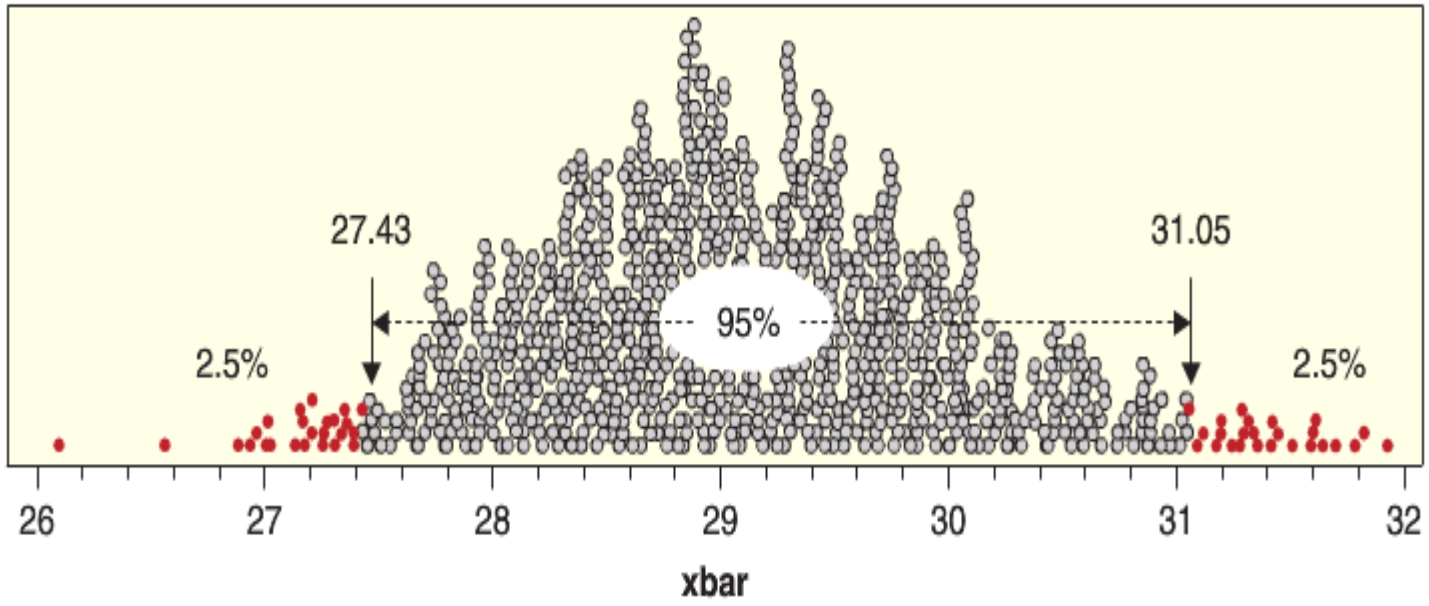
When a bootstrap distribution for a sample statistic is approximately normal, we can estimate a 95% confidence interval using:

$$Statistic \pm 2 \cdot SE^*$$

Where SE* is the standard error estimated using the bootstrap

What if the bootstrap distribution is not normal?

If the bootstrap distribution is approximately symmetric, we can use percentiles in the bootstrap distribution to an interval that matches the desired confidence level.



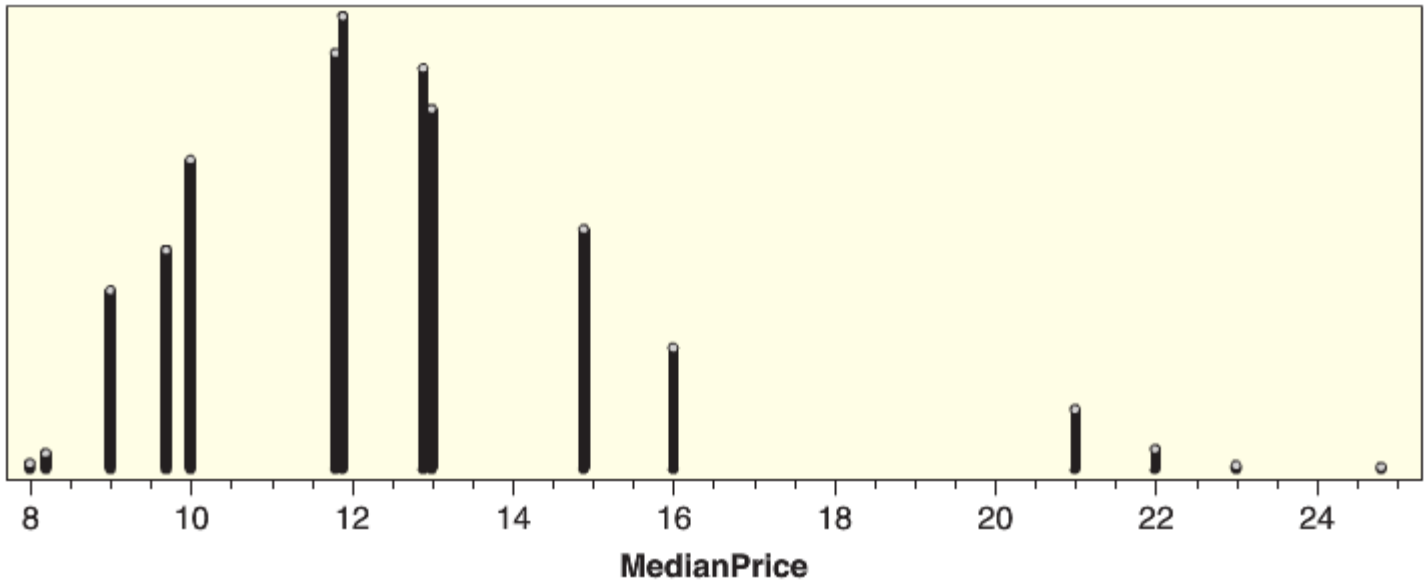
Findings CIs for many different parameters

This bootstrap method works for constructing confidence intervals for many different types of parameters!

Caution: the bootstrap does not always work

Always look at the bootstrap distribution, if it is poorly behaved (e.g., heavily skewed, has isolated clumps of values, etc.), you should not trust the intervals it produces.

Median prices of Mustangs



Homework 4

Homework 4 has been posted

- Use the link on Canvas to access homework 4 on R Studio Cloud
- Due on Gradescope at 11:30pm on Sunday February 16th