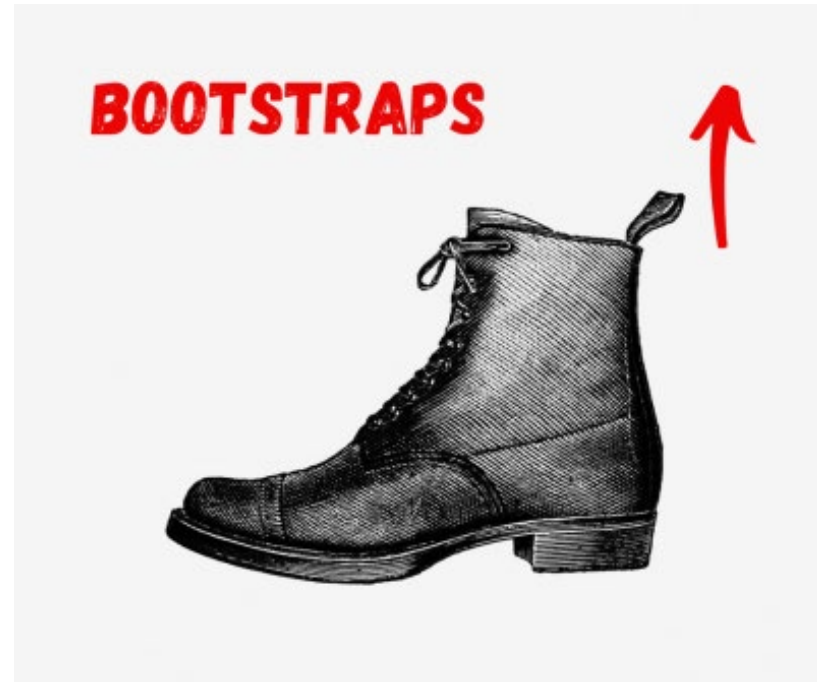


Review of confidence intervals and introduction to the bootstrap

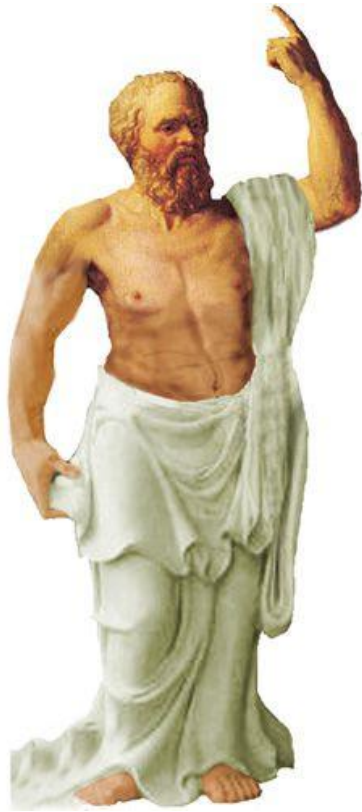


Overview

Review: sampling distributions and confidence intervals

The bootstrap

Review of sampling distributions, standard errors and confidence intervals



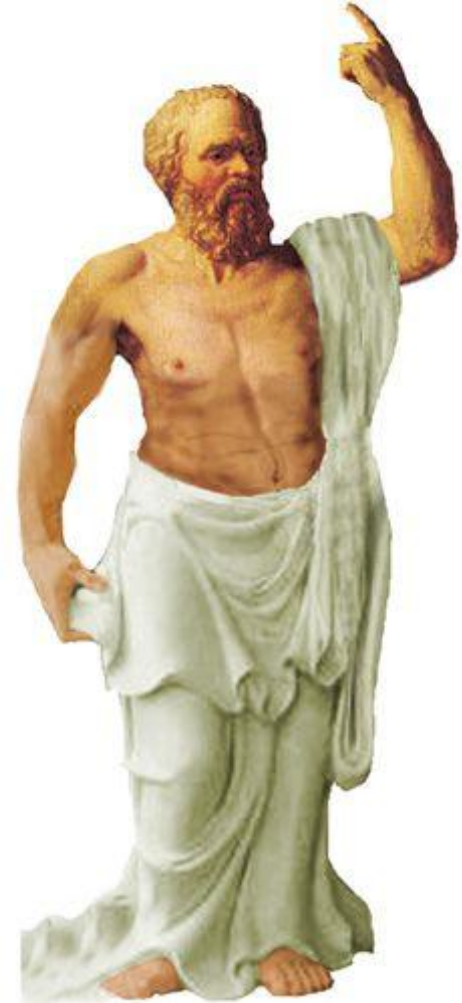
Review of confidence intervals and sampling distributions

Question₁: Who is this?

- A: Socrates!

Pause the video to answer questions on Canvas quiz!

- If you do not know an answer, enter a ? and be sure to review the answer
- Then continue on to the next question in the video
 - Having the quiz and video in separate tabs on your web browser will help reduce scrolling.



Sampling distributions

Q₂: 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

Q₃: What does a sampling distribution show us?

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

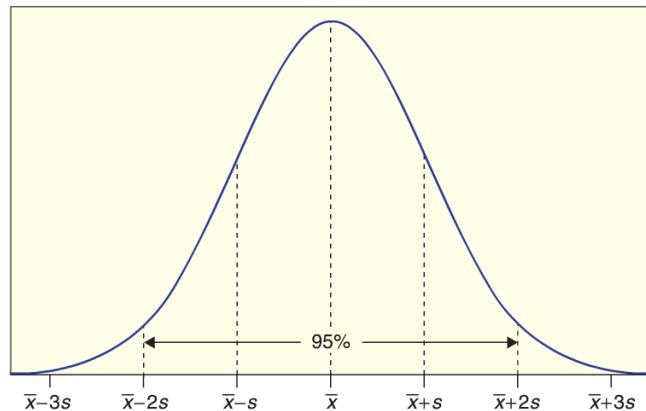
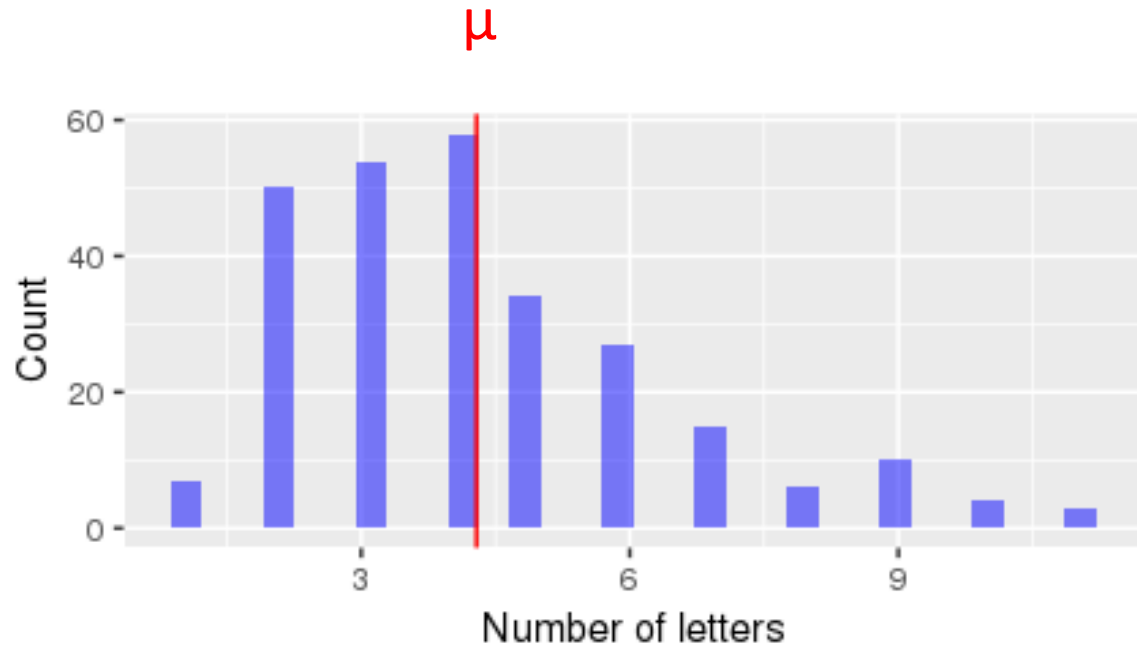
Art time



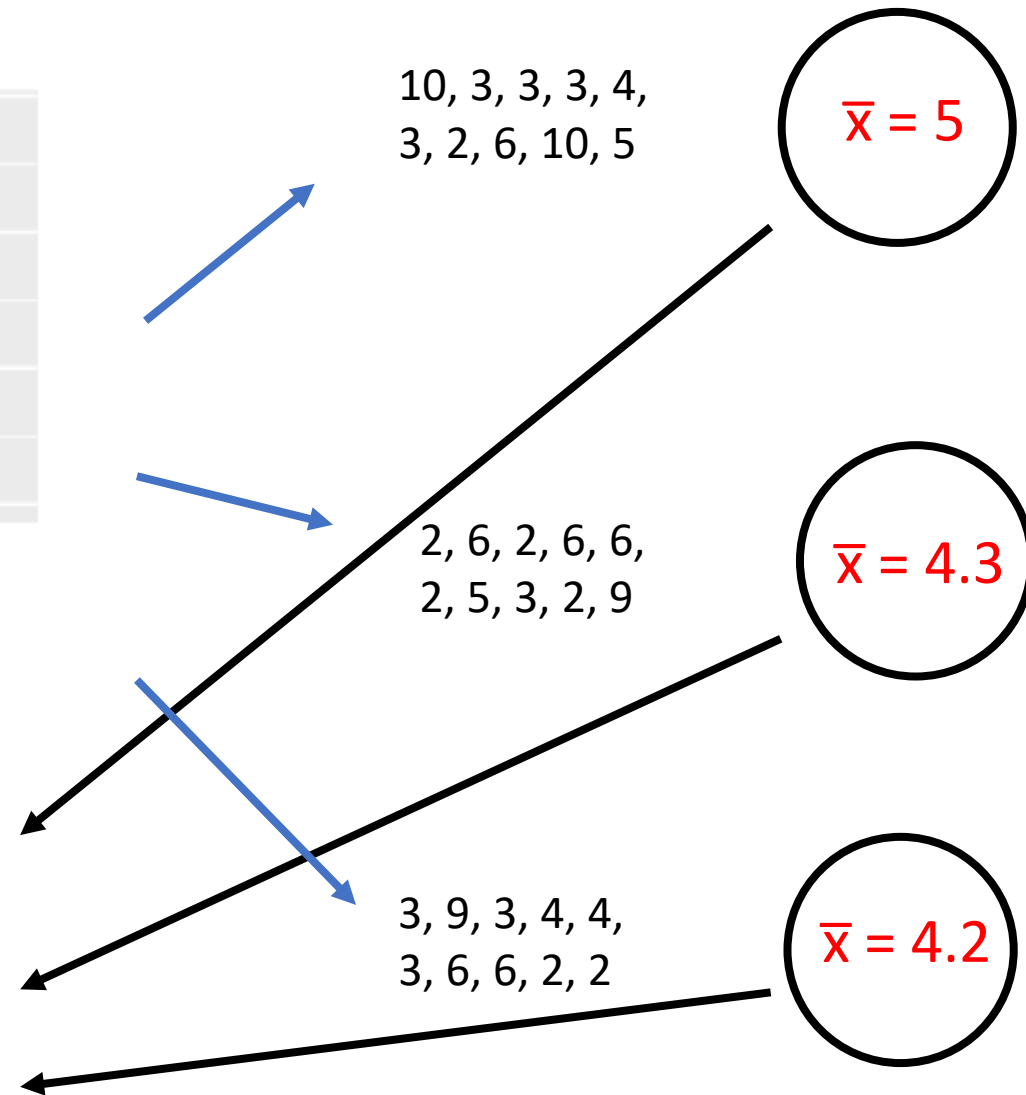
Please draw:

- 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



Sampling distribution!



[Gettysburg sampling distribution app](#)

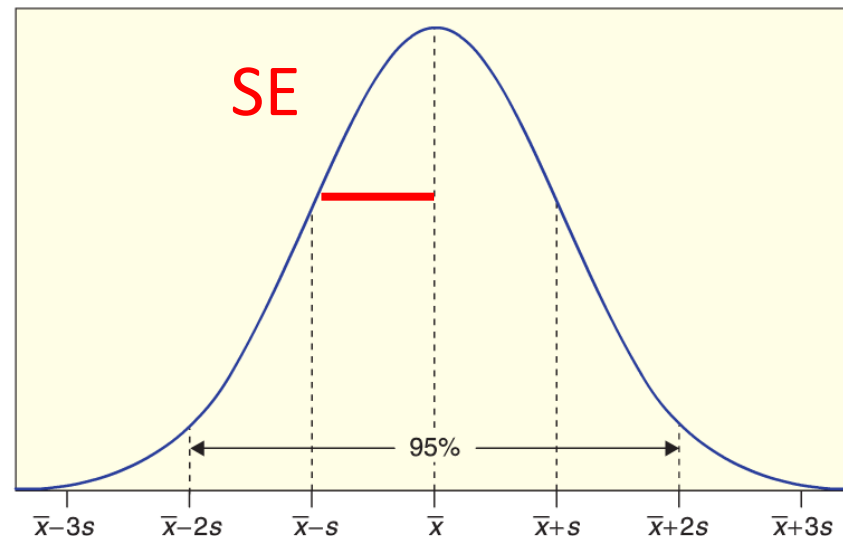
The standard error

Q₄: What is the **standard error**?

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

Q₅: What symbol do we use to denote the standard error?

- SE




Sampling distribution in R

Q₆: 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?

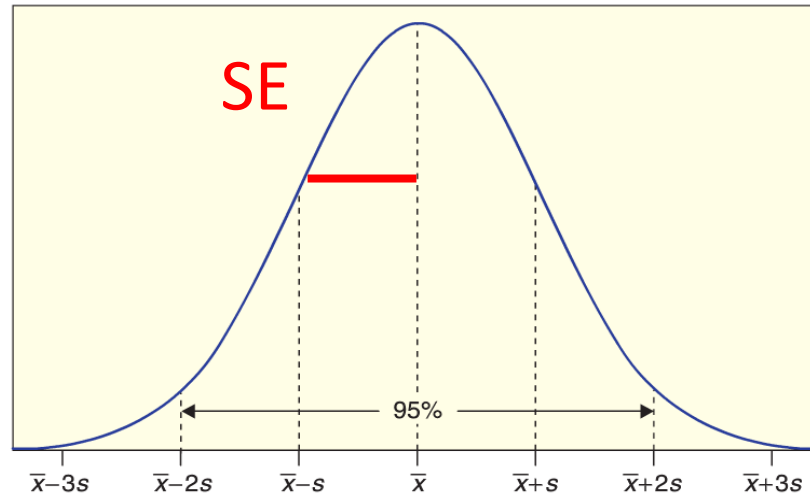
```
sampling_dist <- do_it(10000) * {  
  curr_sample <- get_sample()  
  mean(curr_sample)  
}
```

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



```
SE_mean <- sd(sampling_dist)
```

The standard error



Q₇: What does the size of the standard error tell us?

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

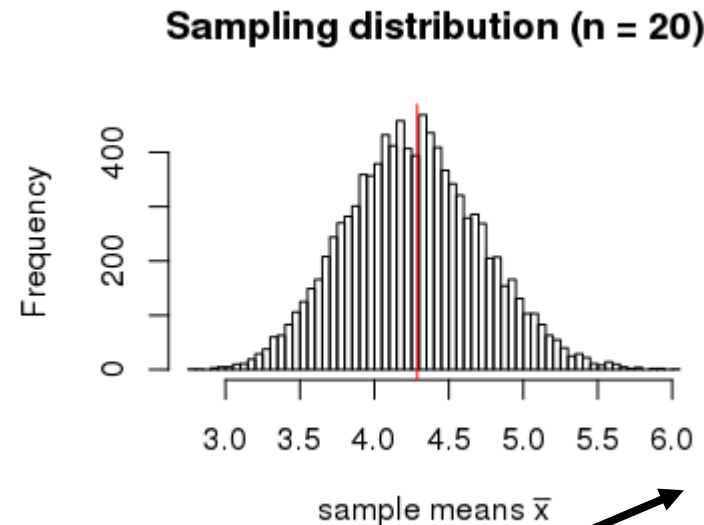
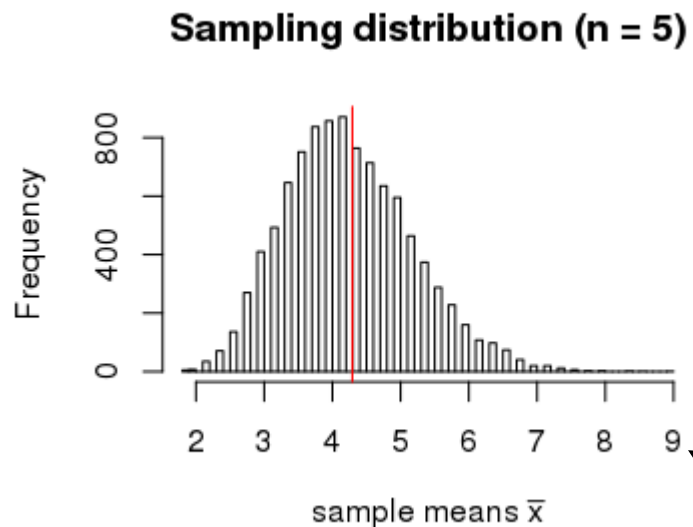
Q₈: What would it mean if there is a large SE?

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

Q₉: What are two ways that sampling distribution for the mean \bar{x} changes 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

Q₁₁: 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



Q₁₂: What is the **confidence level**?

- A₂: The **confidence level** is the percent of all intervals that contain the parameter

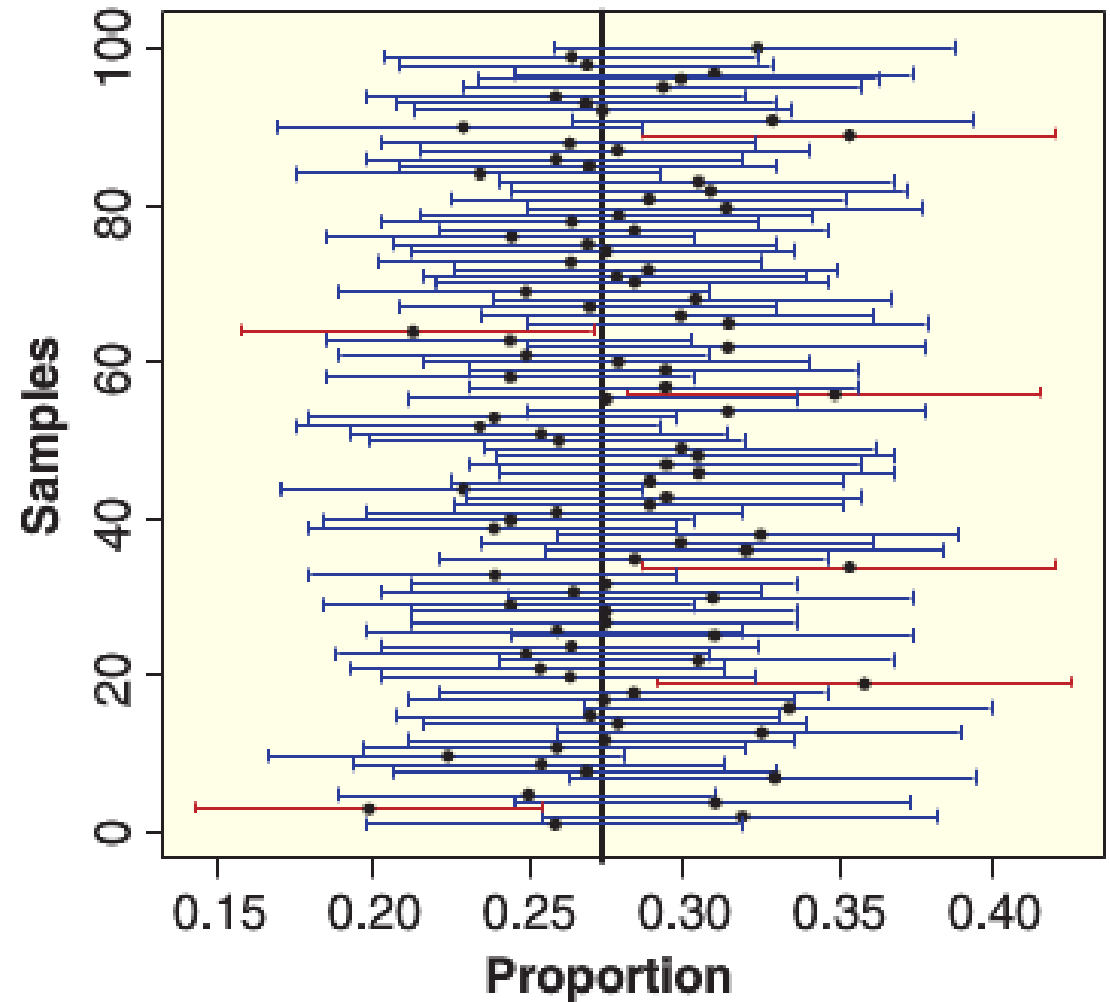
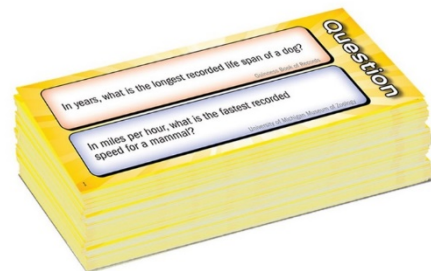


Confidence Intervals

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

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

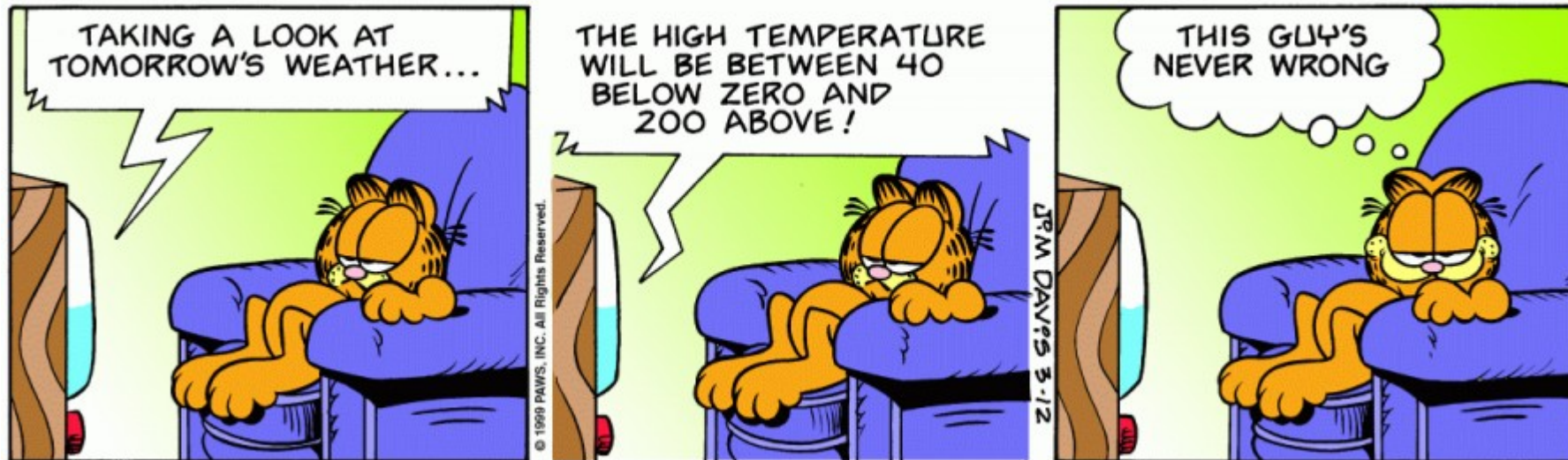
Right???



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!



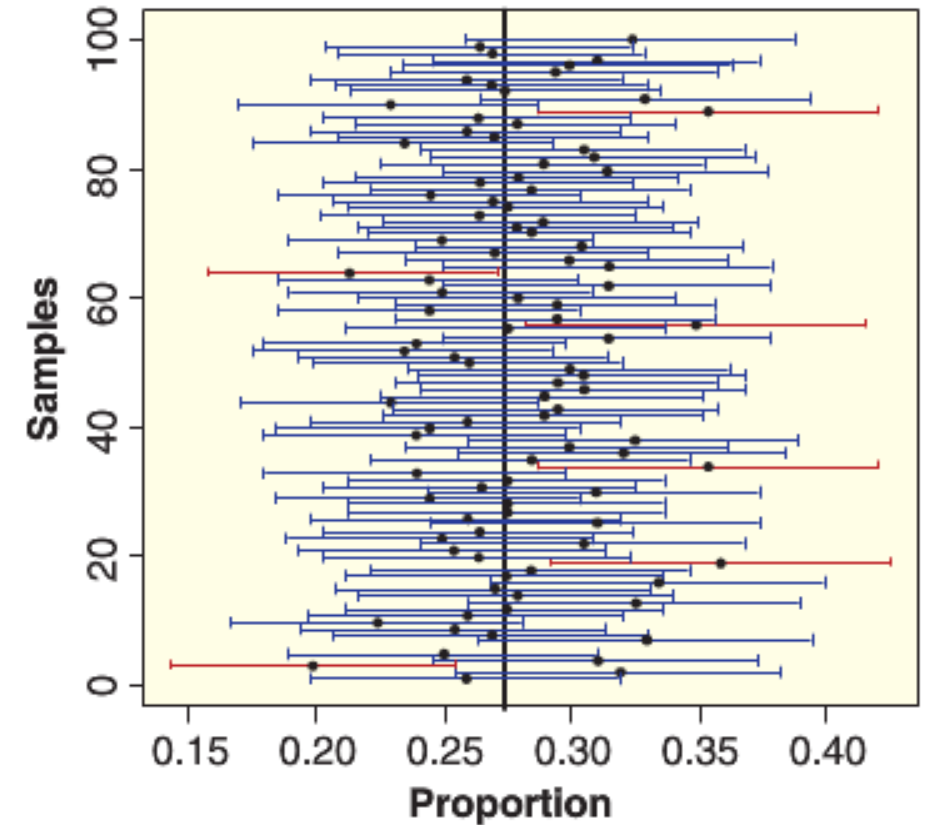
Confidence Intervals

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

- No ☹️

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)



Shapes of sampling distributions

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

A: Normal!



Normal distributions

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

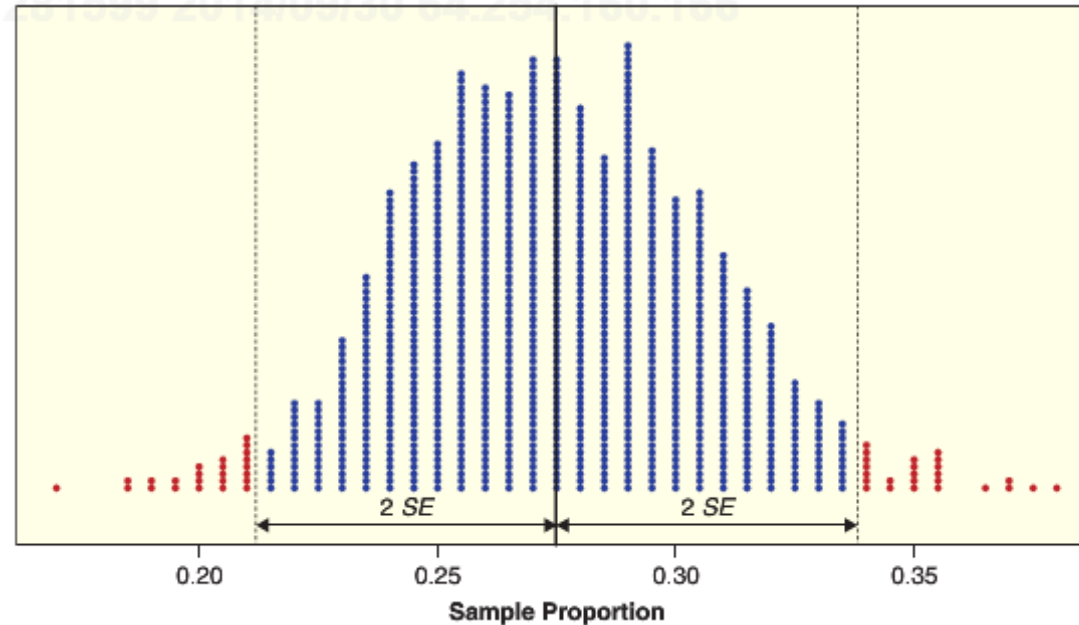
A: 95%



Sampling distributions

Q₁₇: 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%

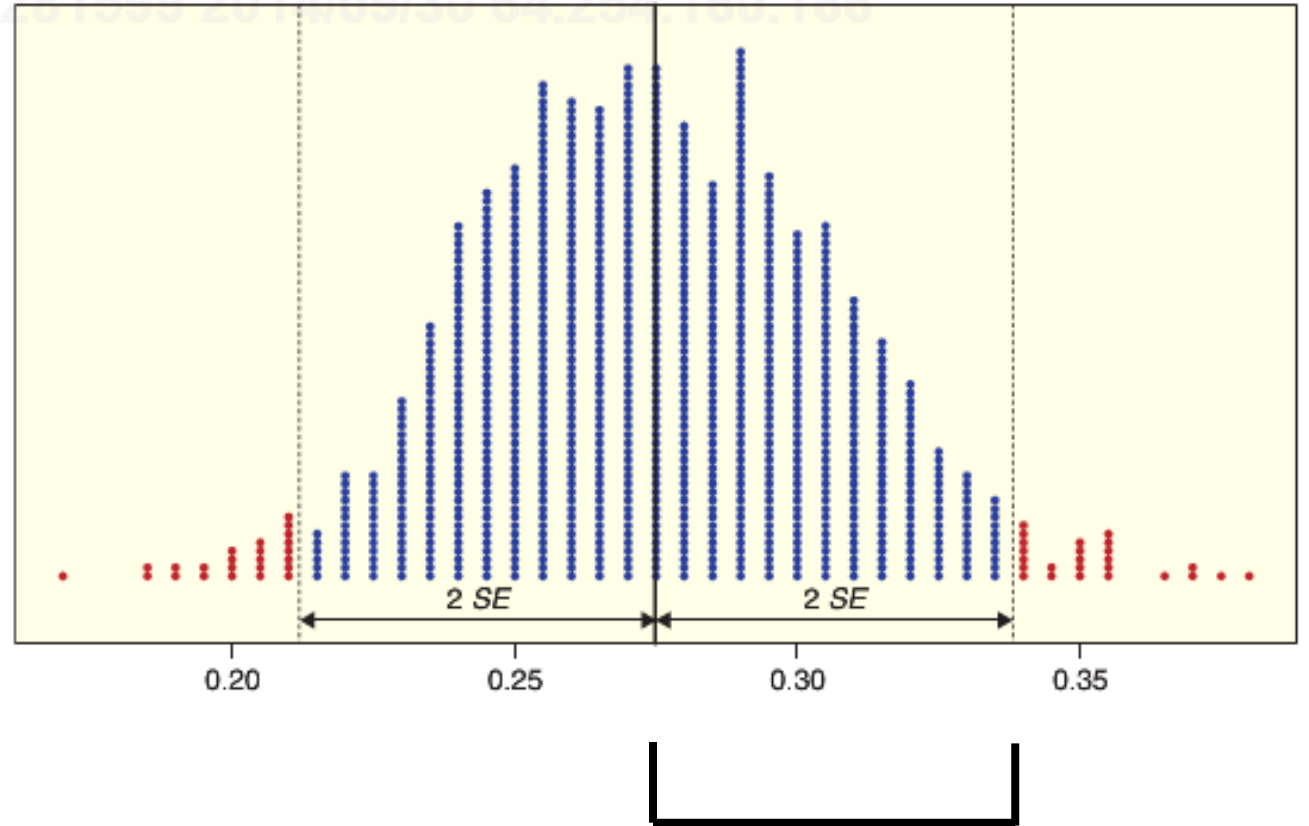


Q₁₈: If we had a *statistic value* and the value of the *SE*, could we compute a 95% confidence interval?

A: Depends/Yes, assuming the sampling distribution is normal, which it often is.

Confidence intervals

Q₁₉: What is a formula we can use to calculate 95% confidence intervals?

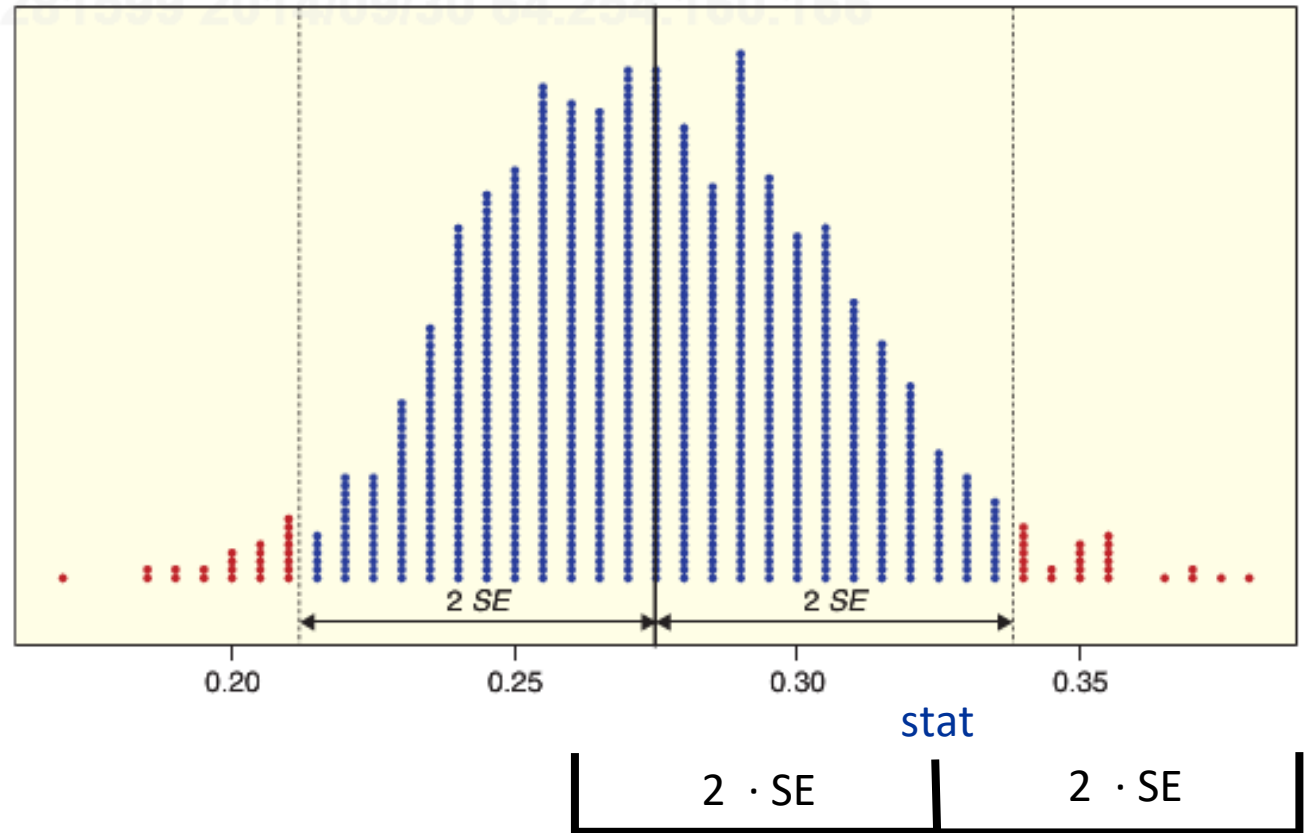


95% confidence interval: $\text{stat} \pm 2 \cdot \text{SE}$ → Q₂₀: What is this quantity called?
A: Margin of error

Confidence intervals

Q₂₁: For a 95% confidence interval, how far can the center of the confidence interval be from the parameter until the interval does overlap with the parameter value?

- A: $2 \cdot SE$



Confidence interval

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

Sampling distributions

Q_{22} : 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...

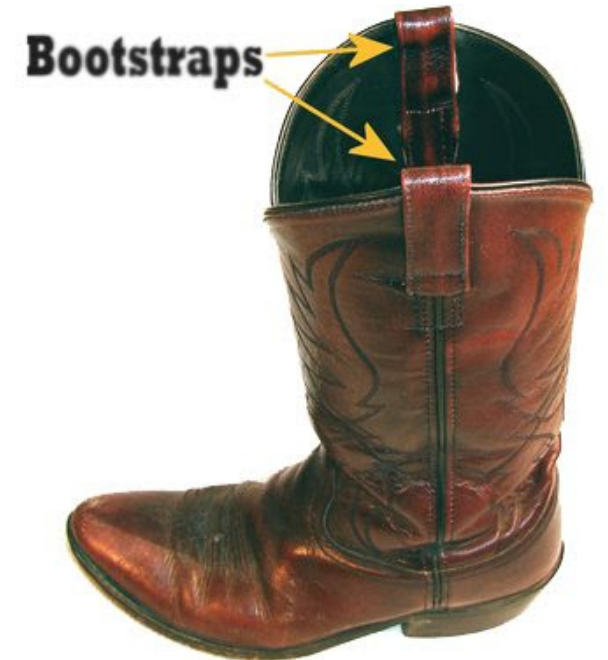


Sampling distributions

Q₂₃: 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} **from a single sample of data**
2. Then use: $\text{stat} \pm 2 \cdot \hat{SE}$ to get the 95% CI



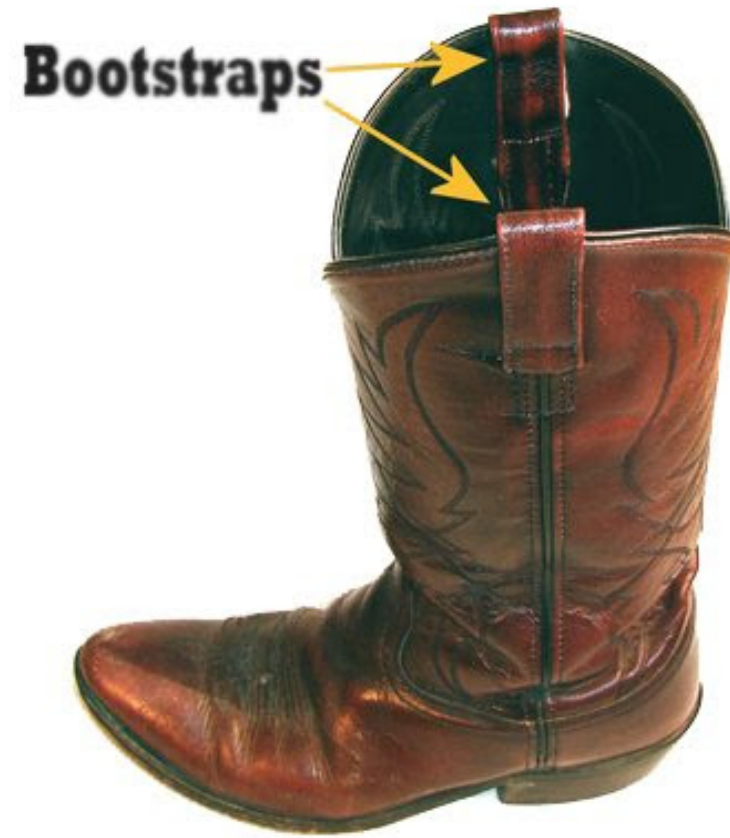
Confident intervals

Q₂₄: Do you feel confident about confidence intervals?

- A: I hope so!
- If not: review material, ask questions on Ed Discussions, and come to office hours!



The bootstrap



The bootstrap

The bootstrap is a method to estimate the standard error

- \hat{SE} is an estimate for SE
- We will use the symbol SE^* as the *bootstrap* estimate for SE (rather than \hat{SE})

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



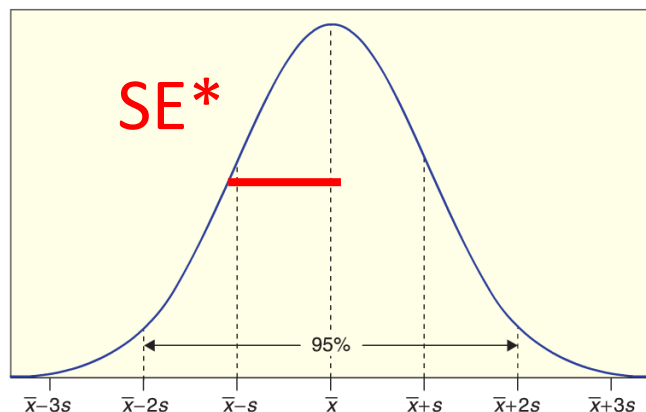
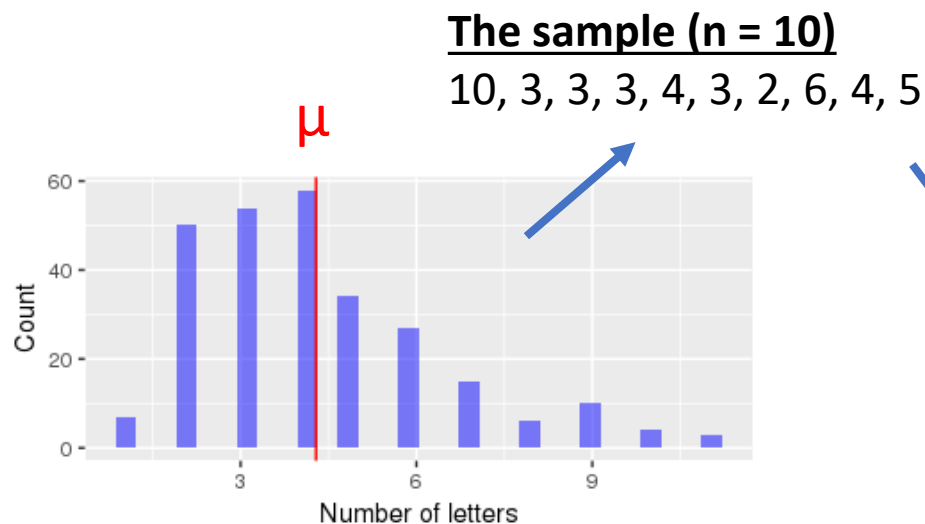
Plug-in principle

Suppose we get one sample of size n from a population

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



Bootstrap distribution!

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

$$\bar{x}^* = 4$$

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

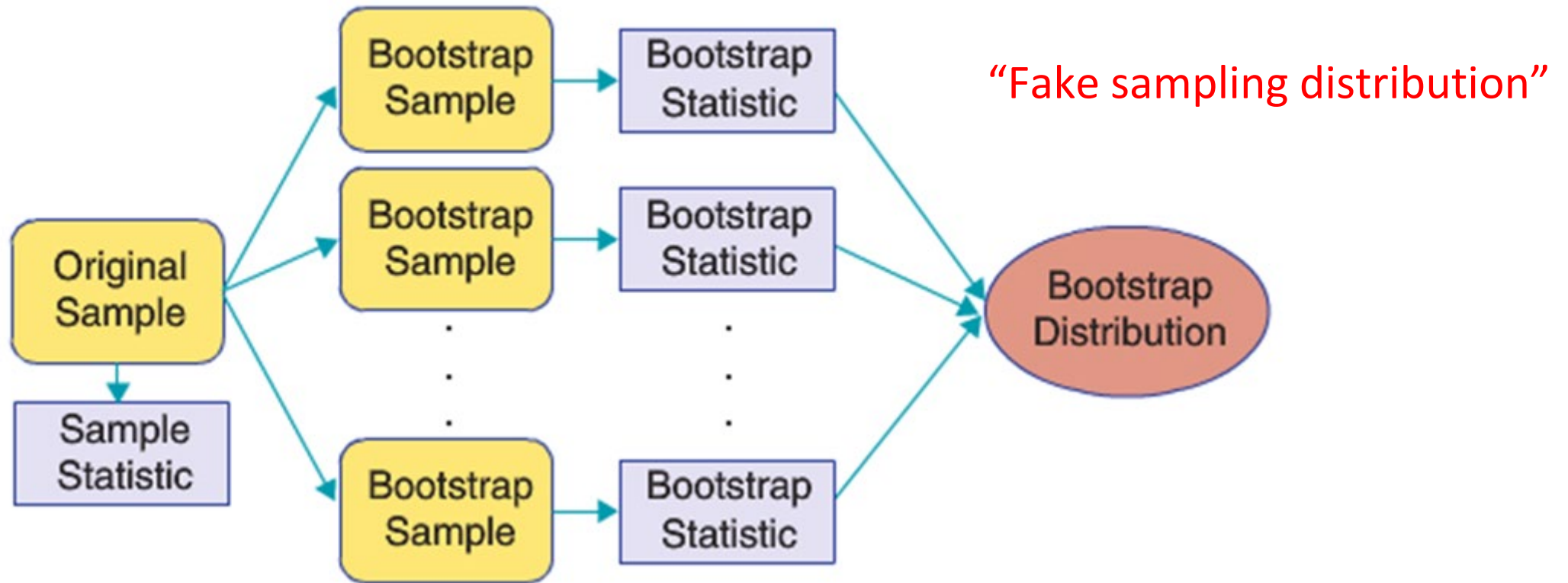
$$\bar{x}^* = 4.1$$

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

$$\bar{x}^* = 3.9$$

Notice there is no 9's in the bootstrap samples

Bootstrap process



95% Confidence Intervals

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

$$\text{Statistic} \pm 2 \cdot SE^*$$

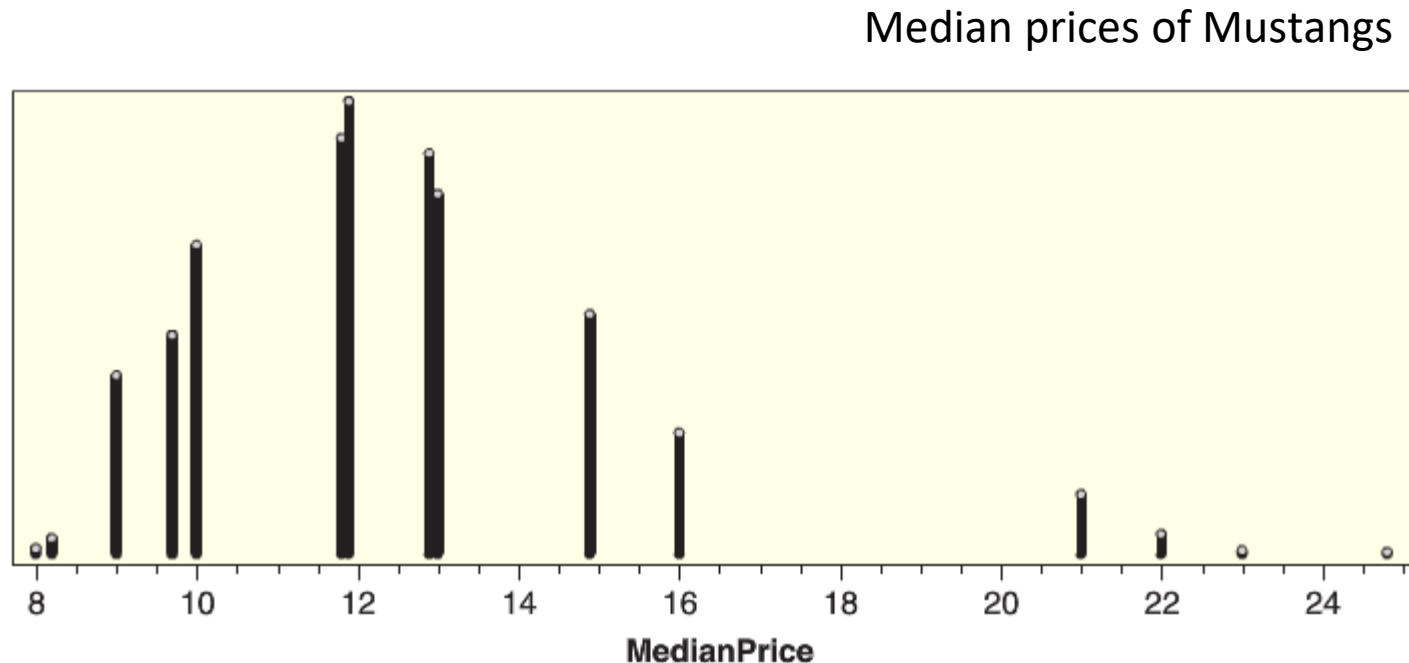
Where SE^* is the standard error estimated using the bootstrap

Findings CIs for many different parameters

The 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.





I believe in pulling yourself up by
your own bootstraps. I believe it is
possible — I saw this guy do it once
in Cirque du Soleil. It was magical.

— *Stephen Colbert* —

AZ QUOTES