Review of confidence intervals and introduction to the bootstrap

Overview

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 February 15th 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 confidence intervals

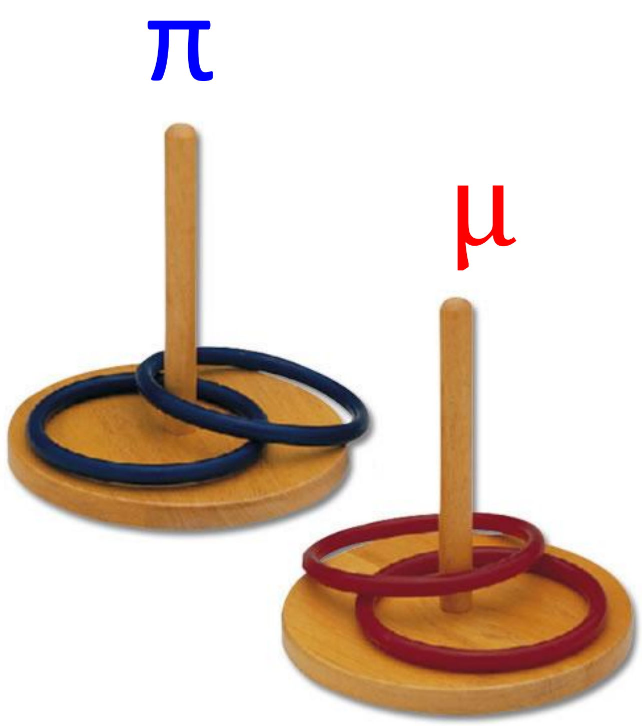
Question: What is a confidence interval?

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

- i.e., if the estimation were repeated many times, the interval will have the parameter x% of the time

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

Think ring toss…



A parameter exists in the world

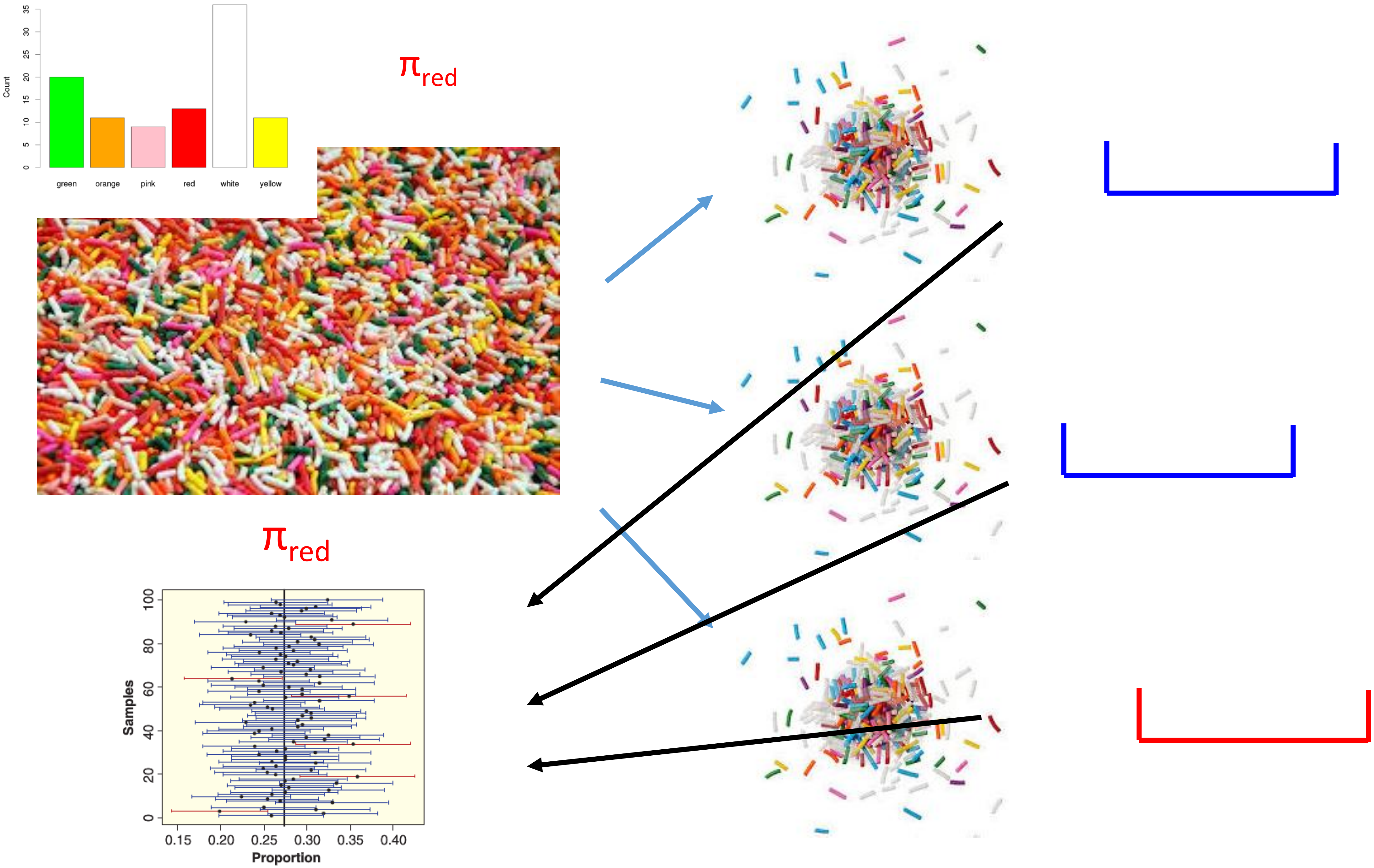
We toss intervals at it

- These are our confidence intervals

95% of those intervals capture the parameter

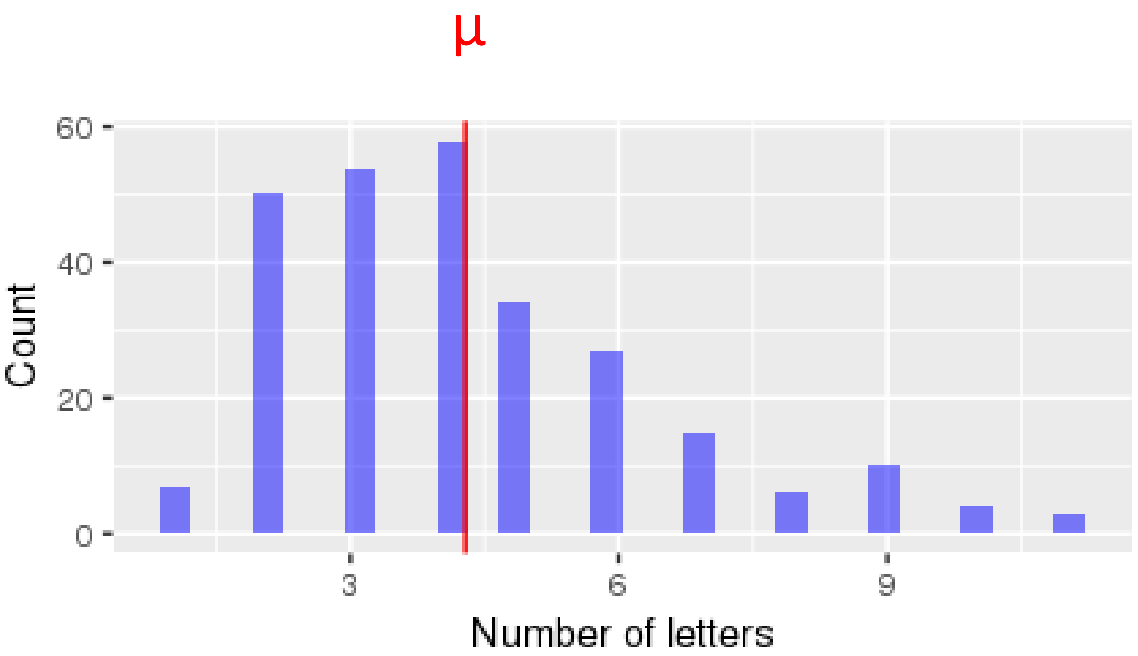
- (95% is our confidence level)

π_{red}


Count
0
5
10
15
20
25
30
35
green
orange
pink
red
white
yellow
Samples
0
20
40
60
80
100
0.15
0.20
0.25
0.30
0.35
0.40
Proportion

π_{red}

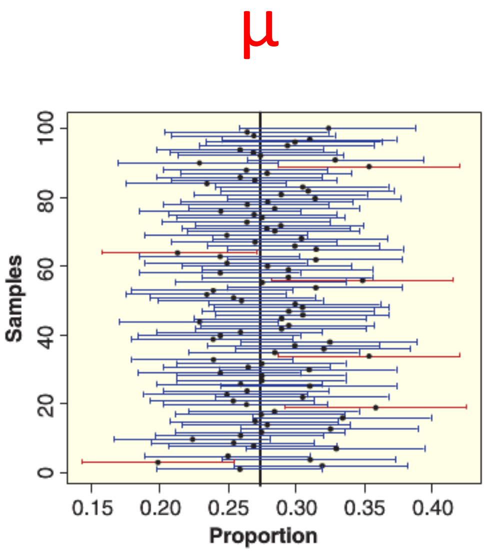
Gettysburg address word length sampling distribution

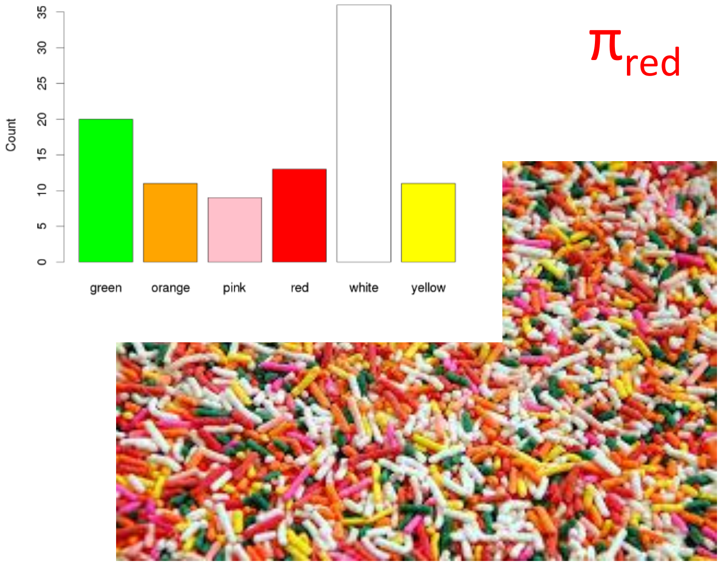


10, 3, 3, 3, 4,
3, 2, 6, 10, 5

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

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

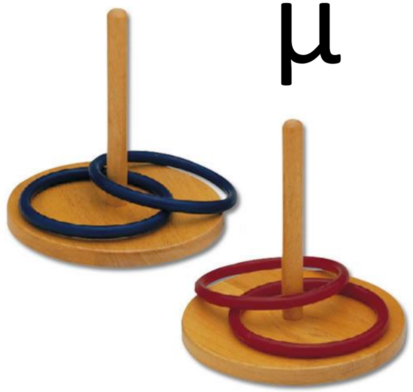




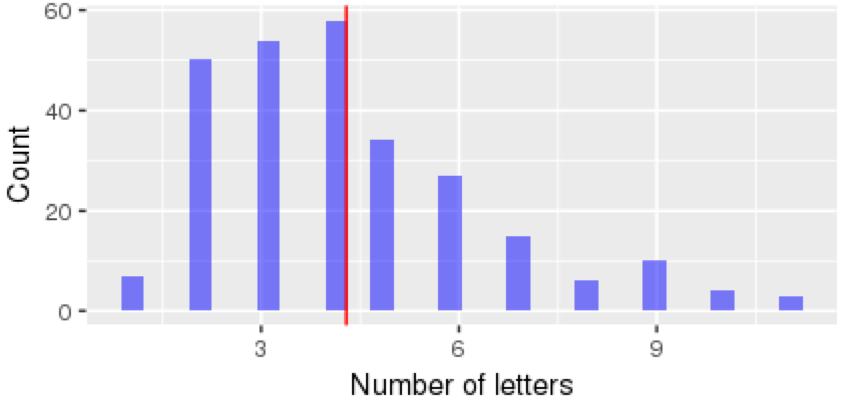
π_{red}

π

μ



μ

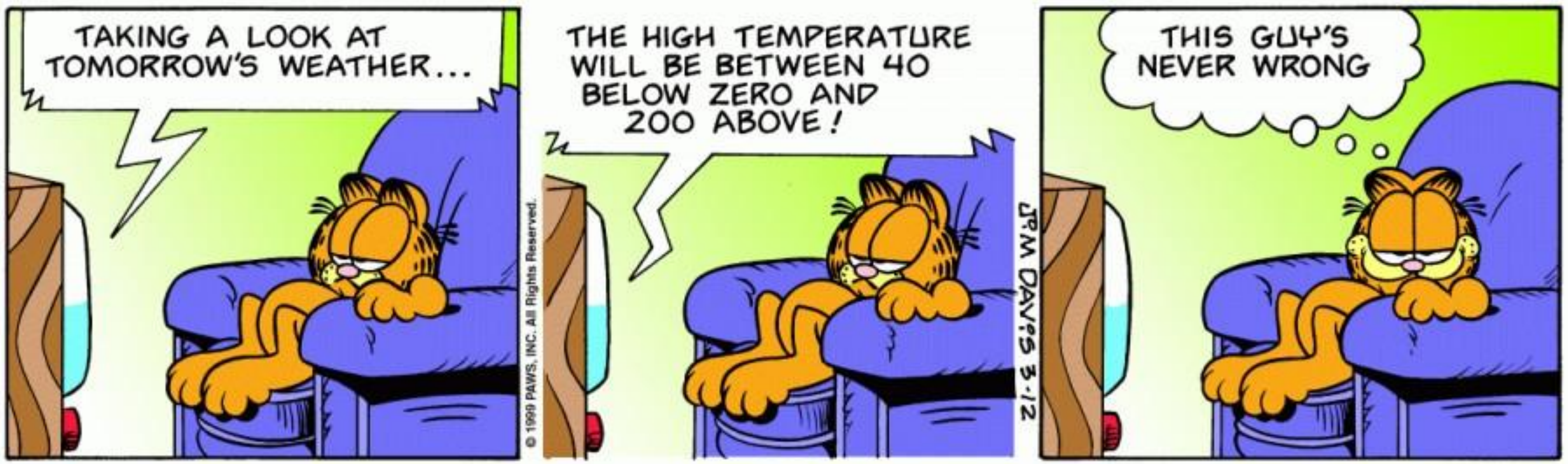


10, 3, 3, 3, 4,
3, 2, 6, 10, 5

For a 95% confidence level, 95% of the intervals we create will have the parameter in them

Confidence Intervals

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

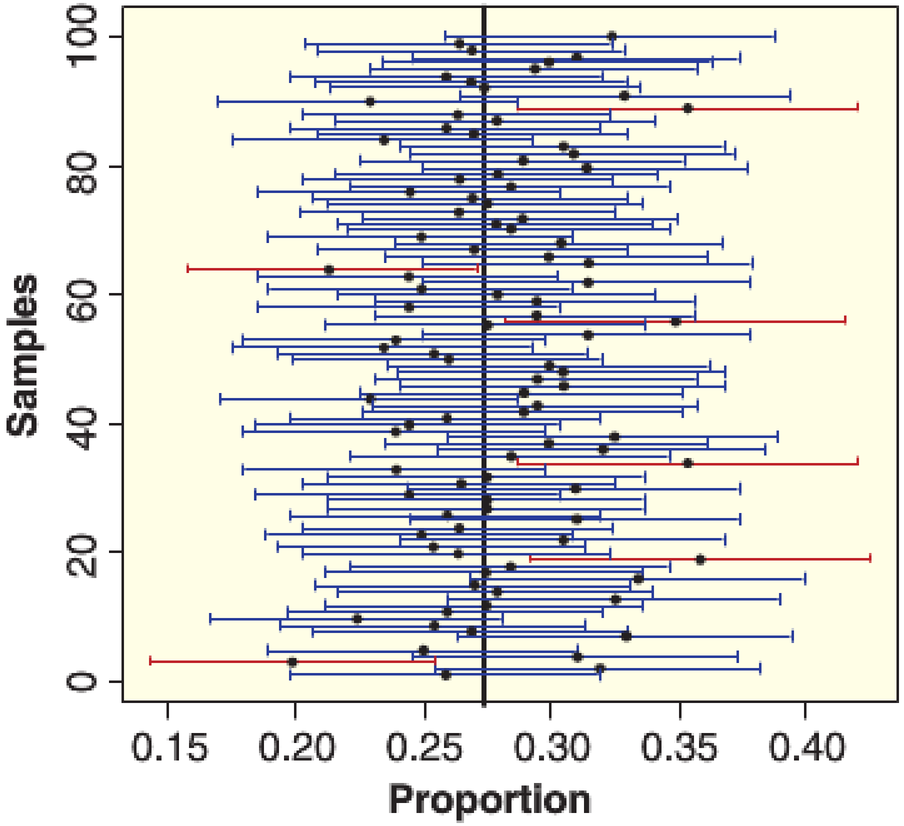

TAKING A LOOK AT TOMORROW'S WEATHER...
THE HIGH TEMPERATURE WILL BE BETWEEN 40 BELOW ZERO AND 200 ABOVE!
THIS GUY'S NEVER WRONG
© 1999 PAWS, INC. All Rights Reserved.
JPM DAVIS 3-12

Confidence Intervals

Q: For any given confidence interval, do we know whether it has really captured the parameter?

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

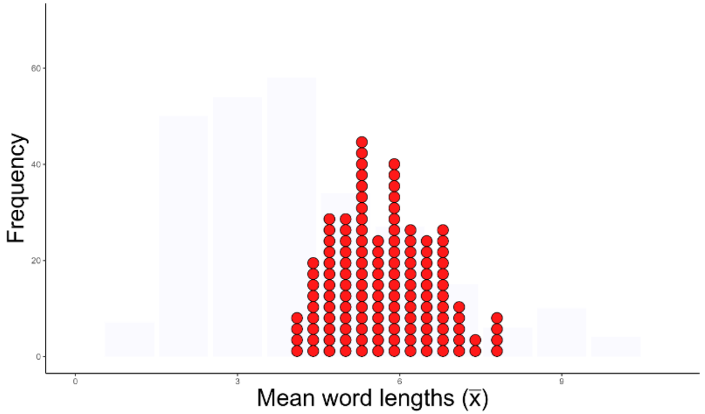
(for a 95% confidence interval)



Review of sampling distributions

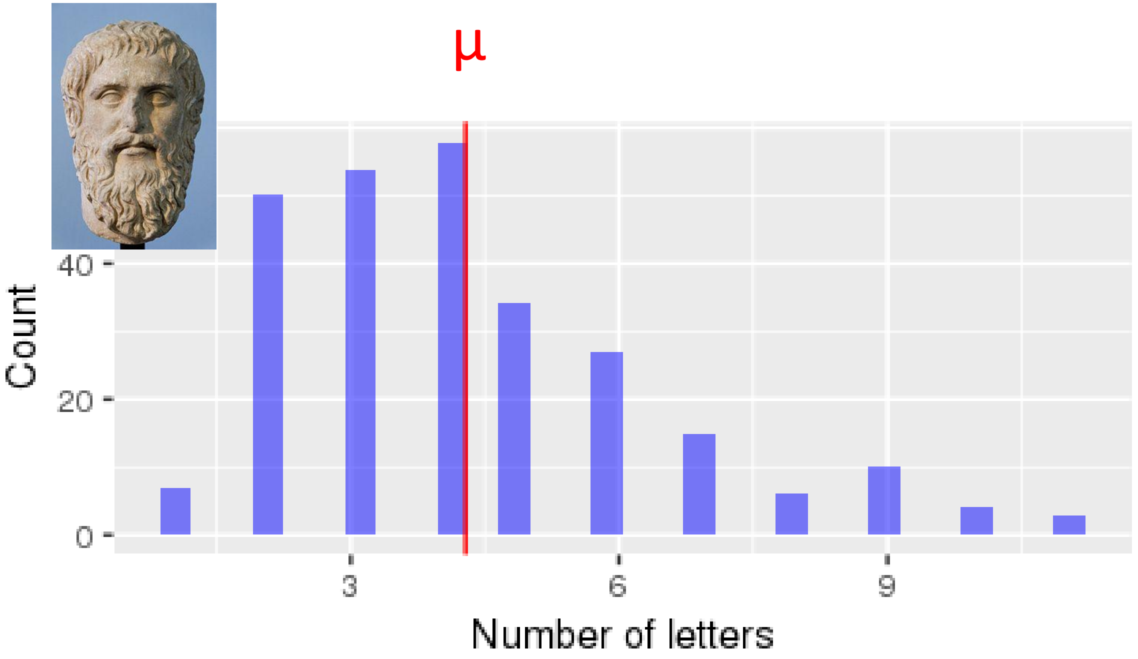
Sampling distributions

Q_2: What is a sampling distribution?



Q_3: What does a sampling distribution show us?

Gettysburg address word length sampling distribution



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

$\bar{x} = 5$

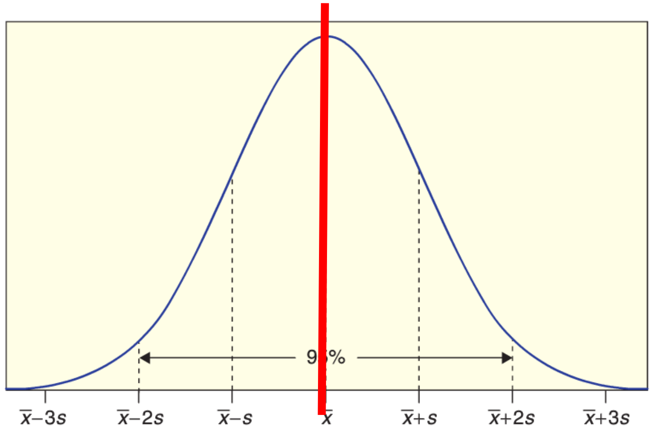
2, 6, 2, 6, 6,
2, 5, 3, 2, 9

$\bar{x} = 4.3$

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

$\bar{x} = 4.2$

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



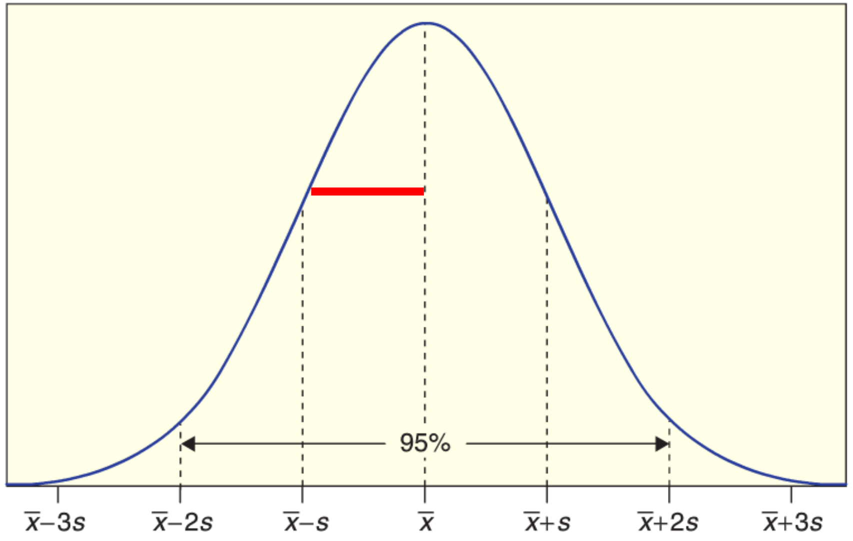
Sampling distribution!

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

The standard error

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

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



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)

}
```

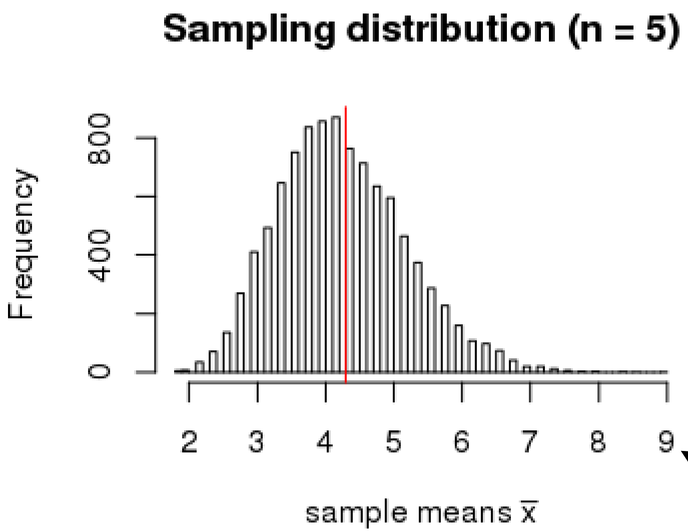
What symbol should
we use for this quantity?

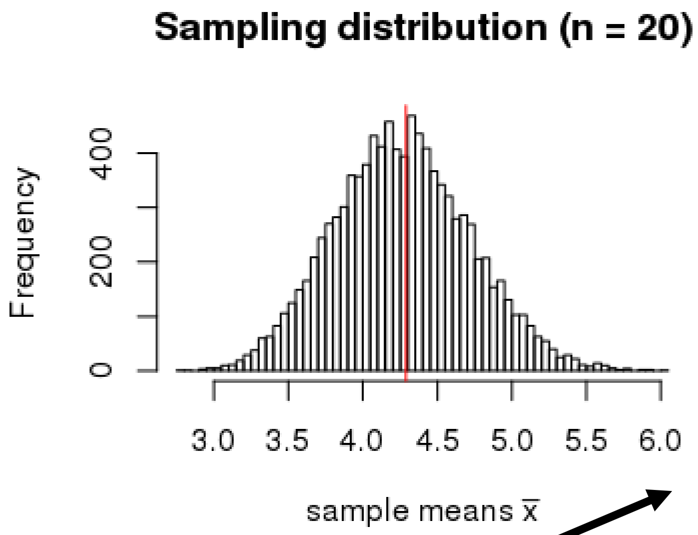
```
SE_mean <- sd(sampling_dist)
```

Q_7: 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



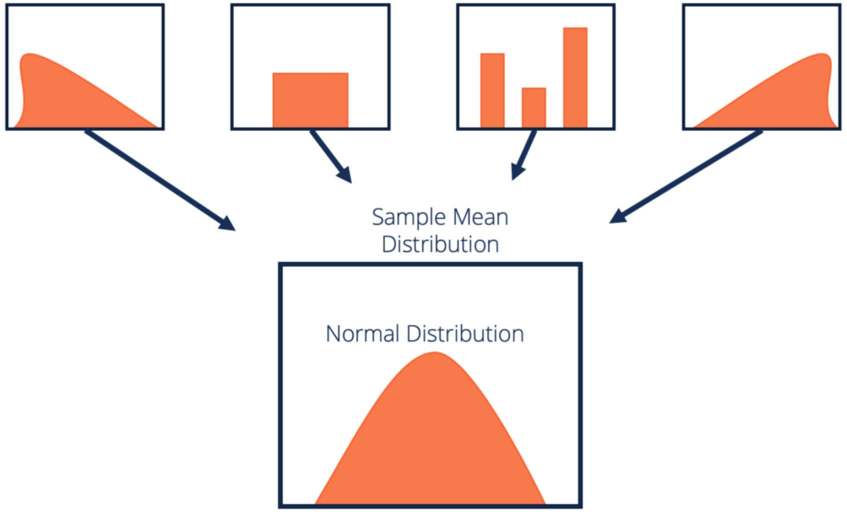


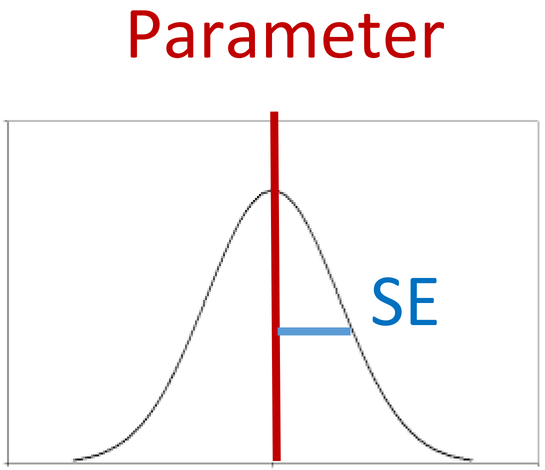
x-axis range 9 vs. 6

Central limit theorem

For random samples with a sufficiently large sample size (n), the distribution of sample statistics for a mean ($\bar{x}$) or a proportion ($\hat{p}$) is:

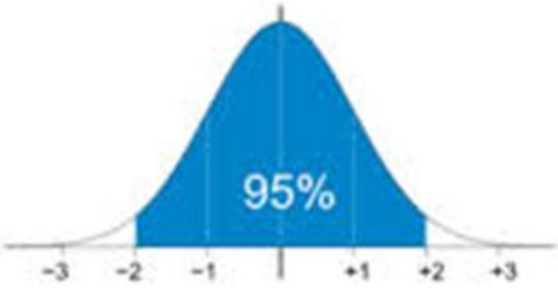
- Normally distributed
- Centered at the value of the population parameter

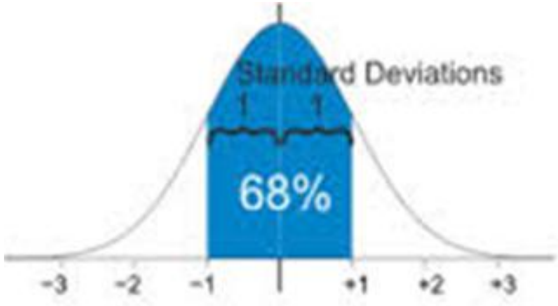


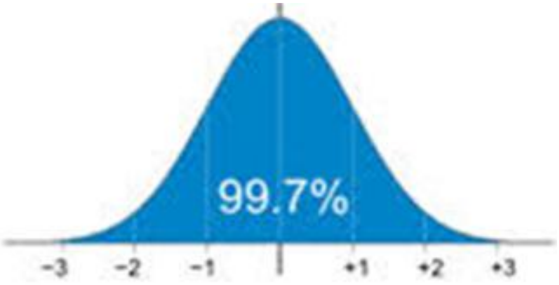


Normal distributions

Q_9: For a normal distribution, what percentage of points lie within 2 standard deviations for the population mean?



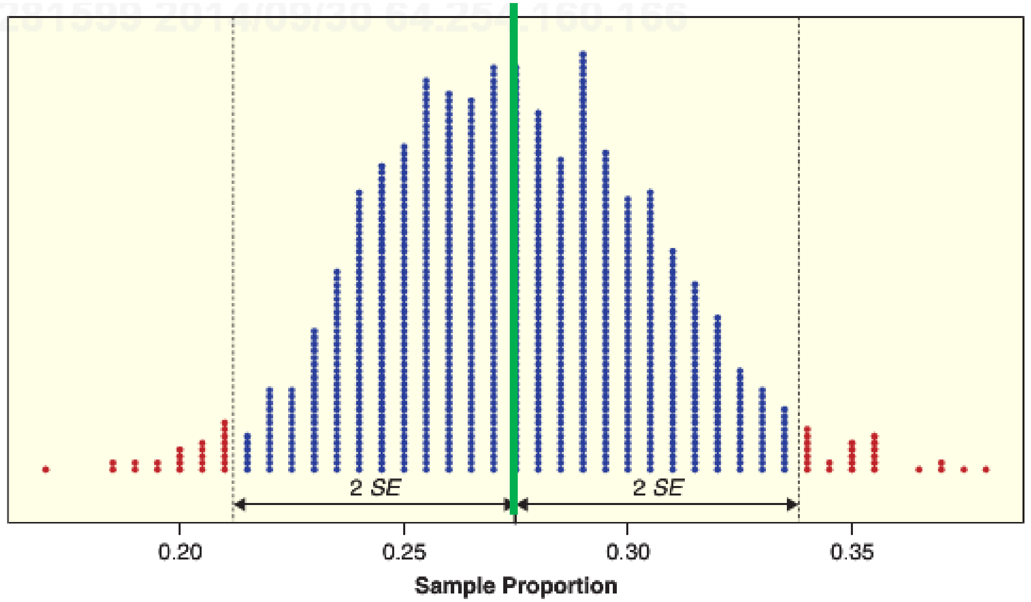




Sampling distributions

Q$_{10}$: For a **sampling distribution** that is a normal distribution, what percentage of ***statistics*** lie within 2 standard deviations (SE) of the parameter value?

Parameter



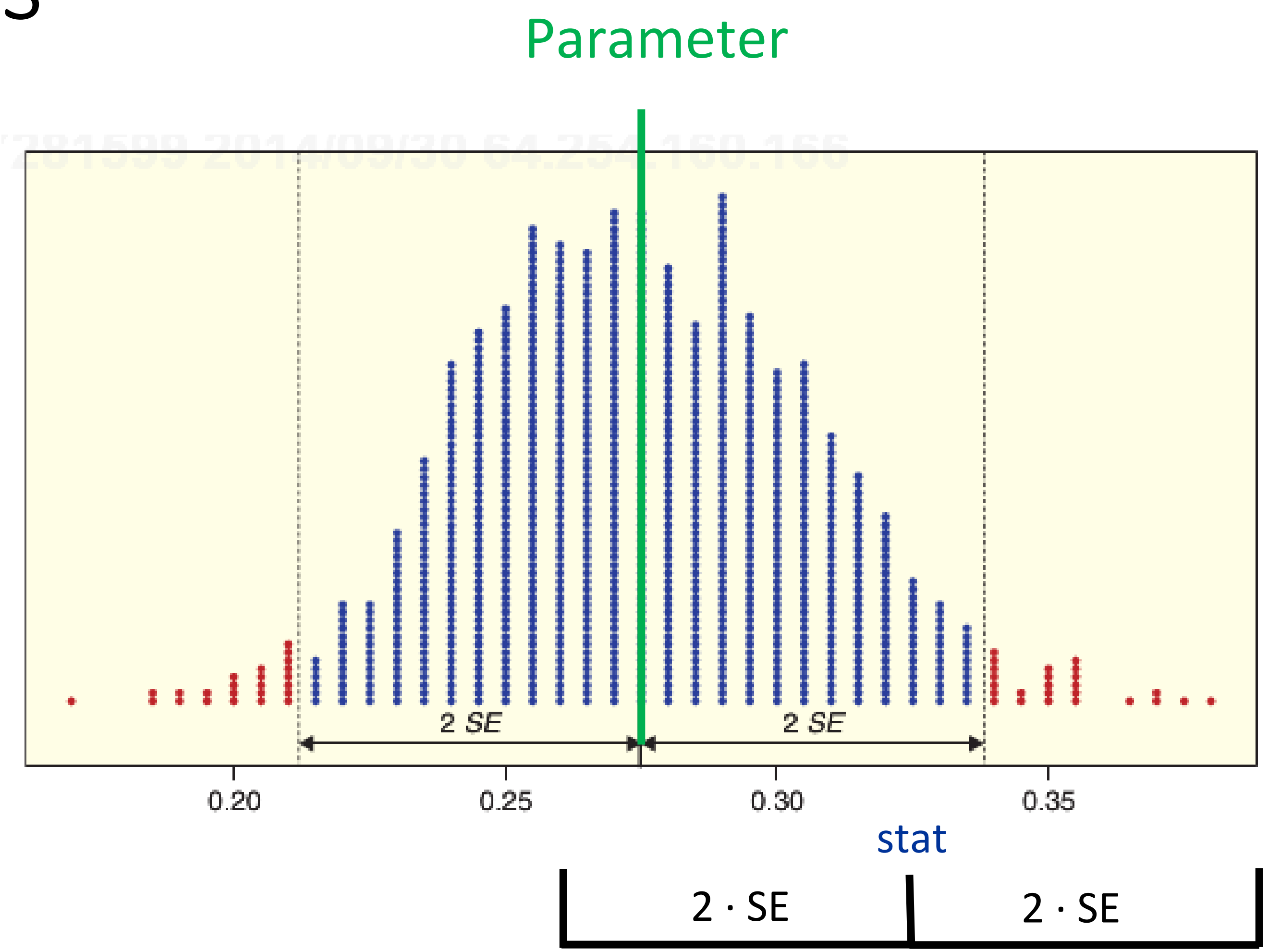
Q$_{11}$: If we had a *statistic value* and the value of the *SE*, could we compute a 95% confidence interval?

Confidence intervals

Q_{11}: Suppose we are going to randomly chosen statistic value

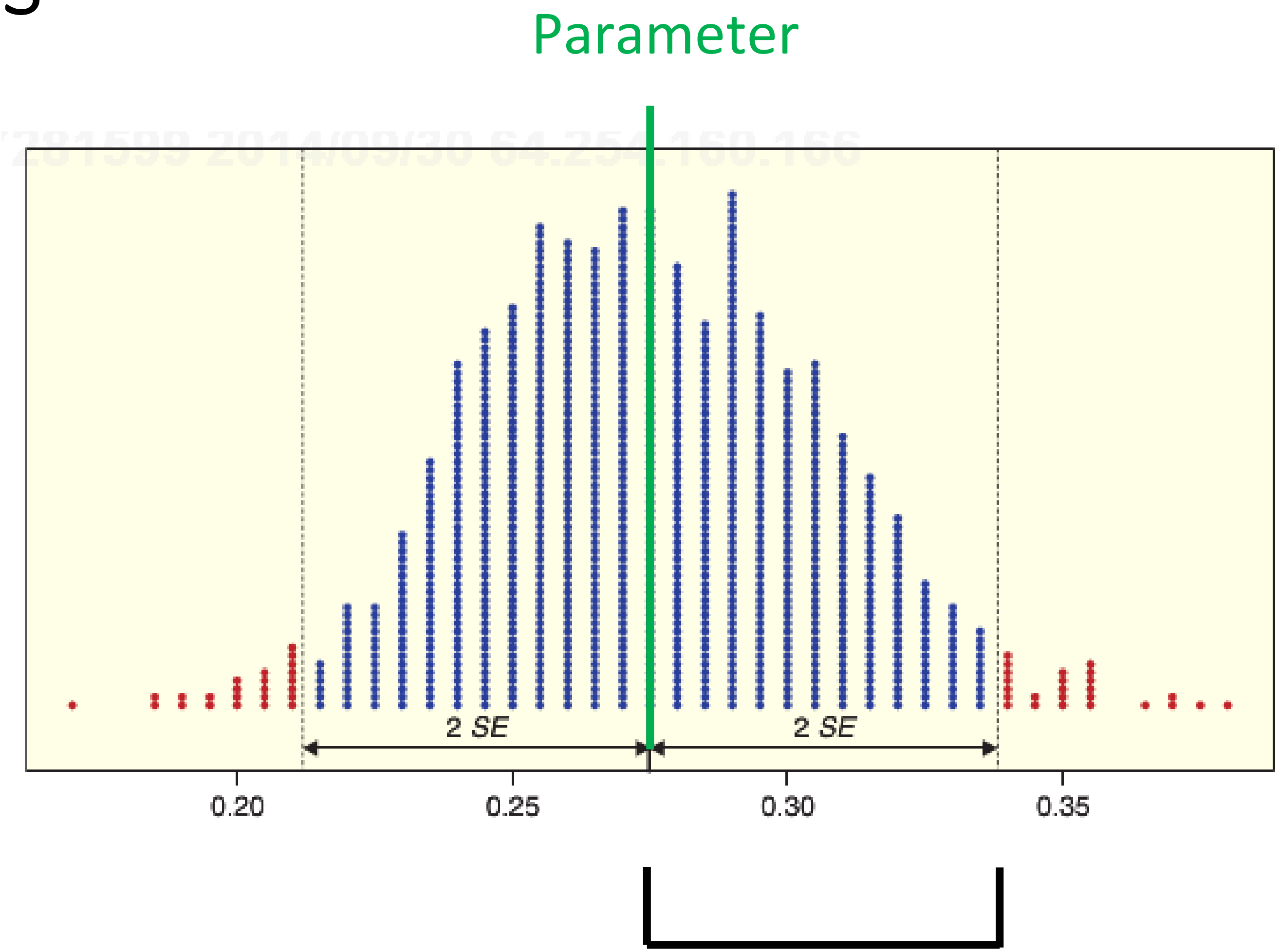
And we are going create an interval centered at the statistic value

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



Confidence interval

Confidence intervals



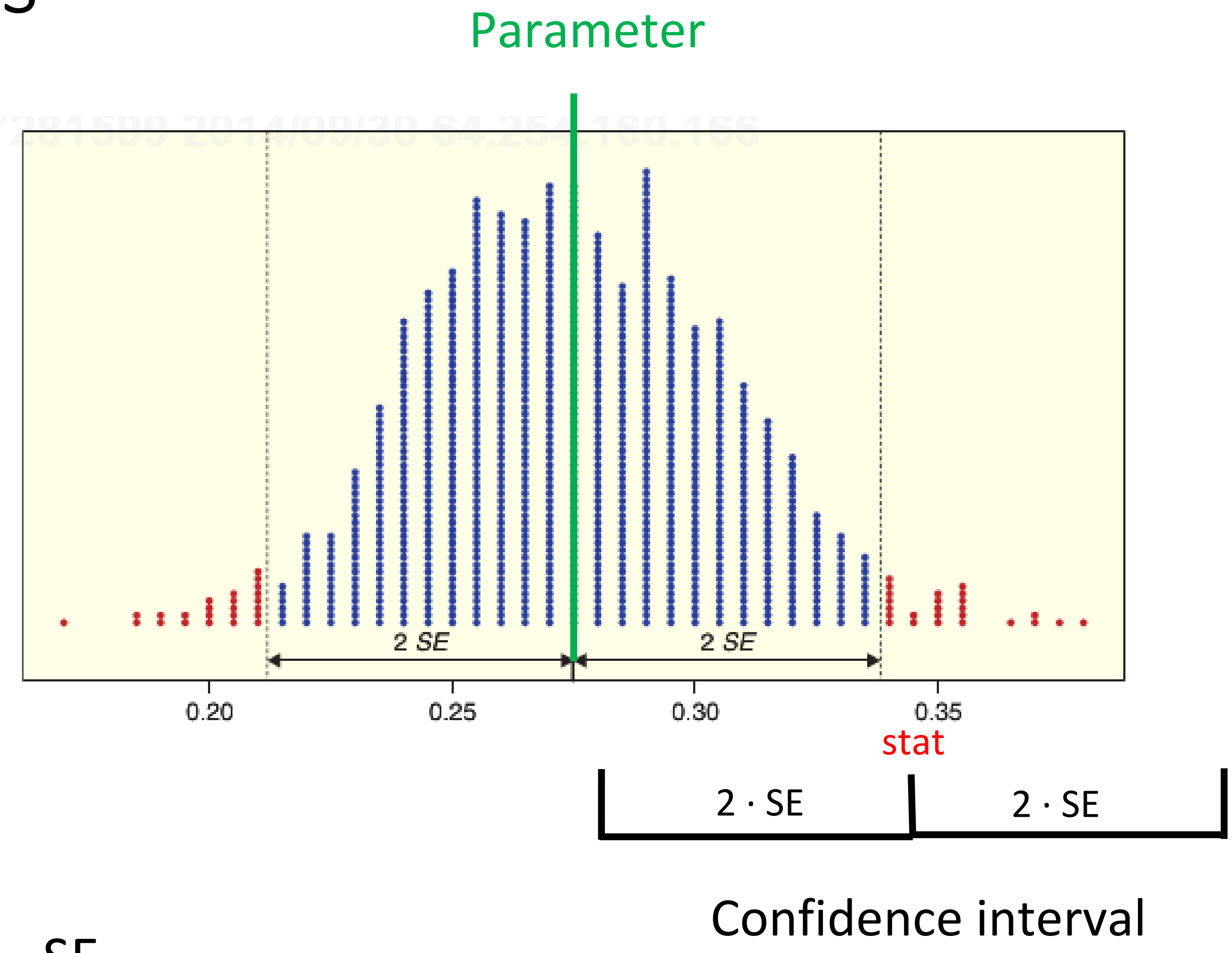
Q_{12}: What is a formula can we use to calculate 95% confidence intervals?

95% confidence interval: stat $\pm\ 2 \cdot SE$ → Q_{13}:What is this quantity called?

Confidence intervals

Q_{14}: How frequently do 95% confidence intervals **<u>fail</u>** to capture the parameter of interest?

- 5% of the time



Parameter

stat

2 · SE

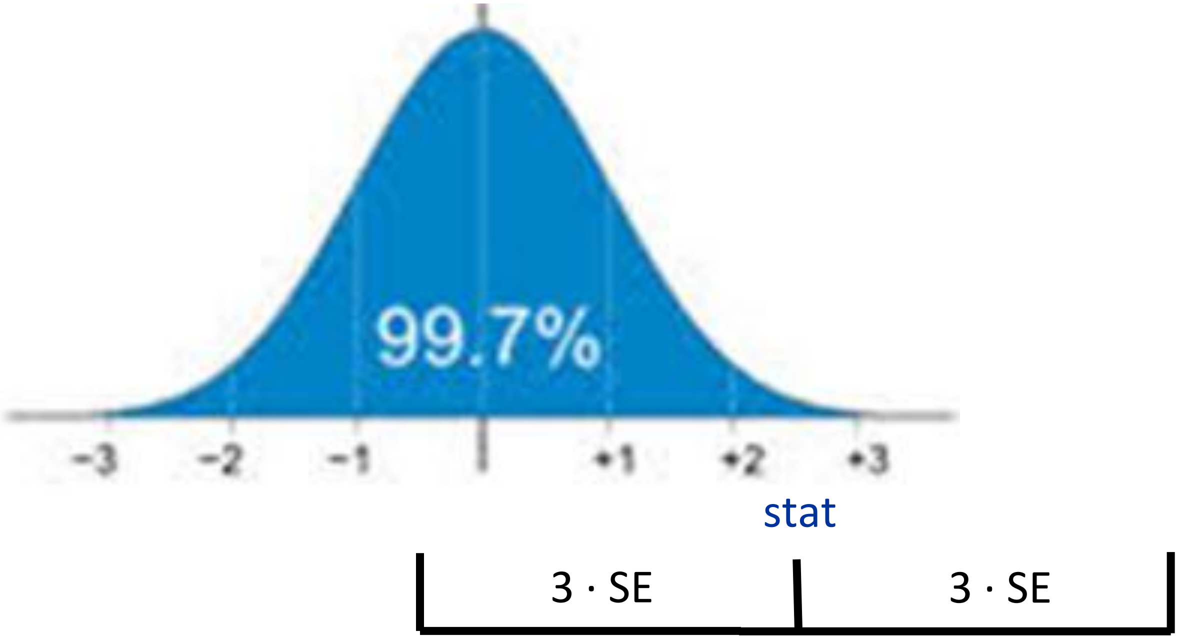
2 · SE

Confidence interval

95% confidence interval: stat ± 2 · SE

Confidence intervals for other confidence levels

Q$_{15}$: How could we get a 99.7% confidence interval confidence level?



stat

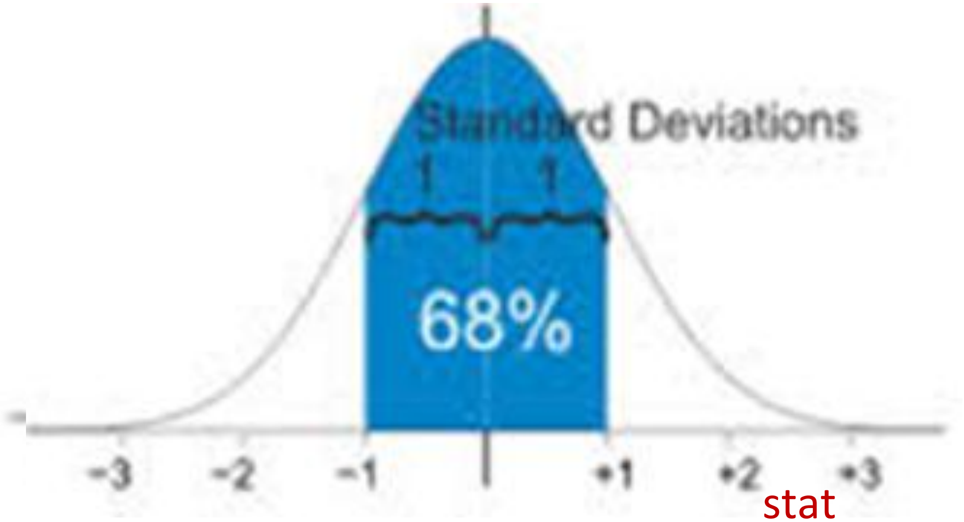
3 · SE

3 · SE

Confidence interval

Confidence intervals for other confidence levels

Q_{16}: How could we get a 68% confidence interval confidence level?



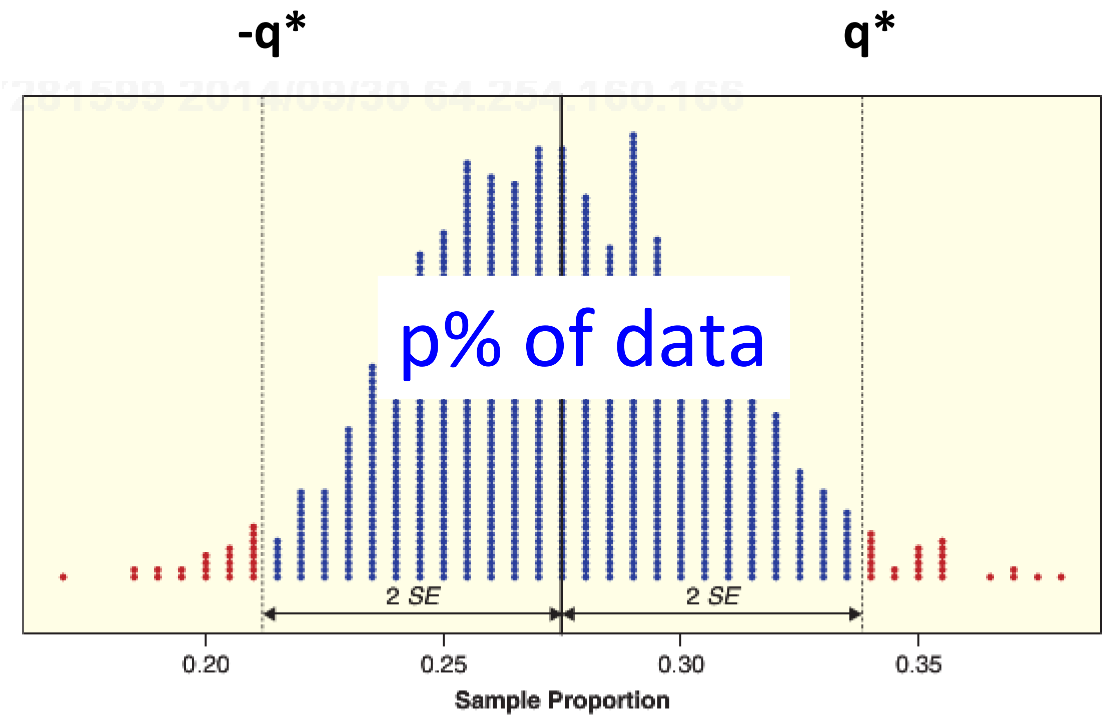
stat

$\underbrace{1 \cdot SE \mid 1 \cdot SE}$

Confidence interval

Confidence intervals for other confidence levels

Q_{16}: How could we get a confidence interval for the q^{th} confidence level?



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

```
library(SDS1000)
cnorm(0.95)
  [1]  1.96
```

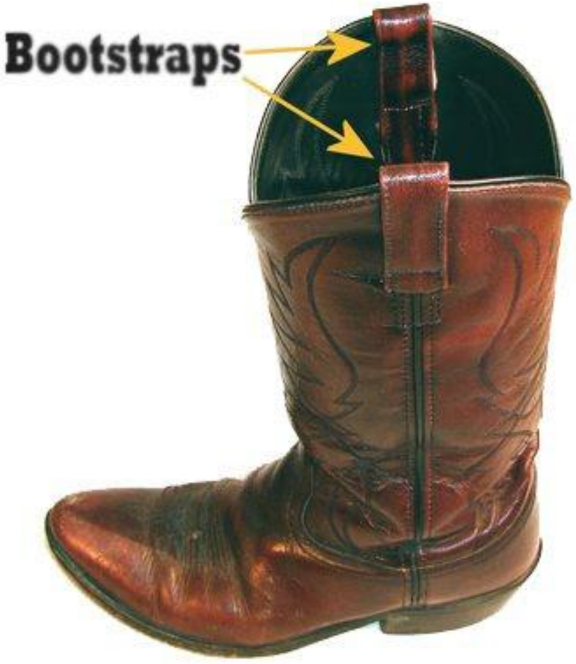
Sampling distributions

Q_{18}: Could we calculate the SE by repeatedly sampling from a population to create sampling distribution, and then take the sd of this sampling distribution?

Sampling distributions

Q_{19}: If we can't calculate the sampling distribution, what else can we do?

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

If you don’t feel confident about confidence intervals…

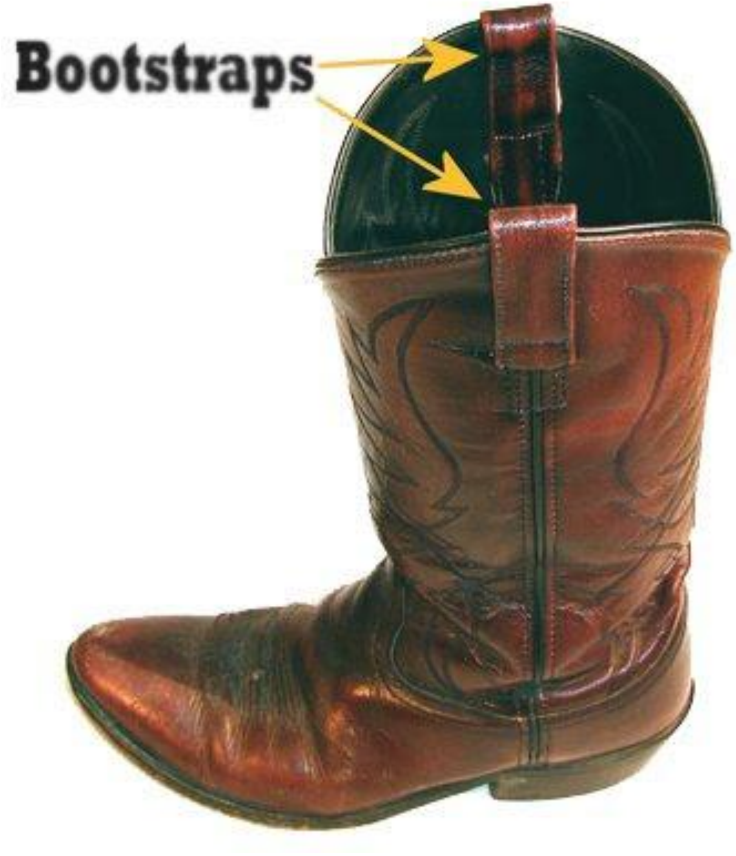
Go to the practice sessions

Review material

Ask questions on Ed Discussion

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

Plug-in principle

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

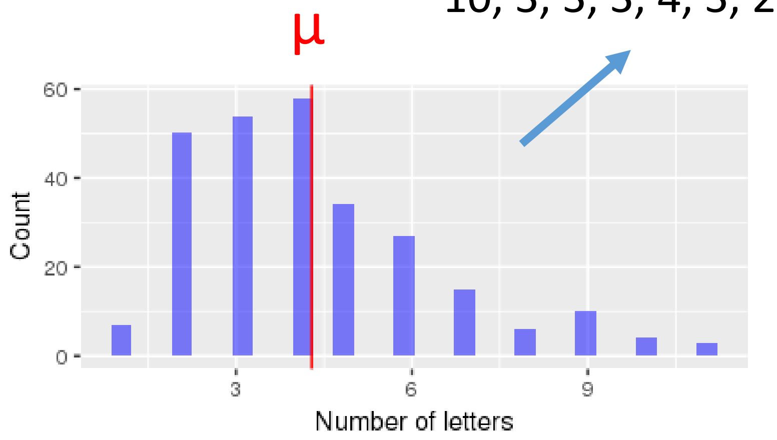
We pretend that this sample is the population (plug-in principle)

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



The sample (n = 10)

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

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

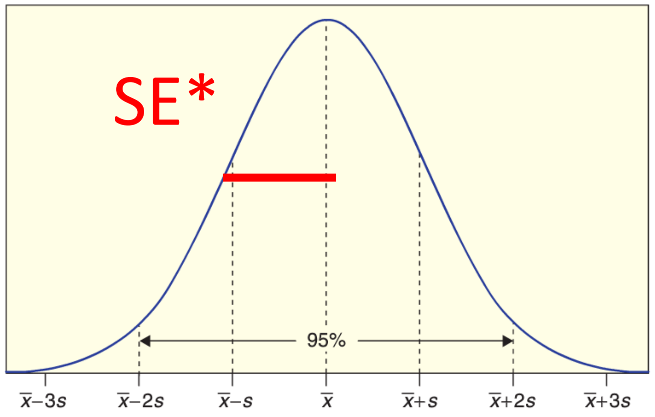
$\bar{x}^* = 4$

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

$\bar{x}^* = 4.1$

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

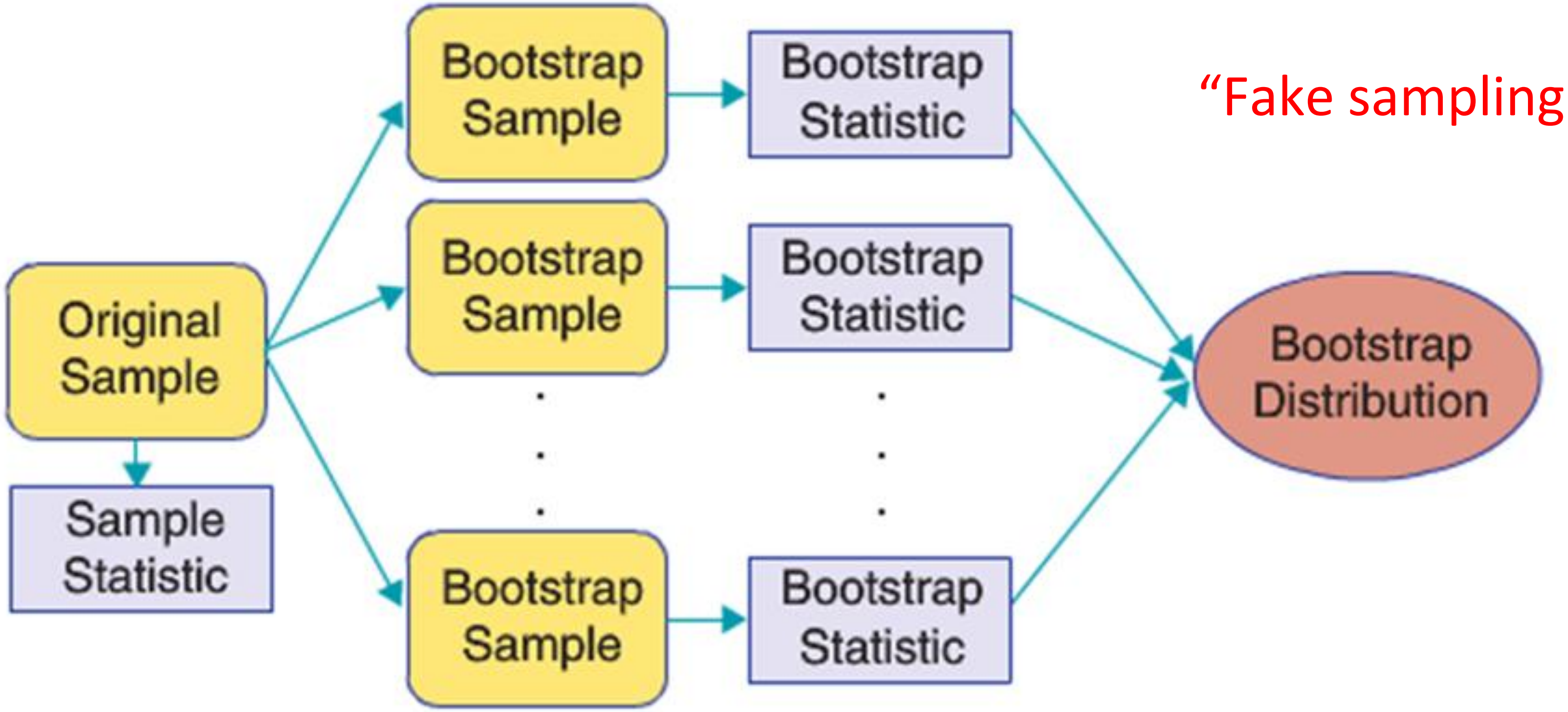
$\bar{x}^* = 3.9$



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

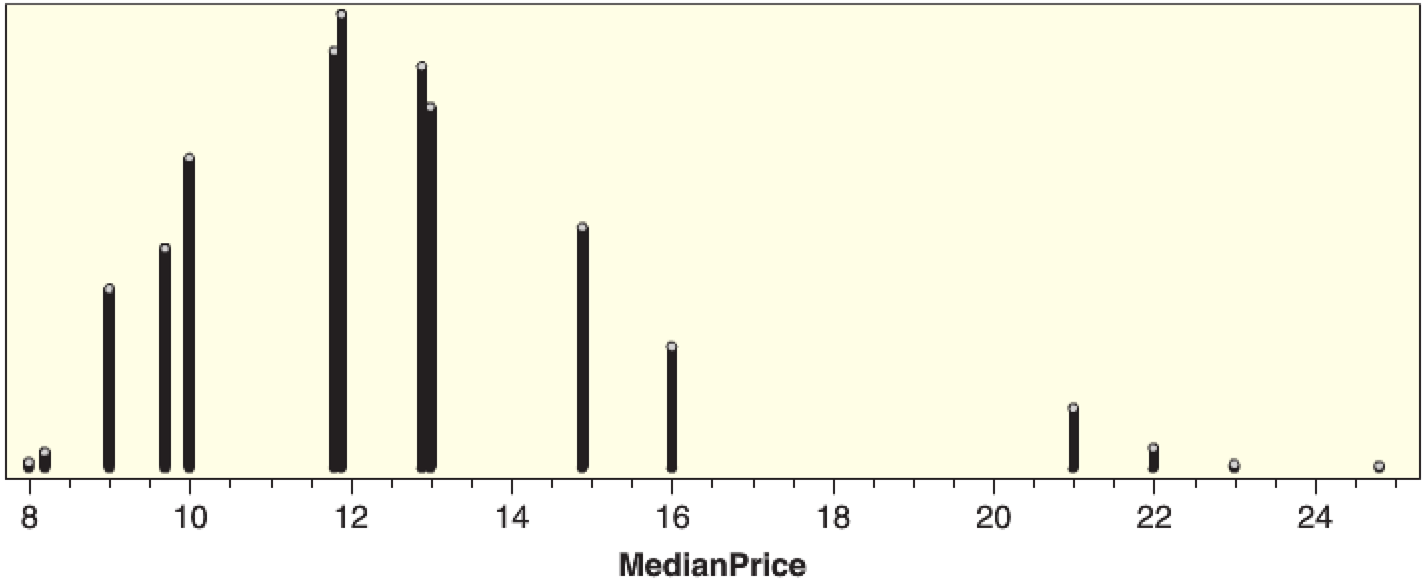
Findings CIs for many different parameters

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
```