Review of confidence intervals and introduction to the bootstrap

Overview

Quickly go through the review of sampling distributions and confidence intervals

- Answer any questions you have

The bootstrap

Calculating bootstrap confidence intervals in R

Announcement: homework 4

Homework 4 has been posted

- Due on Gradescope at 11pm on Sunday February 18th

How was homework 3?

Review of confidence intervals and sampling distributions

Question$_0$: Who is this?

- Socrates (with sunglasses)!

Our goal: to create confidence intervals

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

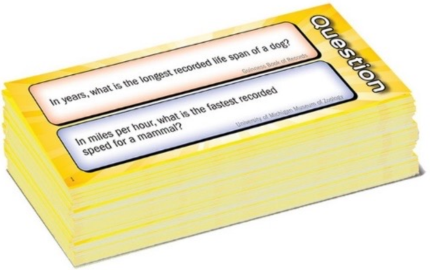
- A_2: The **confidence level** is the percent of all intervals that contain the parameter

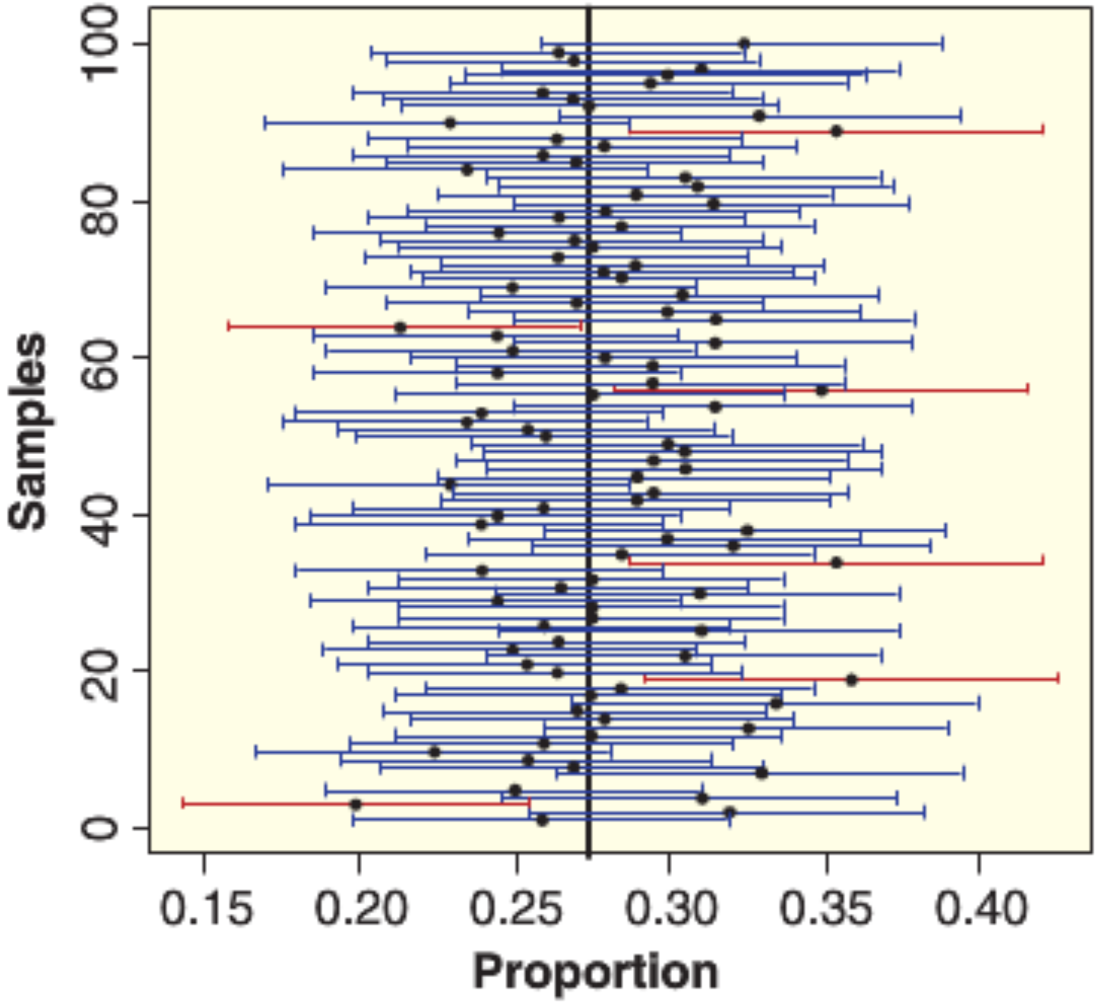
Confidence Intervals

Q: 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???


Question
In years, what is the longest recorded life span of a dog?
Guinness Book of Records
In miles per hour, what is the fastest recorded speed for a mammal?
University of Michigan Museum of Zoology



Confidence Intervals

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

- Yes!

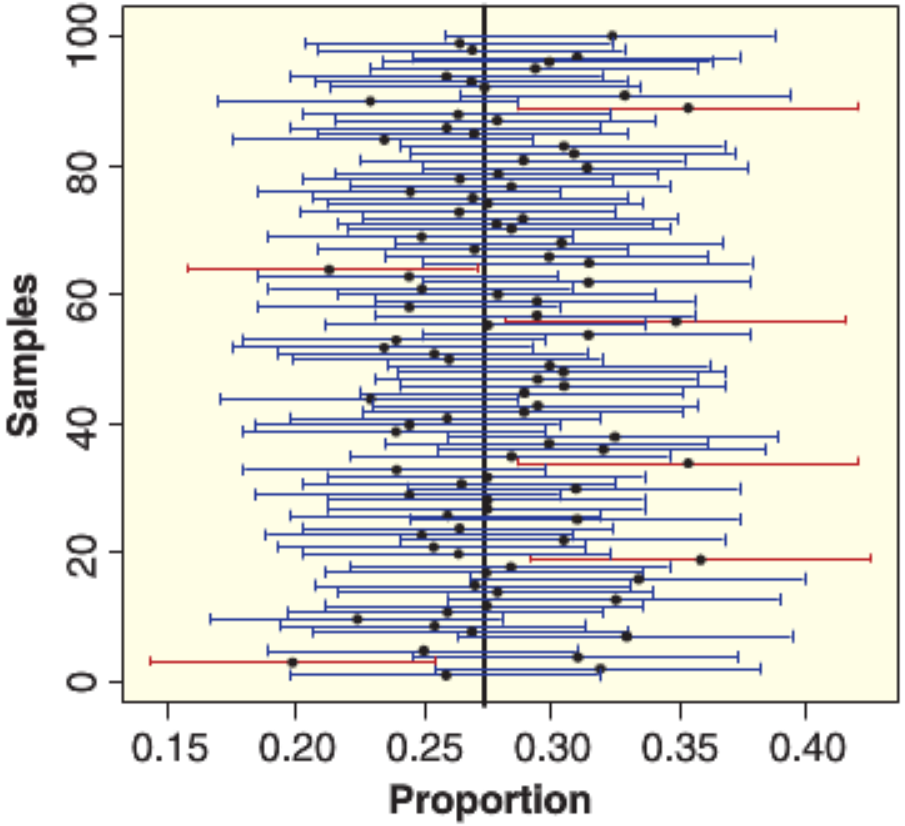
Confidence Intervals

Q: 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)



Confident intervals

Q: Do you feel confident what a confidence interval is?

Questions?

How can we create confidence intervals?

Sampling distributions

Q: 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: What does a sampling distribution show us?

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

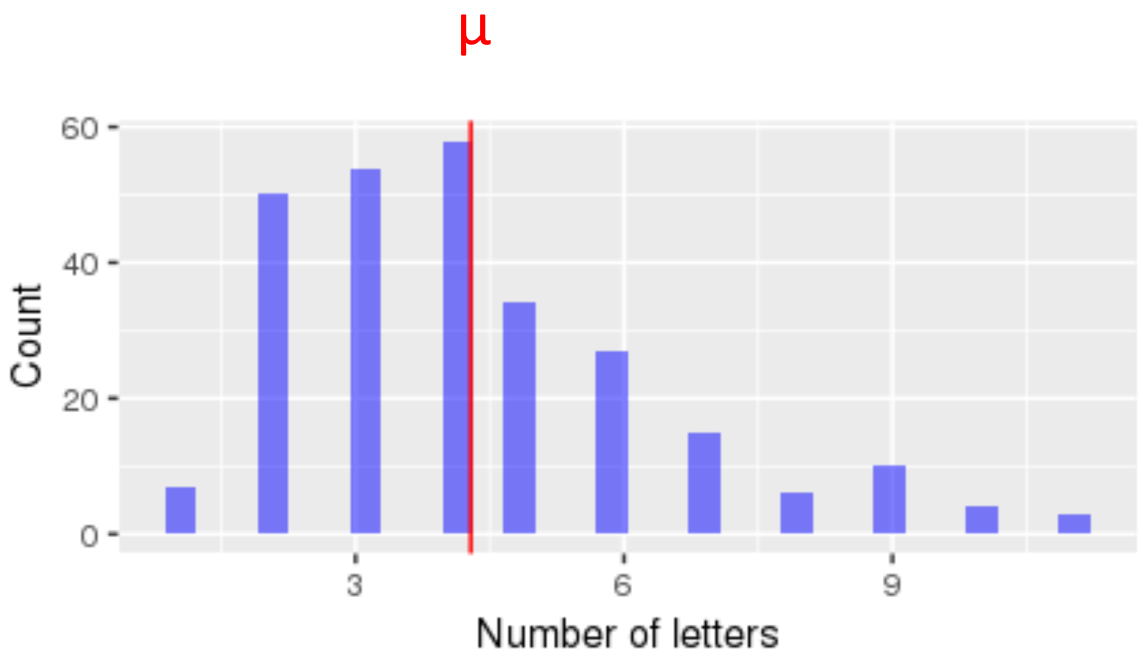
Art show time

On the online class quiz you drew:

- 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

Let's look at your art!

Gettysburg address word length sampling distribution



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

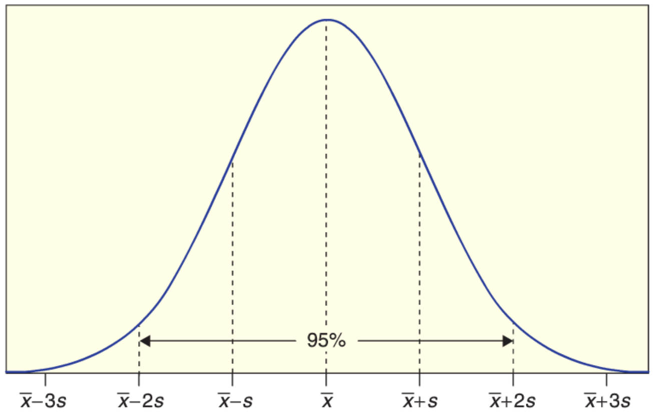
$\bar{x} = 5$

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

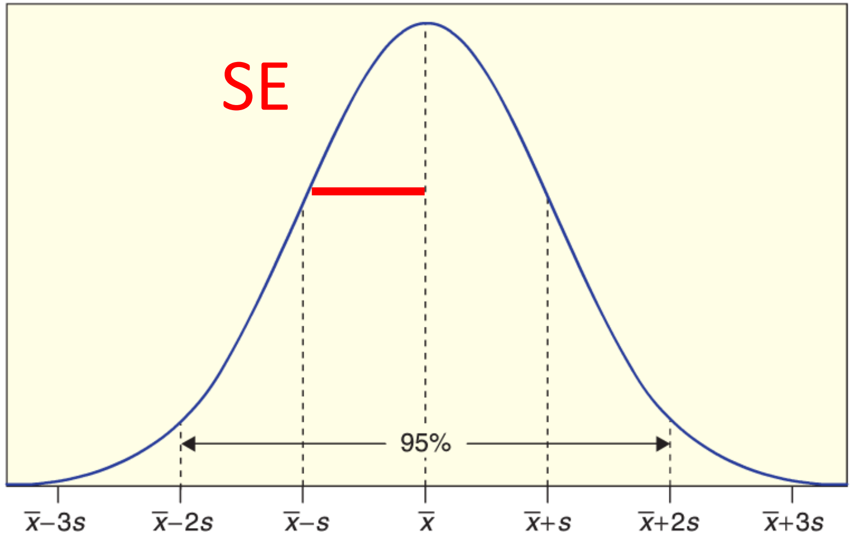
The standard error

Q: What is the **standard error** and **what symbol** do we use to denote it?

- The **standard error** of a statistic is the standard deviation of the sampling distribution
- The symbol we use to denote the standard error is SE

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

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



Q: 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

Sampling distribution in R

Q: 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)

}

SE_mean <- sd(sampling_dist)
```

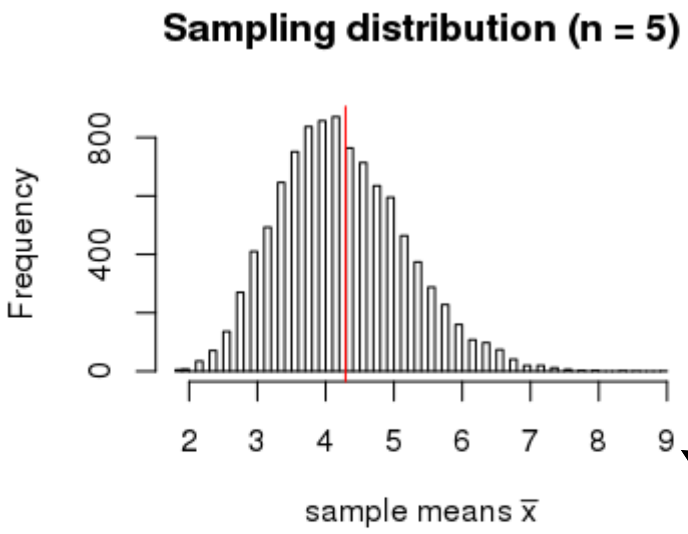
What would happen if we added set.seed(100) here?

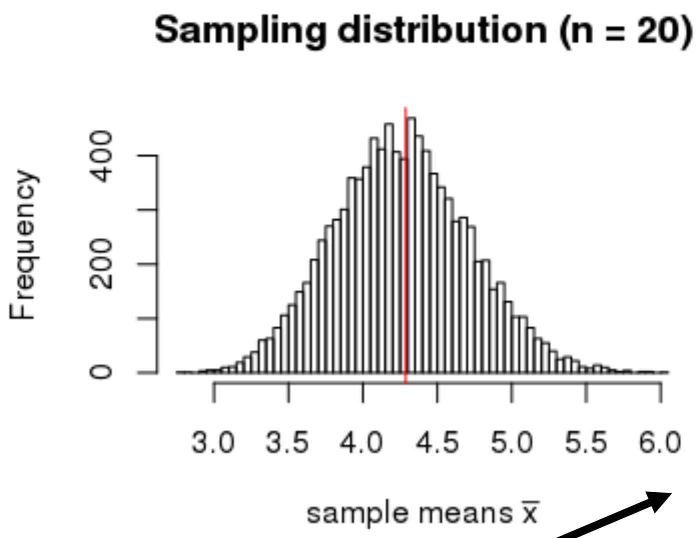
What symbol should we use for this quantity? $\bar{x}_i$

Q: 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



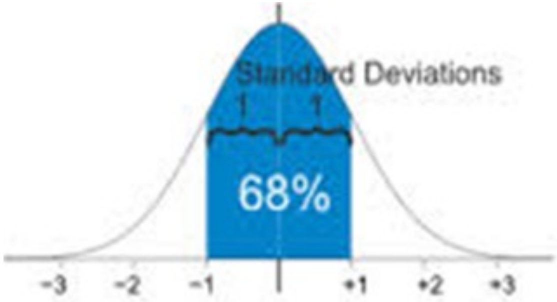


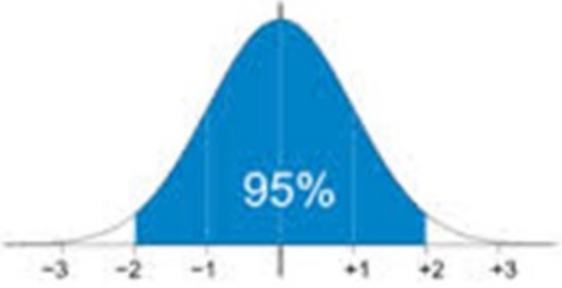
x-axis range 9 vs. 6

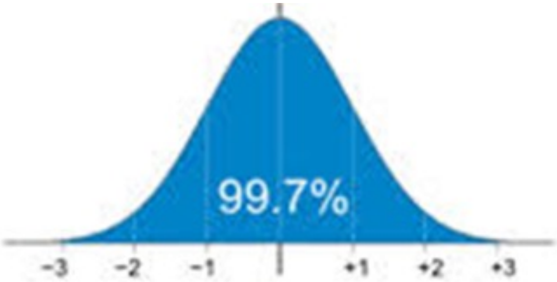
Shapes of sampling distributions

Q: What is a commonly seen shape for sampling distributions?

A: Normal!







Q: What percent of statistics (say $\bar{x}$) will lie 1, 2, and 3 SE away from μ?

A: 68% of $\bar{x}$'s are within ±1 SE from μ

A: 95% of $\bar{x}$'s are within ±2 SE from μ

Confidence intervals

Q: Suppose we create an interval that has these properties:

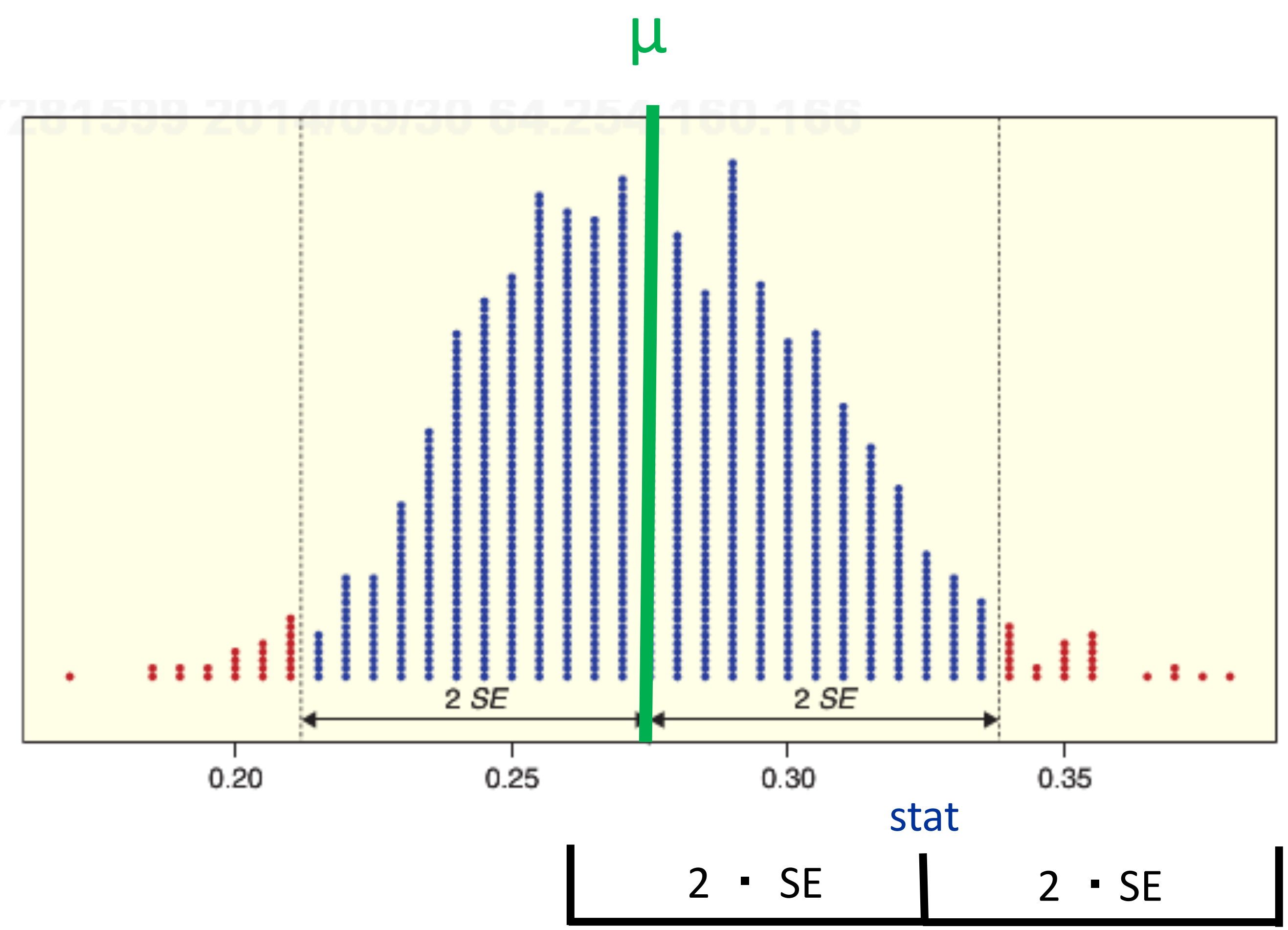
1. Centered at our statistic value
   - E.g. centered at $\bar{x}$

2. That as a width $\pm\ 2 \cdot SE$
   - i.e., our interval is $\bar{x}\ \pm 2 \cdot SE$

What percent of the time will intervals constructed this way overlap with μ?

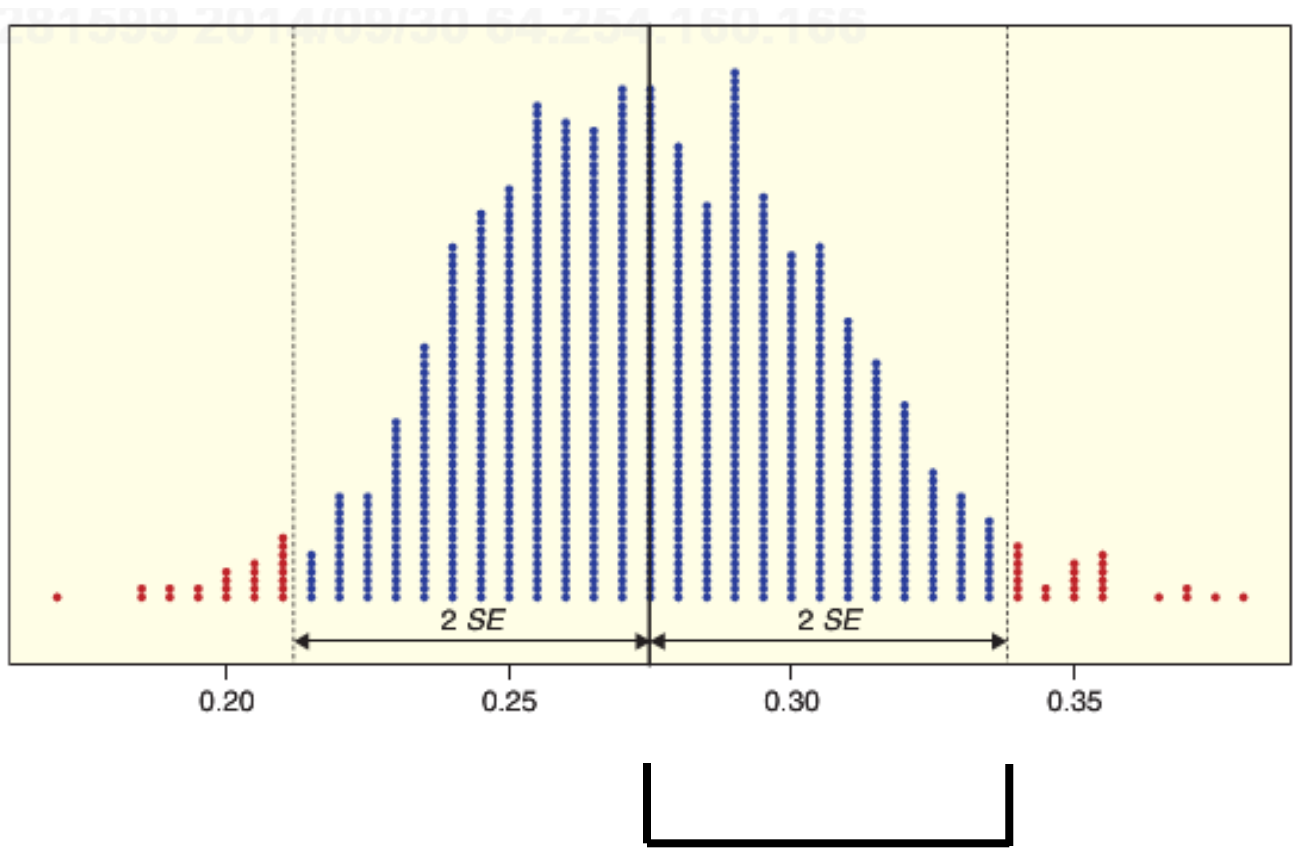
- A: 95% of the time

Q: why?



Confidence interval

Confidence intervals



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

95% confidence interval: $\text{stat} \pm \boxed{2 \cdot \text{SE}}$ → Q:What is this quantity called?

A: Margin of error

Sampling distributions

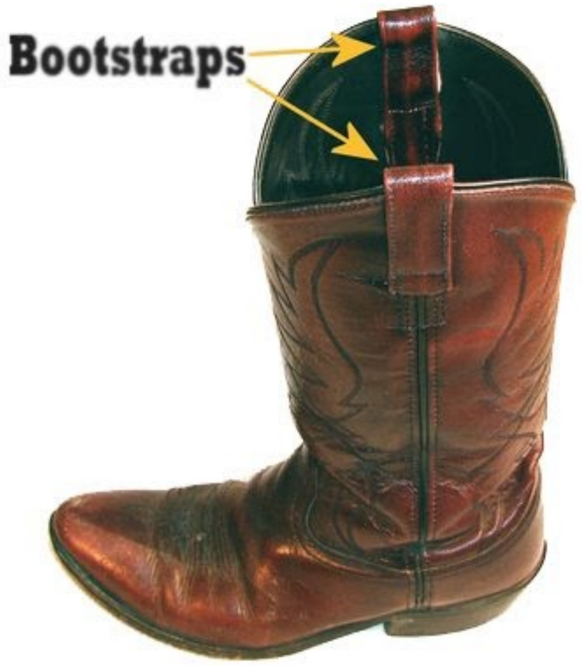
Q: Could we repeat the sampling process many times to create a sampling distribution and then calculate the SE?

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

Sampling distributions

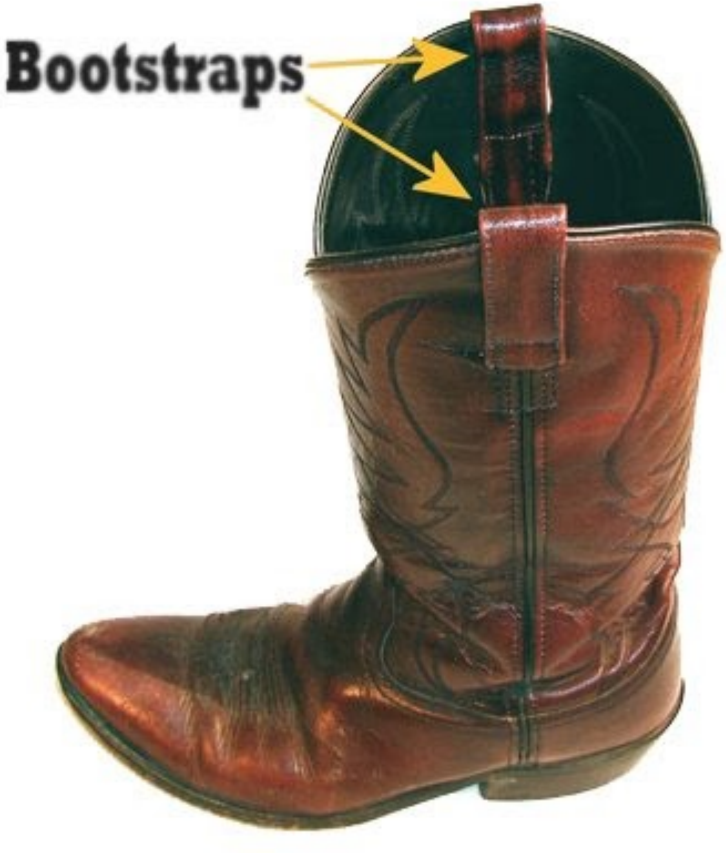
Q: If we can’t calculate the sampling distribution, what’s else could we do?

- A: We could pick ourselves up from the bootstraps



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

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 $\overline{x} \pm 2 \cdot$ SE* to get the 95% CI

Plug-in principle

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

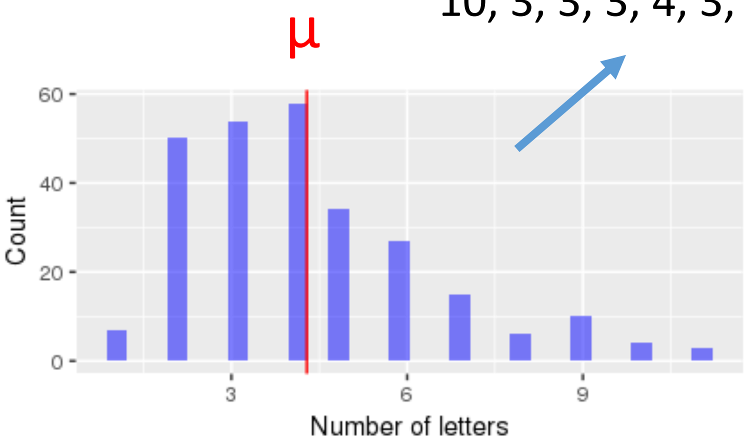
We pretend that this sample is the population   (plug-in principle)

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



<u>The sample (n = 10)</u>
10, 3, 3, 3, 4, 3, 2, 6, 4, 5

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

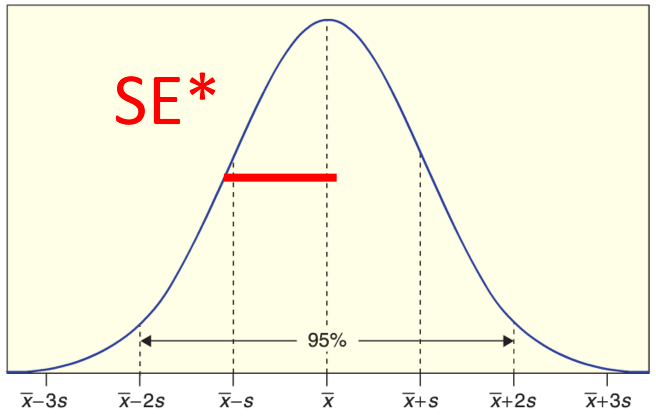
$\overline{x}^* = 4$

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

$\overline{x}^* = 4.1$

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

$\overline{x}^* = 3.9$



Bootstrap distribution!

Notice there is no 9's in the bootstrap samples

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

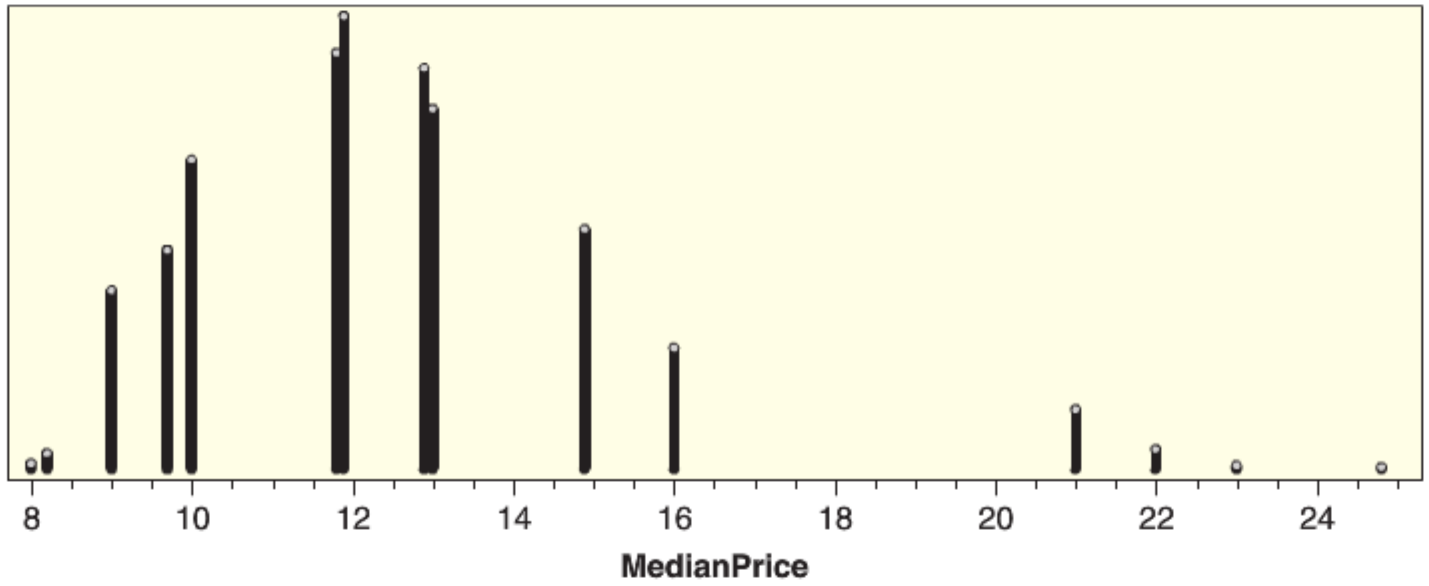
Findings CIs for many different parameters

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.

Median prices of Mustangs



Calculating bootstrap confidence intervals in R

What are the steps needed to create a bootstrap SE?

1. Start with a sample

2. Repeat steps 10,000 times

    a. Resample the points in the sample to get a bootstrap sample

    b. Compute the statistic of interest on the bootstrap sample

3. Take the standard deviation of the bootstrap distribution to get SE*

Sampling with replacement from a vector

```
my_sample <- c(3, 1, 4, 1, 5, 9)
```

To get a sample of size n = 6 with replacement:

```
> boot_sample  <-  sample(my_sample,  6,  replace = TRUE)
```

Sampling distribution in R

```
 my_sample <- c(21, 29, 25, 19, 24, 22, 25, 26, 25, 29)

bootstrap_dist <-  do_it(10000) * {

        curr_boot <- sample(my_sample , 10, replace = TRUE)
        mean(curr_boot)

}

SE_boot <- sd(bootstrap_dist)
```

Bootstrap confidence interval in R

```
obs_mean <- mean(my_sample)

CI_lower <-  obs_mean  - 2 * SE_boot

CI_upper <-  obs_mean  + 2 * SE_boot
```

Let's try it in R!