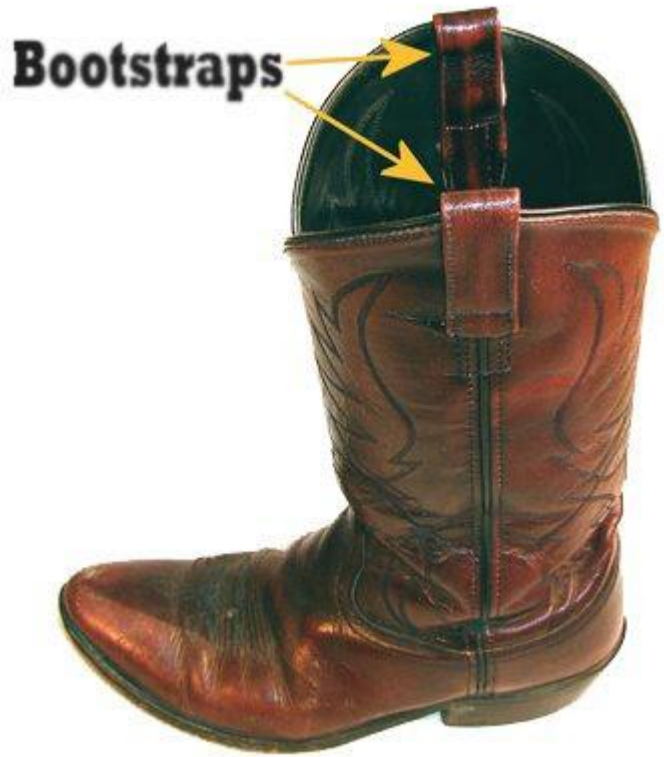
The bootstrap



Overview

Questions about worksheet 6?

Review: confidence intervals and sampling distributions

The bootstrap

Questions about worksheet 6?

How were the Lock5 questions?

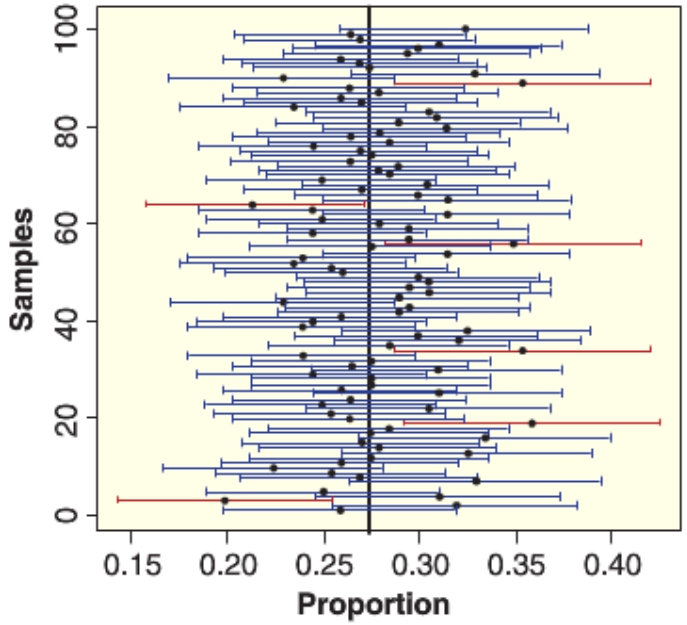
How was the sampling distribution question?

- We will review the code for this later in the class

Thank you everyone for turning in the worksheet on time!

- I realize it's not easy for everyone to turn in work when you are not sure of the answers, but it is important for me to see where everyone is at so that I can adjust the content as necessary

Confidence Intervals



Q_1: What is a **confidence interval**?

- A_1: a **confidence interval** is an interval computed by a method that will contain the *parameter* a specified percent of times

Q_2: What is the **confidence level**?

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

True or False? 95% confidence intervals

Q: A 95% confidence interval contains 95% of the data in the population?

- A: False

Q: The sample statistic (e.g., $\bar{x}$) will fall in the confidence interval 95% of the time?

- A: False - the sample statistic should always be in your CI

Q: For a particular confidence interval, there is a 95% probability that the population parameter (e.g., μ) is in the interval

- A: Actually false – once you have an interval, the parameter is either in it or is not in it.
- It's the method that constructs intervals that generates them in such a way that the parameter is in these intervals 95% of the time

Sampling distributions

Q_7: 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_8: What does a sampling distribution show us?

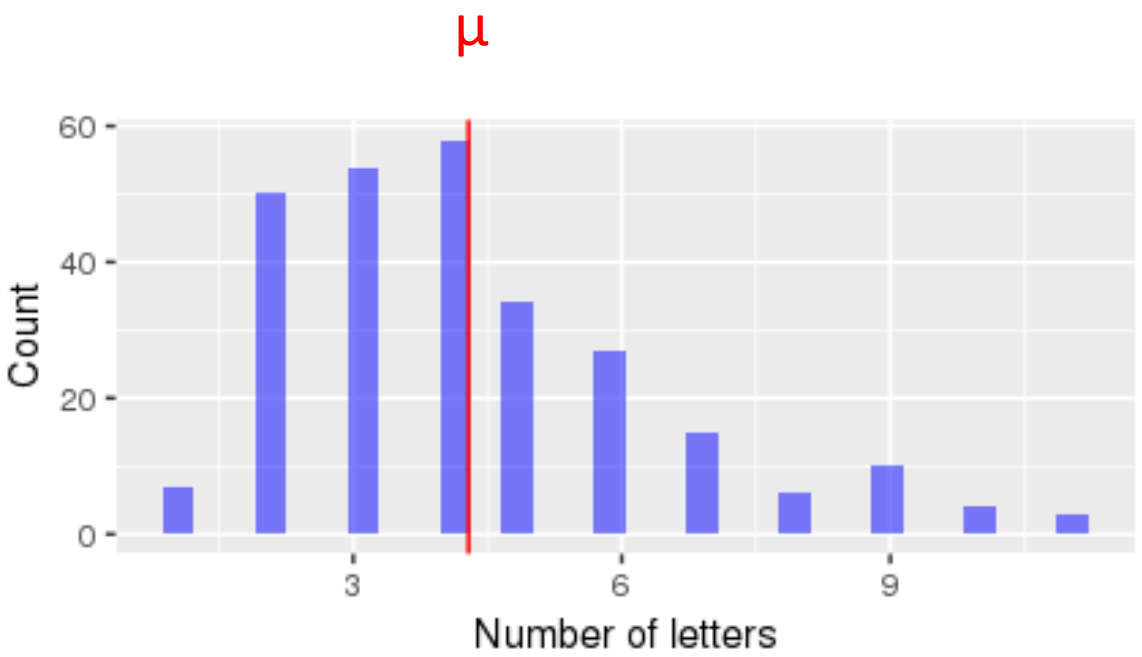
- A: A sampling distribution shows us how the sample statistic varies from sample to sample

Art time

Draw:

- Population
- 1 sample that has 100 points
- 9 more samples that have 100 points
- A sampling distribution
- Plato
- Population parameter with appropriate symbol
- Sample statistic with appropriate symbol

Gettysburg address word length sampling distribution



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

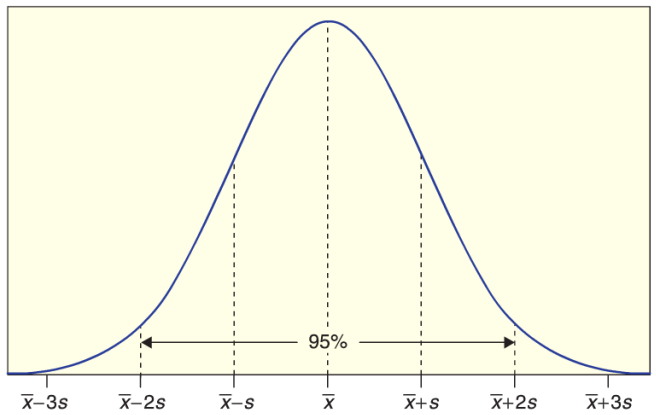
$\bar{x} = 5$

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]

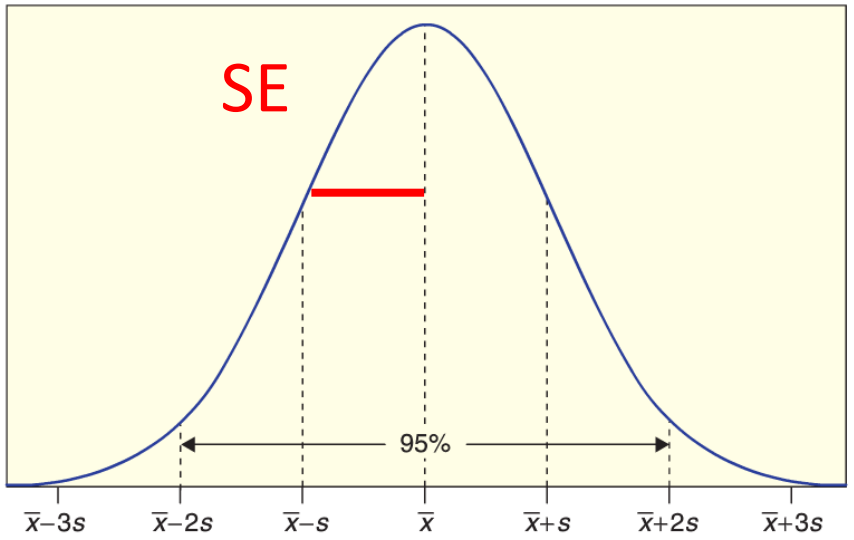
The standard error

Q_9: What is the **standard error (SE)**?

- The **standard error** of a statistic is the standard deviation of the sample statistic

Q_{12}: What does the size of the standard error tell us?

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



Shapes of sampling distributions

Q_{15a}: What is a commonly seen shape for sampling distributions?

A: Normal!

Let’s quickly examine whether sampling distributions really are usually normal

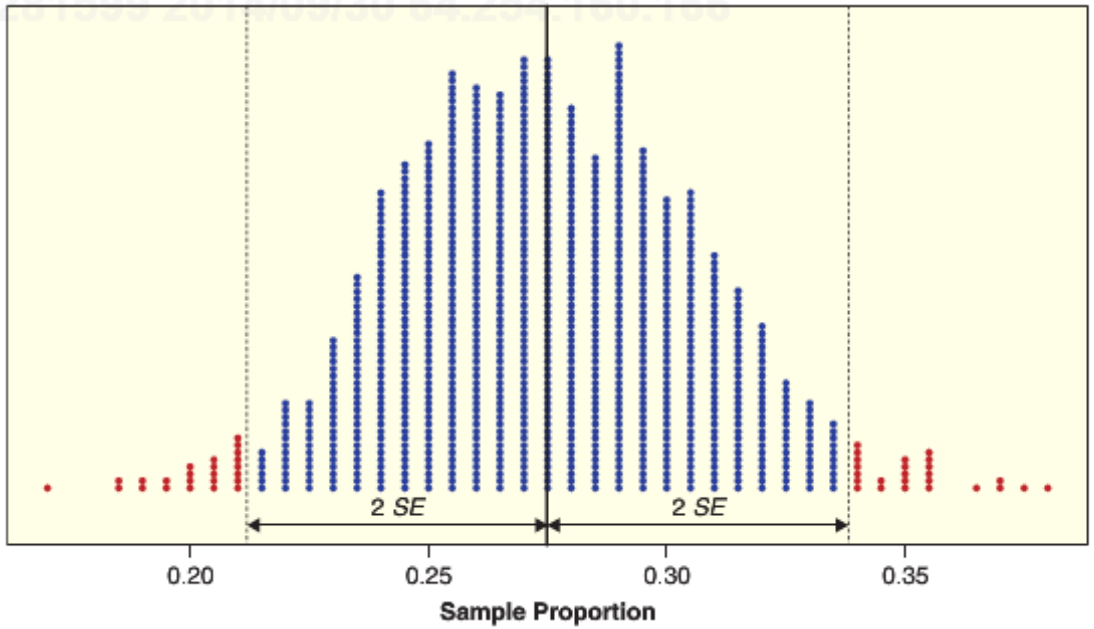
https://asterius.hampshire.edu:3939/intro_stats/sampling_and_bootstrap_distributions/

You will explore whether the sampling distribution appears normal using this app on worksheet 7!

Sampling distributions

Q_{16}: 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_{17}: If we had a statistics value and the value of the SE could we compute a 95% confidence interval?

A: Yes! CI = statistic_value $\pm$ 2 $\cdot$ SE

Worksheet 6: sprinkle sampling distribution

Getting (an approximate) sampling distribution for $\hat{p}_{red}$

```
source("/home/shared/intro_stats/get_sprinkle_sample.R")
```

Get $\hat{p}_{red}$ for one sample

```
my_sample <- get_sprinkle_sample(100)   # get the sample
sprinkle_table <- table(my_sample)      # count number of sprinkles
sprinkle_prop <- prop.table(sprinkle_table)  # get proportions
p_hat_stat <- sprinkle_prop[4]    # get red proportion (save it for later)
```

Worksheet 6: sprinkle sampling distribution

Getting (an approximate) sampling distribution for $\hat{p}_{red}$

```
red_prop_sampling_dist <- NULL
```

Get $\hat{p}_{red}$ for one sample

```
my_sample <- get_sprinkle_sample(100)   # get the sample
sprinkle_table <- table(my_sample)      # count number of sprinkles
sprinkle_prop <- prop.table(sprinkle_table)  # get proportions
p_hat_stat <- sprinkle_prop[4]
```

Worksheet 6: sprinkle sampling distribution

Getting (an approximate) sampling distribution for $\hat{p}_{red}$

```
red_prop_sampling_dist <- NULL

for (i in 1:10000) {
  my_sample <- get_sprinkle_sample(100)   # get the sample
  sprinkle_table <- table(my_sample)      # count number of sprinkles
  sprinkle_prop <- prop.table(sprinkle_table)  # get proportions
  p_hat_stat <- sprinkle_prop[4]
}
```

Worksheet 6: sprinkle sampling distribution

Getting (an approximate) sampling distribution for $\hat{p}_{red}$

```
   red_prop_sampling_dist <- NULL

for (i in 1:10000) {
   my_sample <- get_sprinkle_sample(100)   # get the sample
   sprinkle_table <- table(my_sample)      # count number of sprinkles
   sprinkle_prop <- prop.table(sprinkle_table)  # get proportions
   red_prop_sampling_dist[i] <- sprinkle_prop[4]
}
```

Worksheet 6: sprinkle sampling distribution

Plot the sampling distribution

```
hist(red_prop_sampling_dist, nclass = 100)
```

Calculate the standard error

```
SE_100 <- sd(red_prop_sampling_dist)
```

Create confidence interval

```
CI_100_lower <- p_hat_stat  - 2 * SE_100
CI_100_upper <- p_hat_stat  + 2 * SE_100
CI_100 <- c(CI_100_lower, CI_100_upper)
```

Sampling distributions

Q_{18}: Could we repeat the sampling process many times to create a sampling distribution and then calculate the SE?

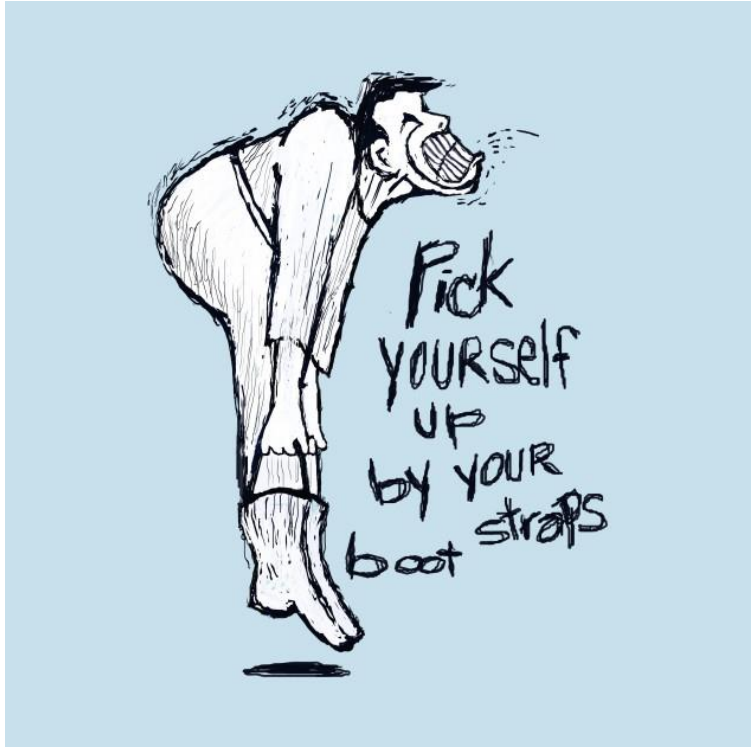
- A: Not in the real world because it would require running our experiment over and over again…

Sampling distributions

Q_{19}: 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*

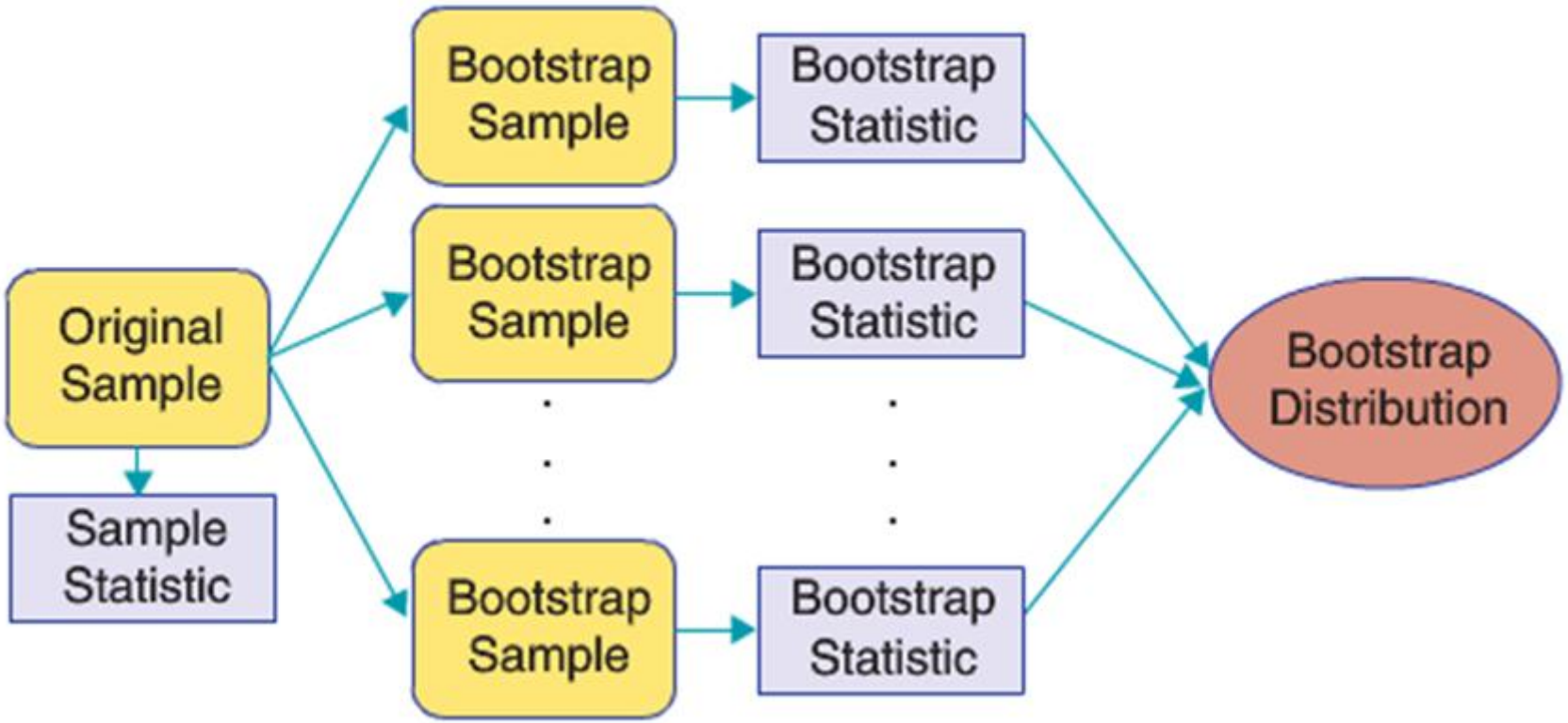
We pretend that this sample is the population (plug-in principle)

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

Bootstrap process



95% Confidence Intervals

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

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

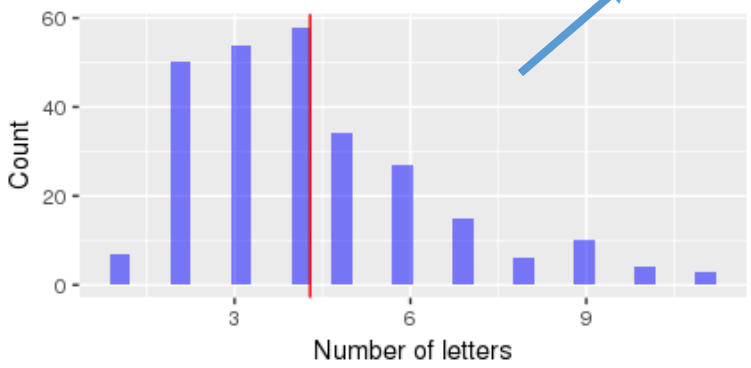
Where SE* is the standard error estimated using the bootstrap

Gettysburg address word length bootstrap distribution

The sample (n = 10)

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

μ



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

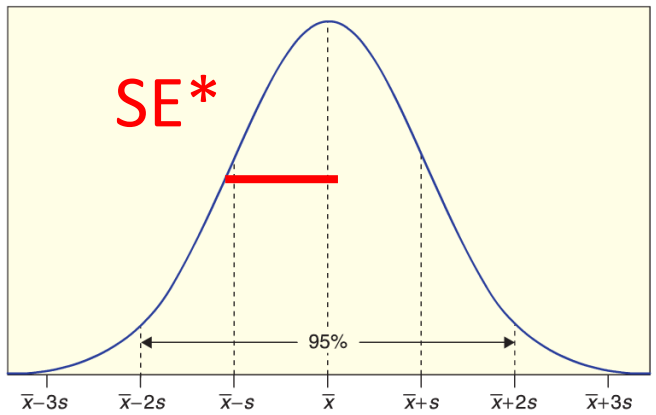
$\bar{x}^* = 4$

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

Let’s quickly explore bootstrap distributions

https://asterius.hampshire.edu:3939/intro_stats/sampling_and_bootstrap_distributions/

You will compare bootstrap distributions to population distributions and sampling distributions using this app on worksheet 7!

Resampling from a vector in R

```
my_sample <- c(3, 2, 4, 2, 5, 9, 10, 2, 4, 3)    # n = 10 points here
```

To get a sample of size 10 with replacement:

```
> boot_sample  <-  sample(my_sample,  10,  replace = TRUE)
```

To create a bootstrap SE, what else do we need to do apart from drawing a sample with replacement?

What are the steps needed to create a bootstrap SE?

1. Start with a sample for size *n*

2. Repeat steps 10,000 times

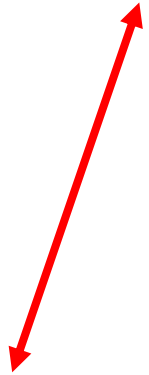
    a. Resample with replacement *n* points in ***the original sample*** to get a *bootstrap sample*

    b. Compute the statistic of interest on the bootstrap sample

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

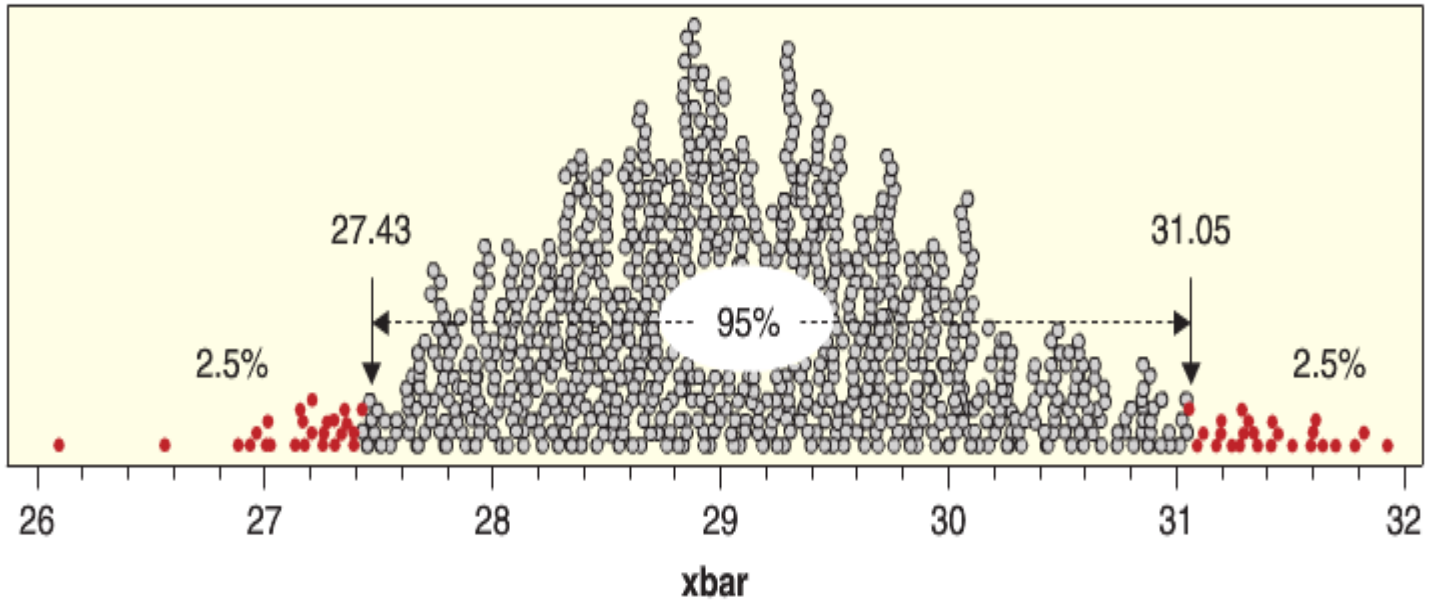
Bootstrap distribution in R for μ

```
my_sample <- c(3, 2, 4, 2, 5, 9, 10, 2, 4, 3)    # n = 10 points here
my_stat <- mean(my_sample)

bootstrap_dist <- NULL

for (i in 1:100000) {

        curr_boot <- sample(my_sample , 10, replace = TRUE)
        bootstrap_dist[i] <-  mean(curr_boot)

}

SE_boot <- sd(bootstrap_dist)
CI <- c(my_stat - 2  * SE_boot, my_stat  + 2  * SE_boot)
```



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.



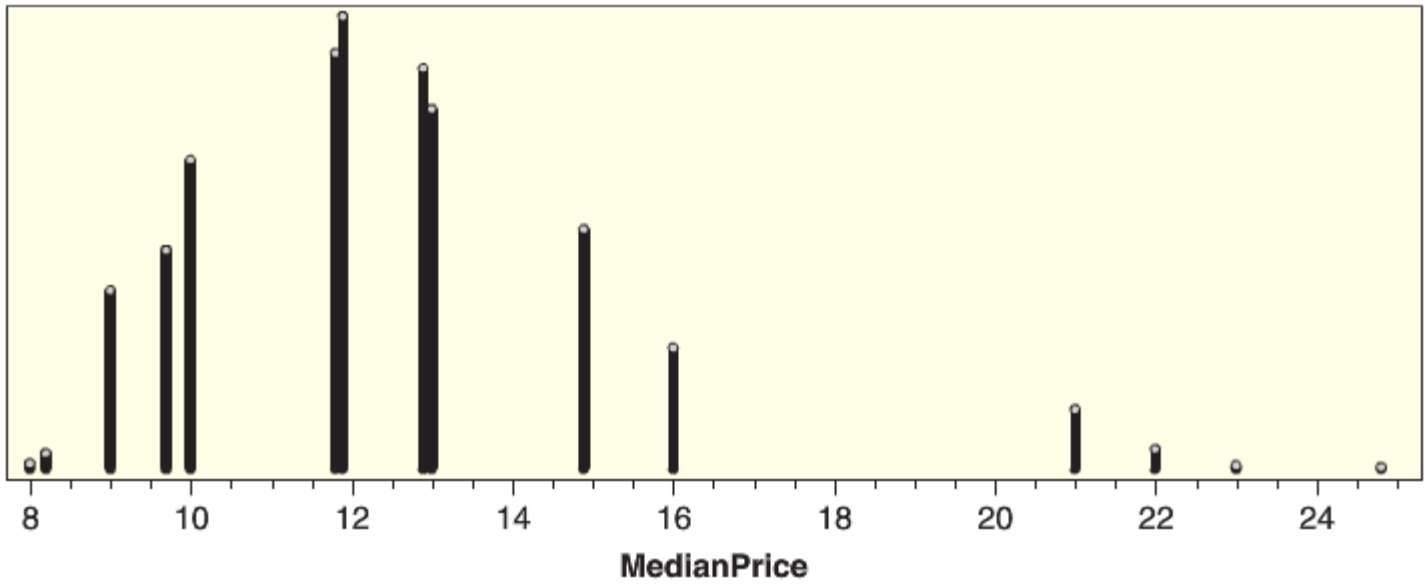
Findings CIs for many different parameters

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.

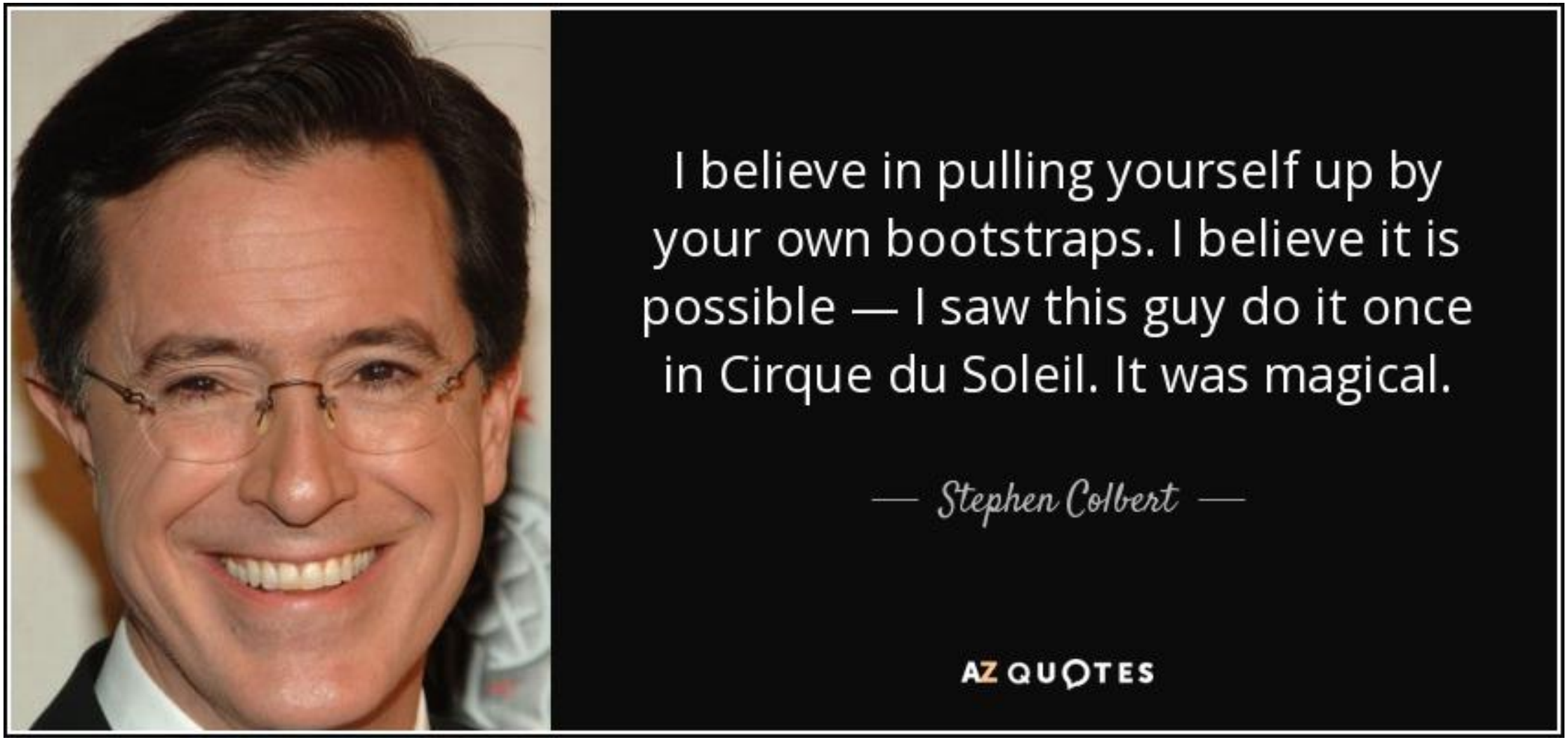
Median prices of Mustangs



Worksheet 7

Due 11:59 on Sunday Oct 28st

```
> source('/home/shared/intro_stats/cs206_functions.R')
> get_worksheet(7)
```


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