

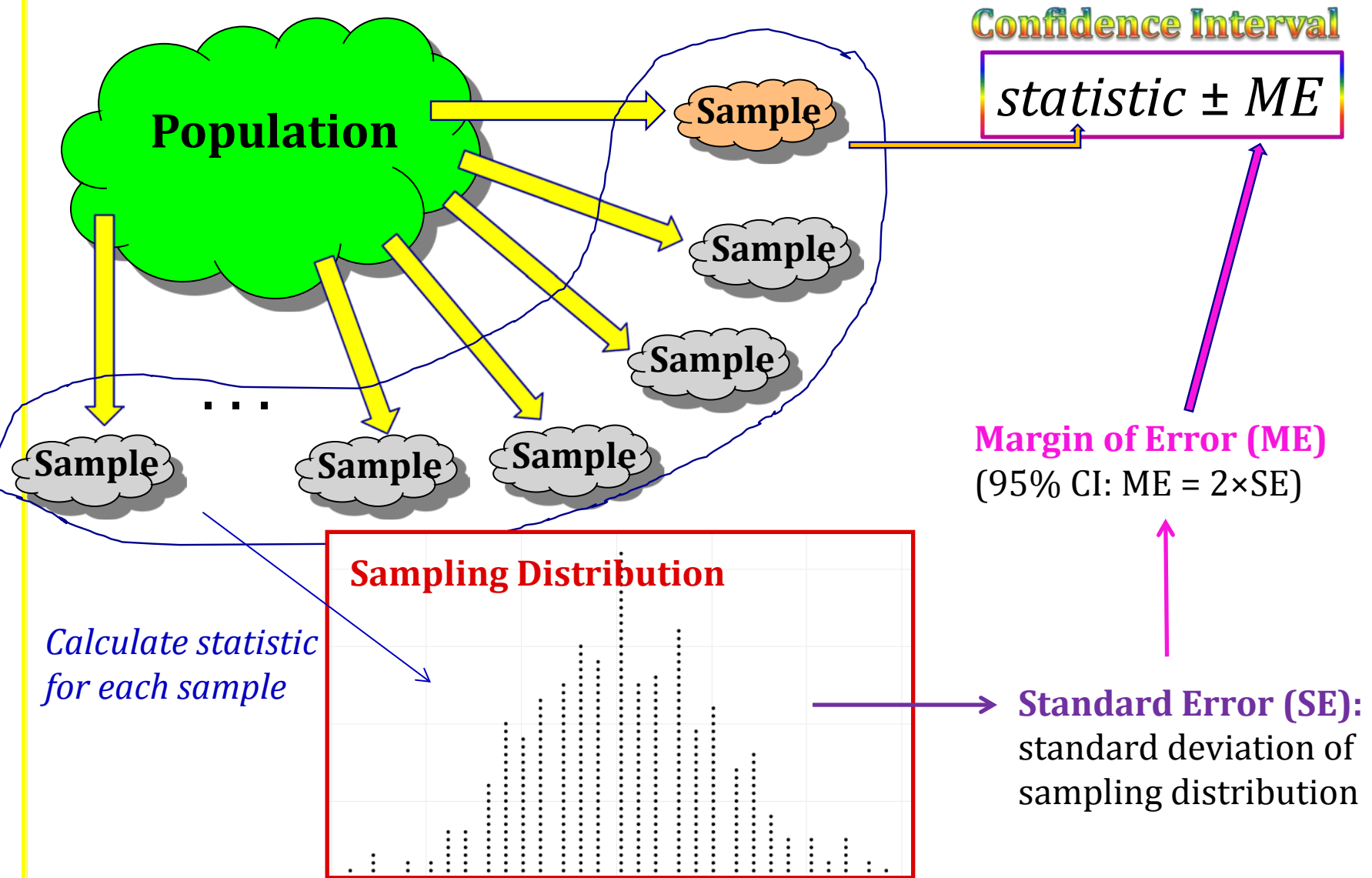
Section 3.3

Constructing Bootstrap Confidence Intervals

Outline

- Bootstrap samples
- Bootstrap distribution
- Standard error from bootstrap distribution
- 95% confidence interval using SE from bootstrap distribution

Confidence Intervals



Summary

- To create a plausible range of values for a parameter:
 - Take many random samples from the population, and compute the sample statistic for each sample
 - Compute the standard error as the standard deviation of all these statistics
 - Use statistic $\pm 2 \times \text{SE}$

- One small problem...

Reality

... WE ONLY HAVE ONE SAMPLE!!!!

- How do we know how much sample statistics vary, if we only have one sample?!?

BOOTSTRAP!

ONE Reese's Pieces Sample

Sample: 52/100 orange

$$\hat{p} = 0.52$$


Where might the “true” p be?

“Population”

- Imagine the “population” is many, many copies of the original sample
- (What do you have to assume?)

Reese's Pieces "Population"

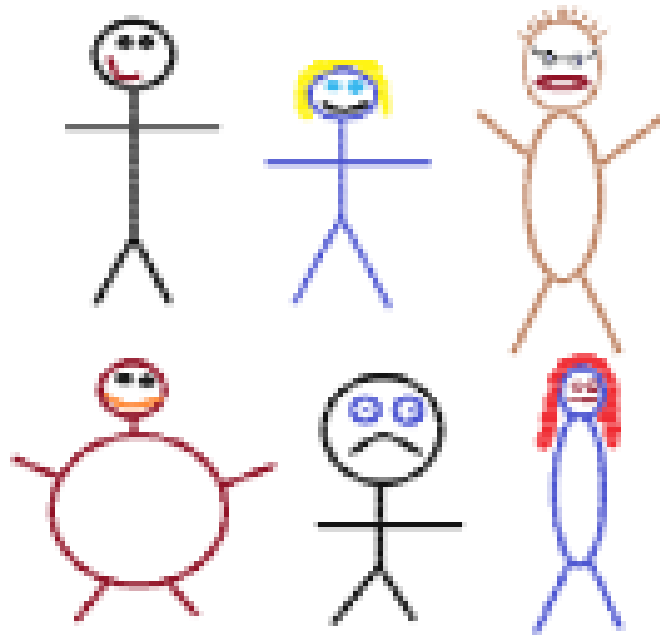
Sample repeatedly from
this "population"

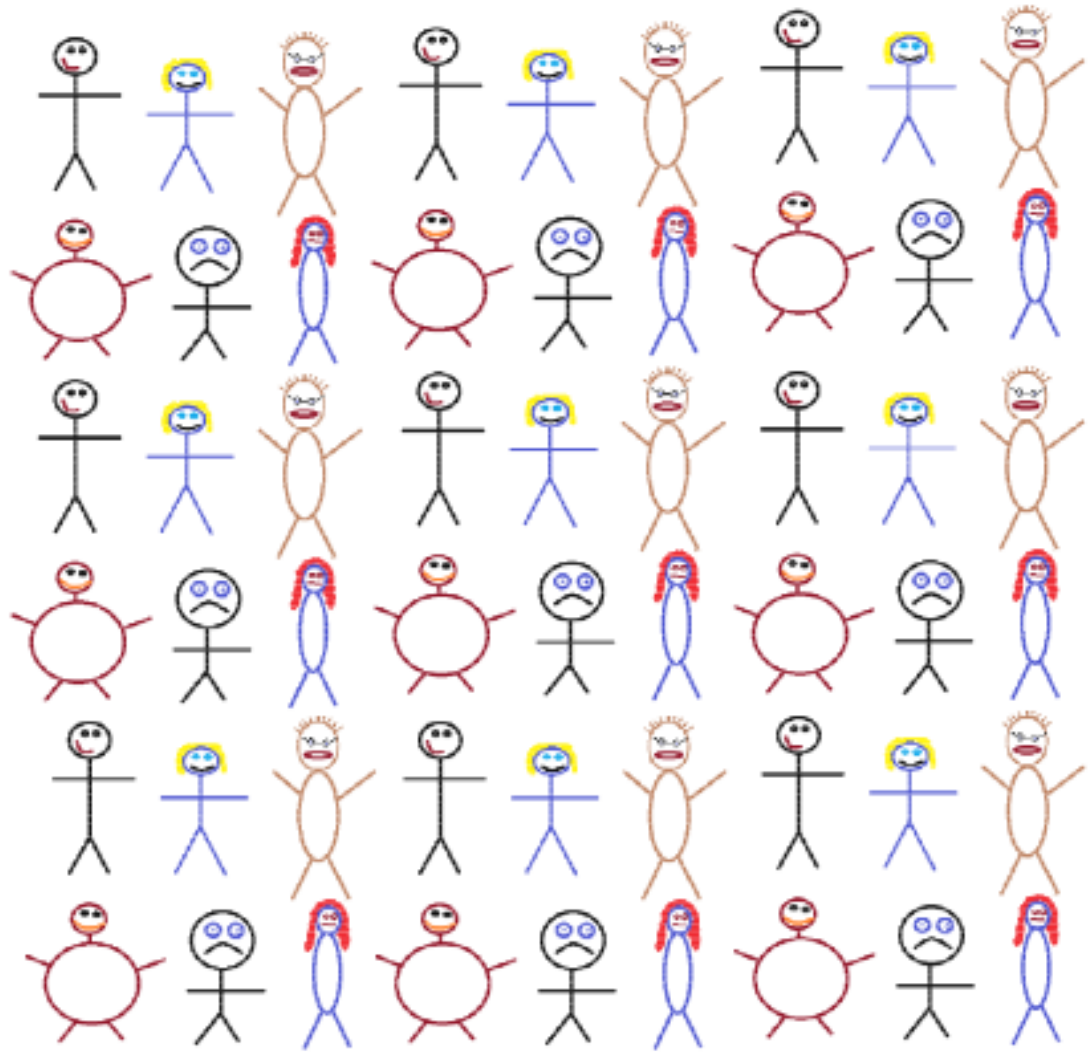
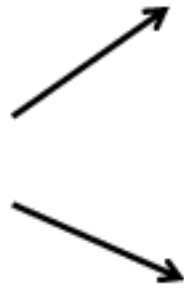
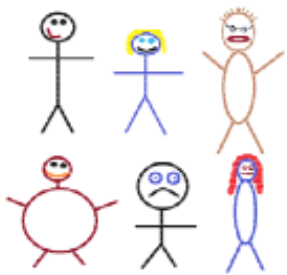


Sampling with Replacement

- To simulate a sampling distribution, we can just take repeated random samples from this “population” made up of many copies of the sample
- In practice, we can't actually make infinite copies of the sample...
- ... but we can do this by sampling *with replacement* from the sample we have (each unit can be selected more than once)

Suppose we have a random sample
of 6 people:

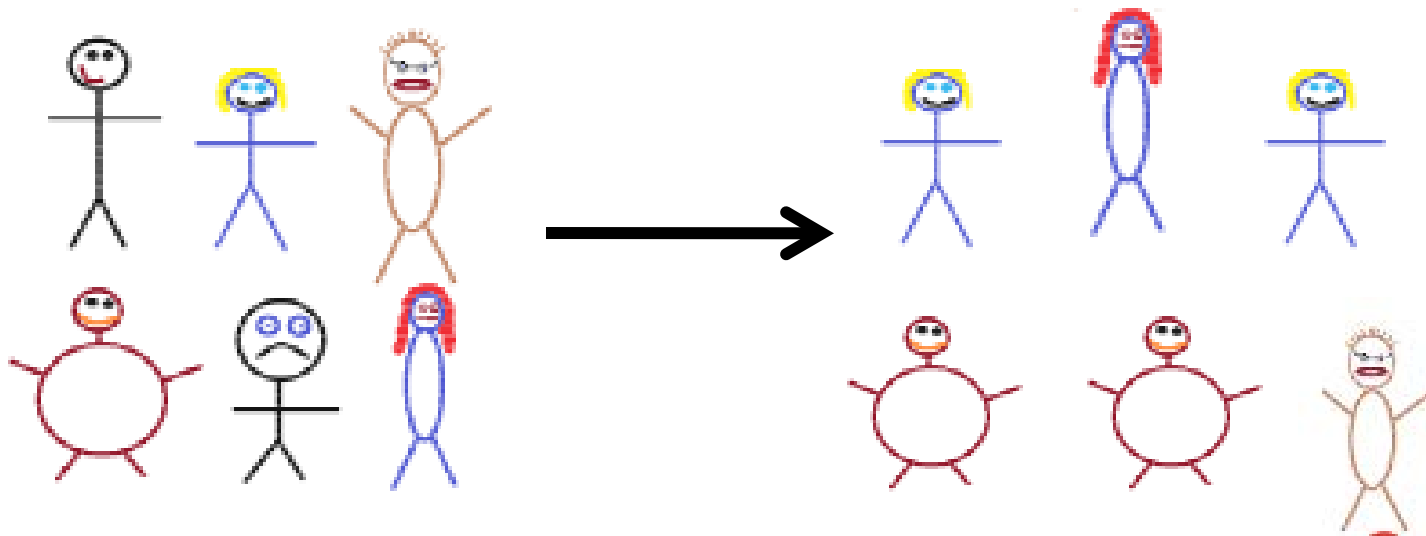




Original
Sample

A simulated “population” to sample from

Bootstrap Sample: Sample with replacement from the original sample, using the same sample size.



Original
Sample

Bootstrap Sample

Reese's Pieces

- How would you take a bootstrap sample from your sample of Reese's Pieces?



Bootstrap Sample

Your original sample has data values

18, 19, 19, 20, 21

Is the following a possible bootstrap sample?

18, 19, 20, 21, 22

*NO. 22 is not a value from
the original sample*

Bootstrap Sample

Your original sample has data values

18, 19, 19, 20, 21

Is the following a possible bootstrap sample?

18, 19, 20, 21

NO. Bootstrap samples must be the same size as the original sample

Bootstrap Sample

Your original sample has data values

18, 19, 19, 20, 21

Is the following a possible bootstrap sample?

18, 18, 19, 20, 21

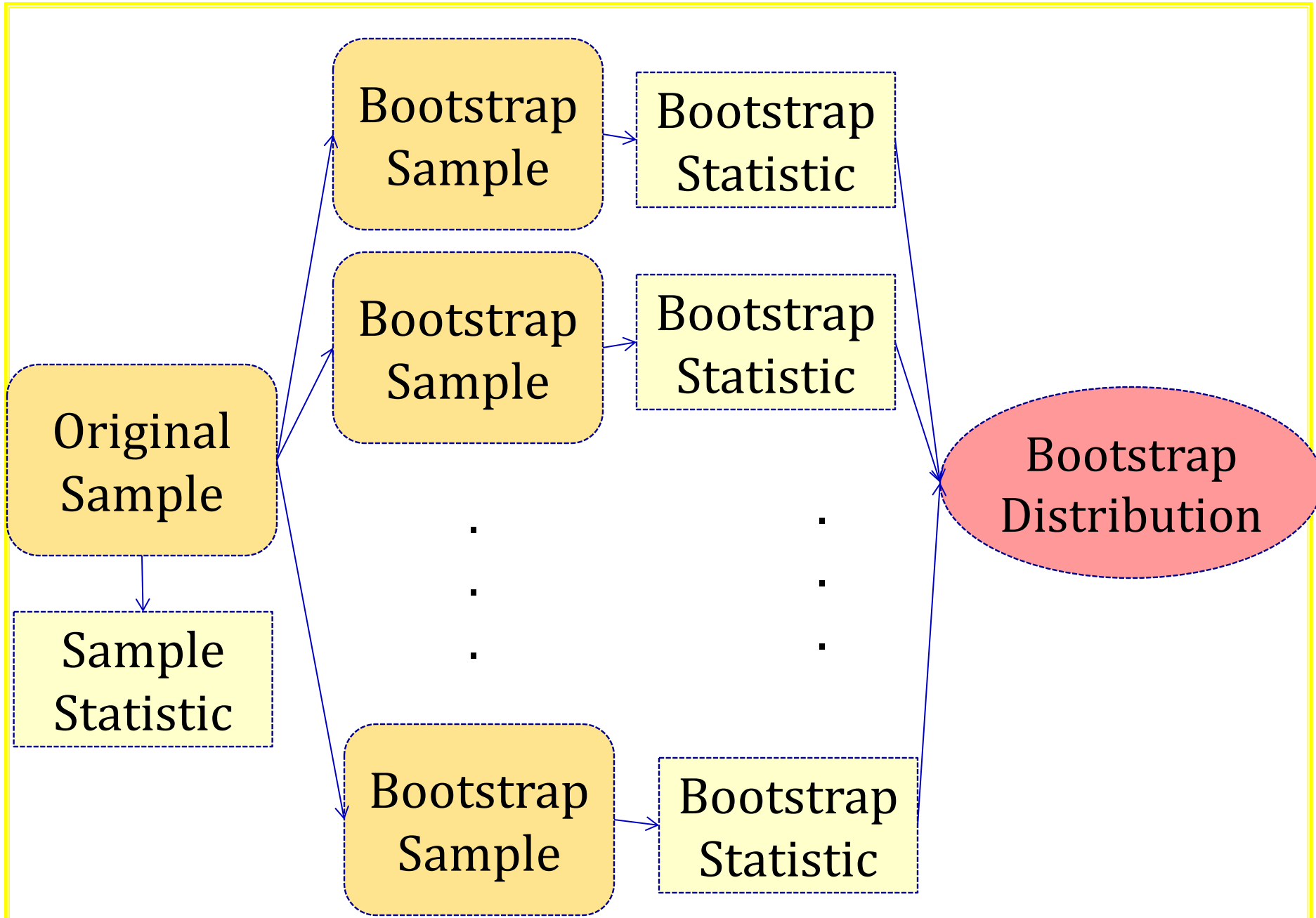
*YES. Same size, could be gotten
by sampling with replacement*

Bootstrap

A ***bootstrap sample*** is a random sample taken with replacement from the original sample, of the same size as the original sample

A ***bootstrap statistic*** is the statistic computed on a bootstrap sample

A ***bootstrap distribution*** is the distribution of many bootstrap statistics



Bootstrap Distribution

lock5stat.com/statkey/

Bootstrap For One Categorical Variable [\[Return to StatKey Index\]](#)

Custom Data ▾

Edit Data

Generate 1 Sample

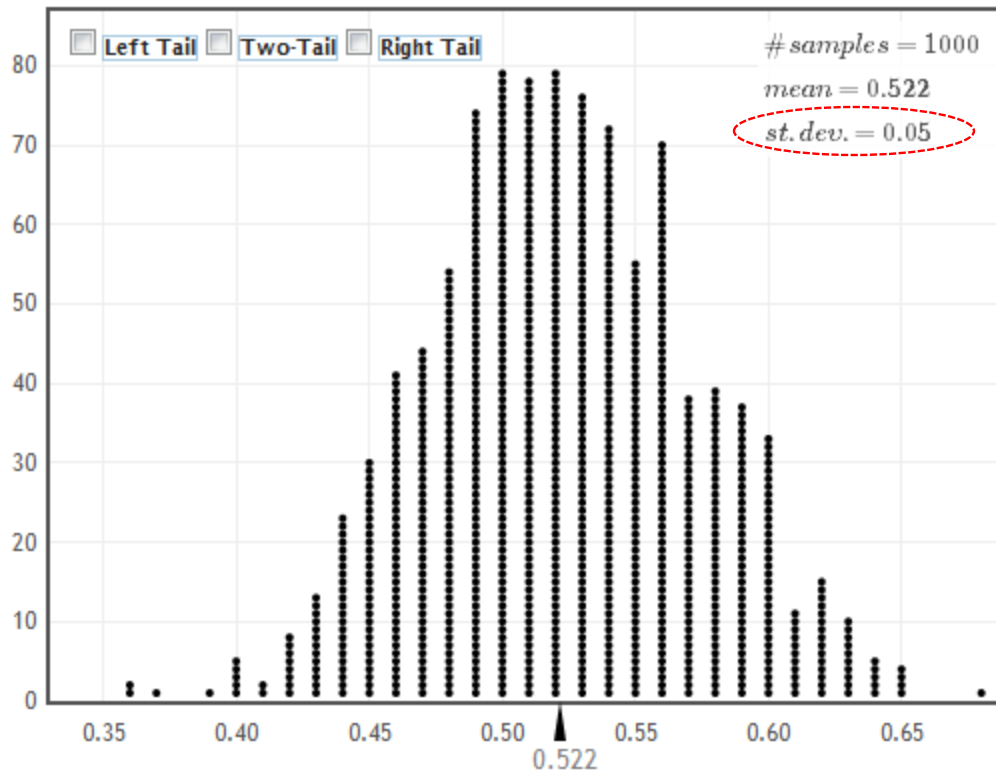
Generate 10 Samples

Generate 100 Samples

Generate 1000 Samples

Reset Plot

Bootstrap Dotplot of **Proportion** ▾



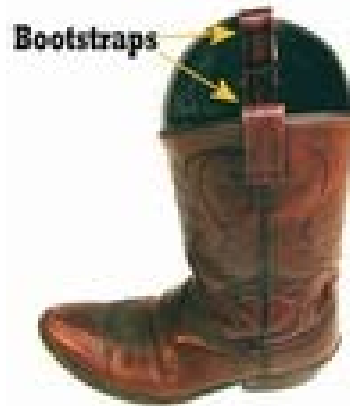
Original Sample

Count	n	Proportion
52	100	0.52

Bootstrap Sample

Count	n	Proportion
46	100	0.46

Why “bootstrap”?

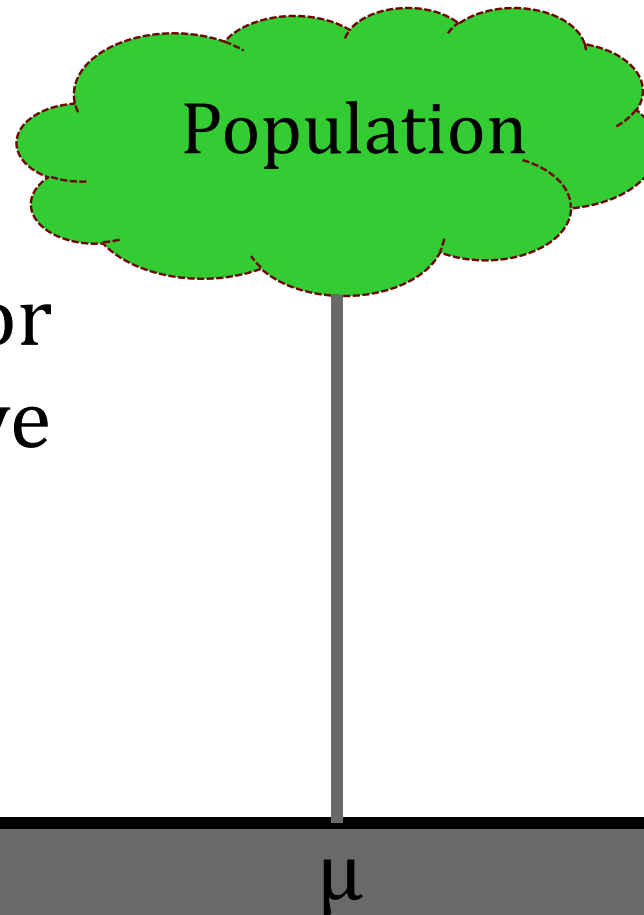


“Pull yourself up by your bootstraps”

- Lift yourself in the air simply by pulling up on the laces of your boots
- Metaphor for accomplishing an “impossible” task without any outside help

Sampling Distribution

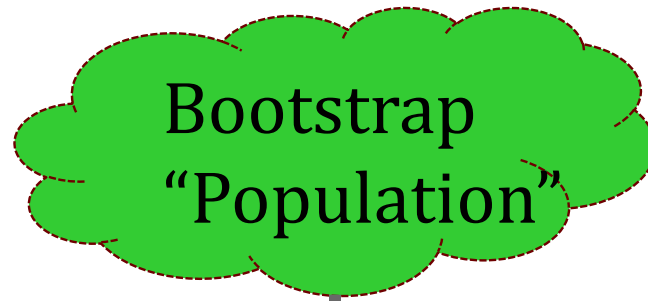
BUT, in practice we don't see the "tree" or all of the "seeds" – we only have ONE seed



Bootstrap Distribution

What can we do with just one seed?

Grow a NEW tree!



Estimate the distribution and variability (SE) of \bar{x} 's from the bootstraps



Golden Rule of Bootstrapping

Bootstrap statistics are to the
original sample statistic

as

the original sample statistic is to
the population parameter

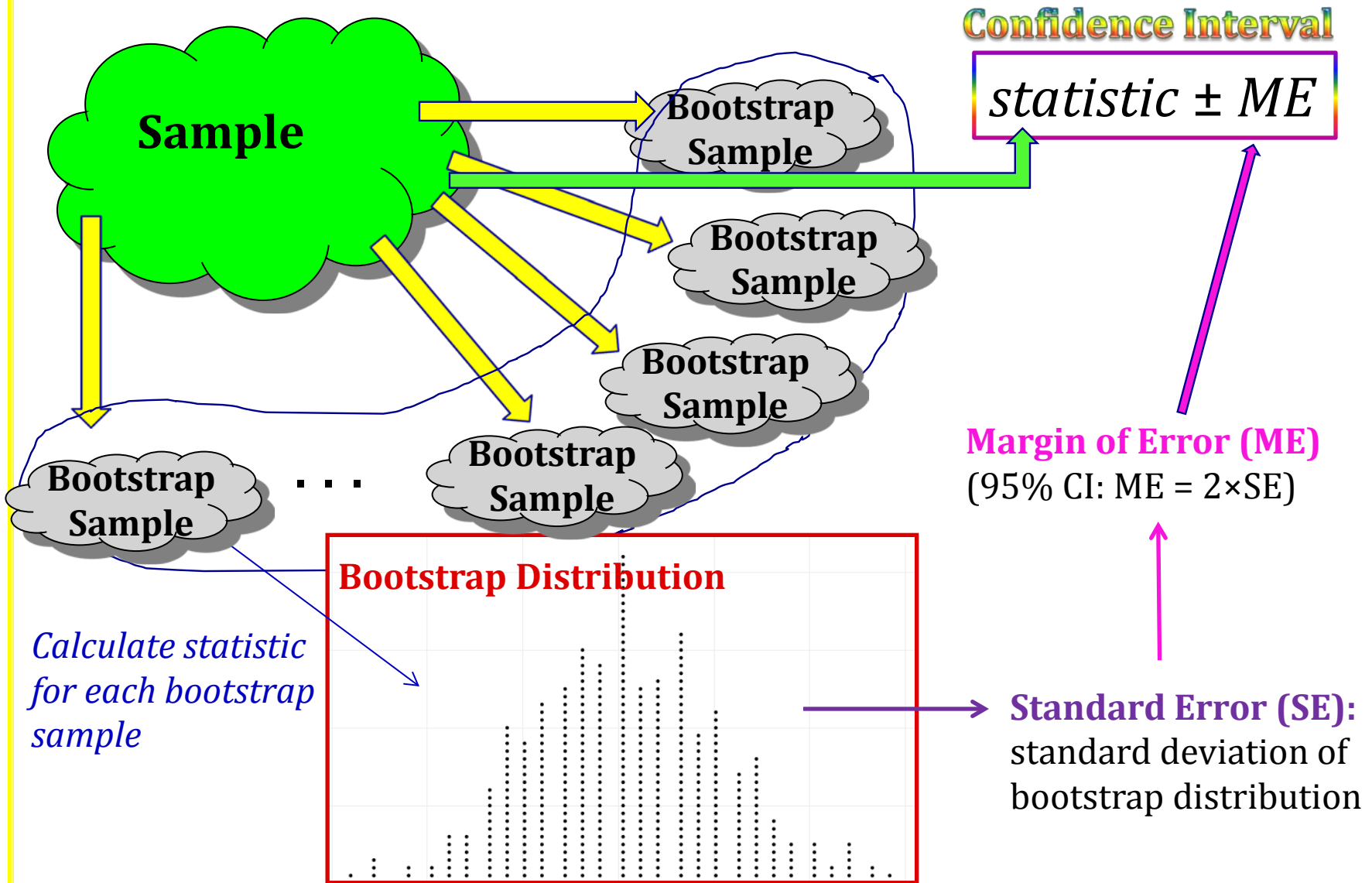
Center

- The sampling distribution is centered around the population parameter
- The bootstrap distribution is centered around the sample statistic
- Luckily, we don't care about the center... we care about the *variability!*

Standard Error

- The variability of the bootstrap statistics is similar to the variability of the sample statistics
- The standard error of a statistic can be estimated using the standard deviation of the bootstrap distribution!

Confidence Intervals



What about Other Parameters?

Estimate the standard error and/or a confidence interval for...

- proportion (p)
- difference in means ($\mu_1 - \mu_2$)
- difference in proportions ($p_1 - p_2$)
- standard deviation (σ)
- correlation (ρ)
- ...

Generate samples with replacement
Calculate sample statistic
Repeat...

The Magic of Bootstrapping

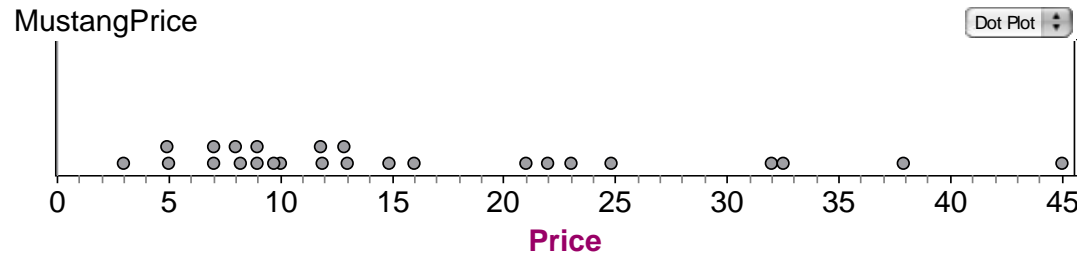
- We can use bootstrapping to assess the uncertainty surrounding ANY sample statistic!
- If we have sample data, we can use bootstrapping to create a 95% confidence interval for any parameter!

(well, almost...)

Used Mustangs

- What's the average price of a used Mustang car?
- Select a random sample of $n = 25$ Mustangs from a website (autotrader.com) and record the price (in \$1,000's) for each car.

Sample of Mustangs:



$$n = 25 \quad \bar{x} = 15.98 \quad s = 11.11$$

Our best estimate for the average price of used Mustangs is \$15,980, but how accurate is that estimate?

BOOTSTRAP!

Original Sample



1. Bootstrap Sample



2. Calculate mean price of bootstrap sample

3. Repeat many times!

Used Mustangs

StatKey Confidence Interval for a Mean, Median, Std. Dev.

Mustang Price (Price) ▾

Show Data Table

Edit Data

Generate 1 Sample

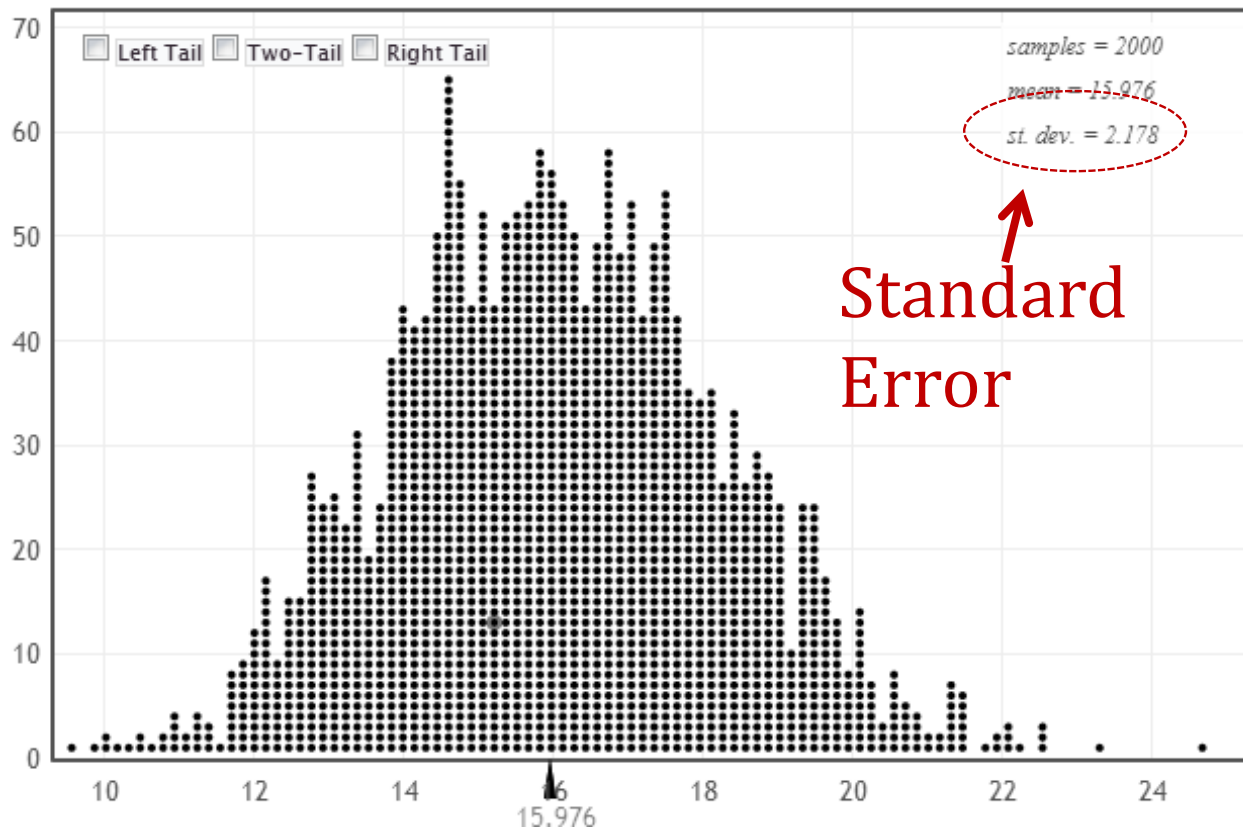
Generate 10 Samples

Generate 100 Samples

Generate 1000 Samples

Reset Plot

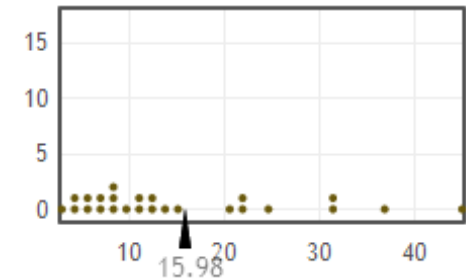
Bootstrap Dotplot of Mean ▾



Original Sample

$n = 25$, mean = 15.98

median = 11.9, stdev = 11.114

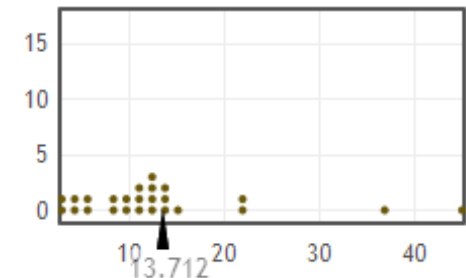


Bootstrap Sample

Show Data Table

$n = 25$, mean = 13.712

median = 11.9, stdev = 9.723



Used Mustangs

- 95% CI:

$$\text{statistic} \pm 2 \cdot SE$$

$$\$15,980 \pm 2 \cdot \$2,178$$

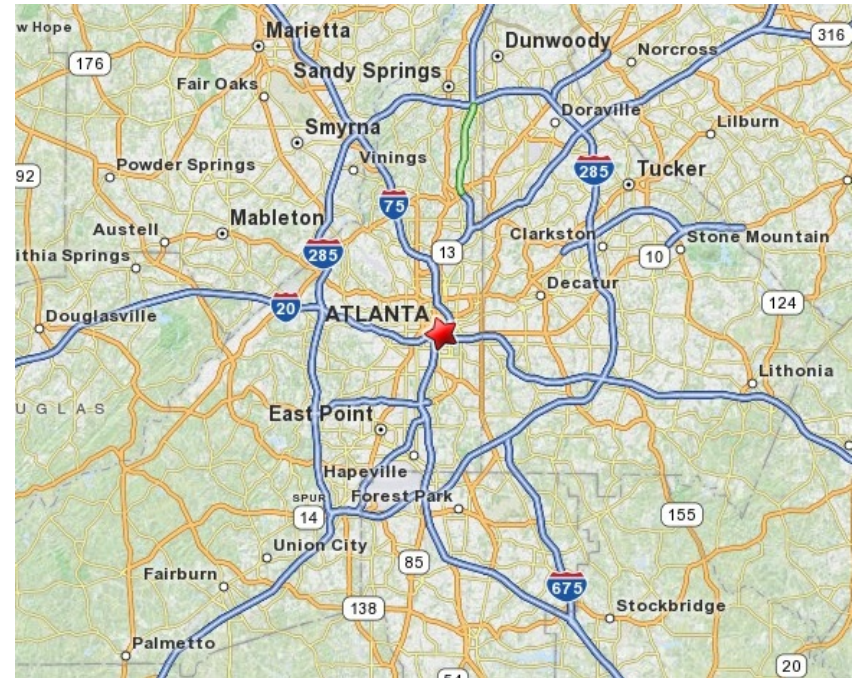
$$(\$11,624, \$20,336)$$

- *We are 95% confident that the average price of a used Mustang on autotrader.com is between \$11,624 and \$20,336*

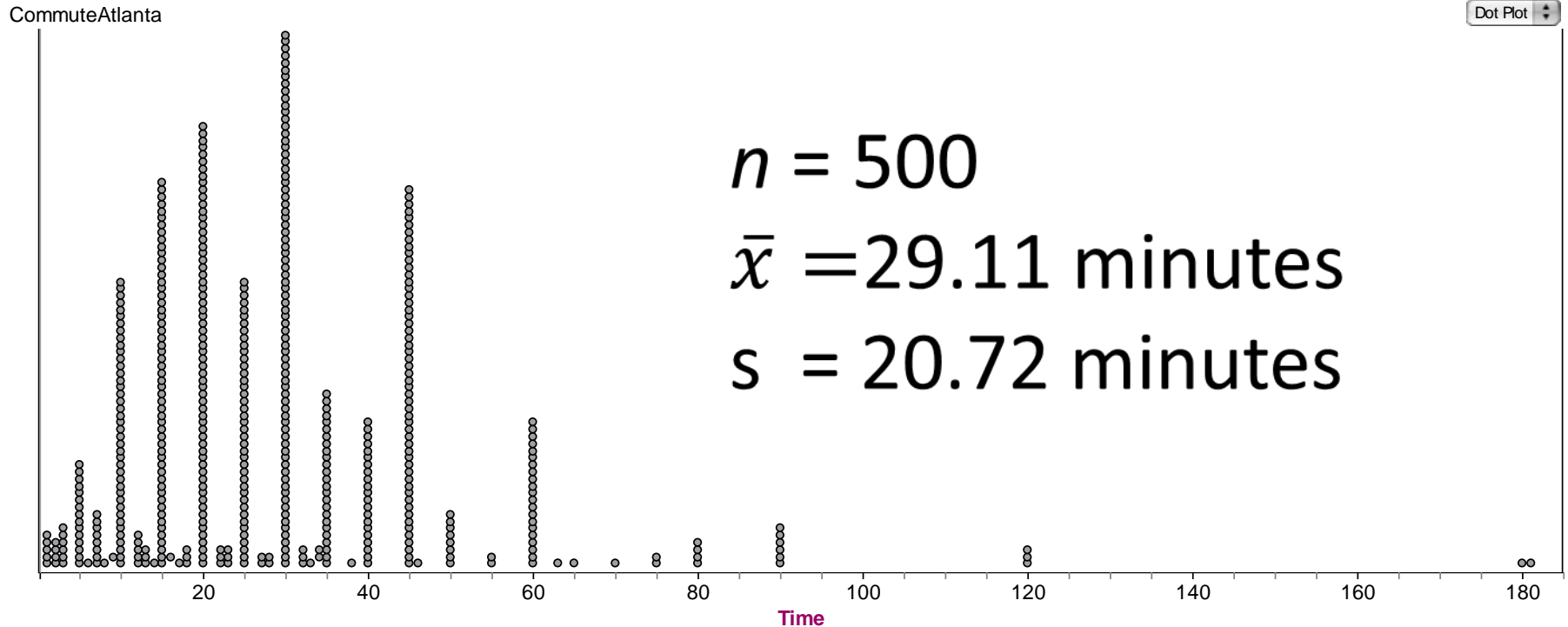
Atlanta Commutes

What's the mean commute time for workers in metropolitan Atlanta?

Data: The American Housing Survey (AHS) collected data from Atlanta in 2004



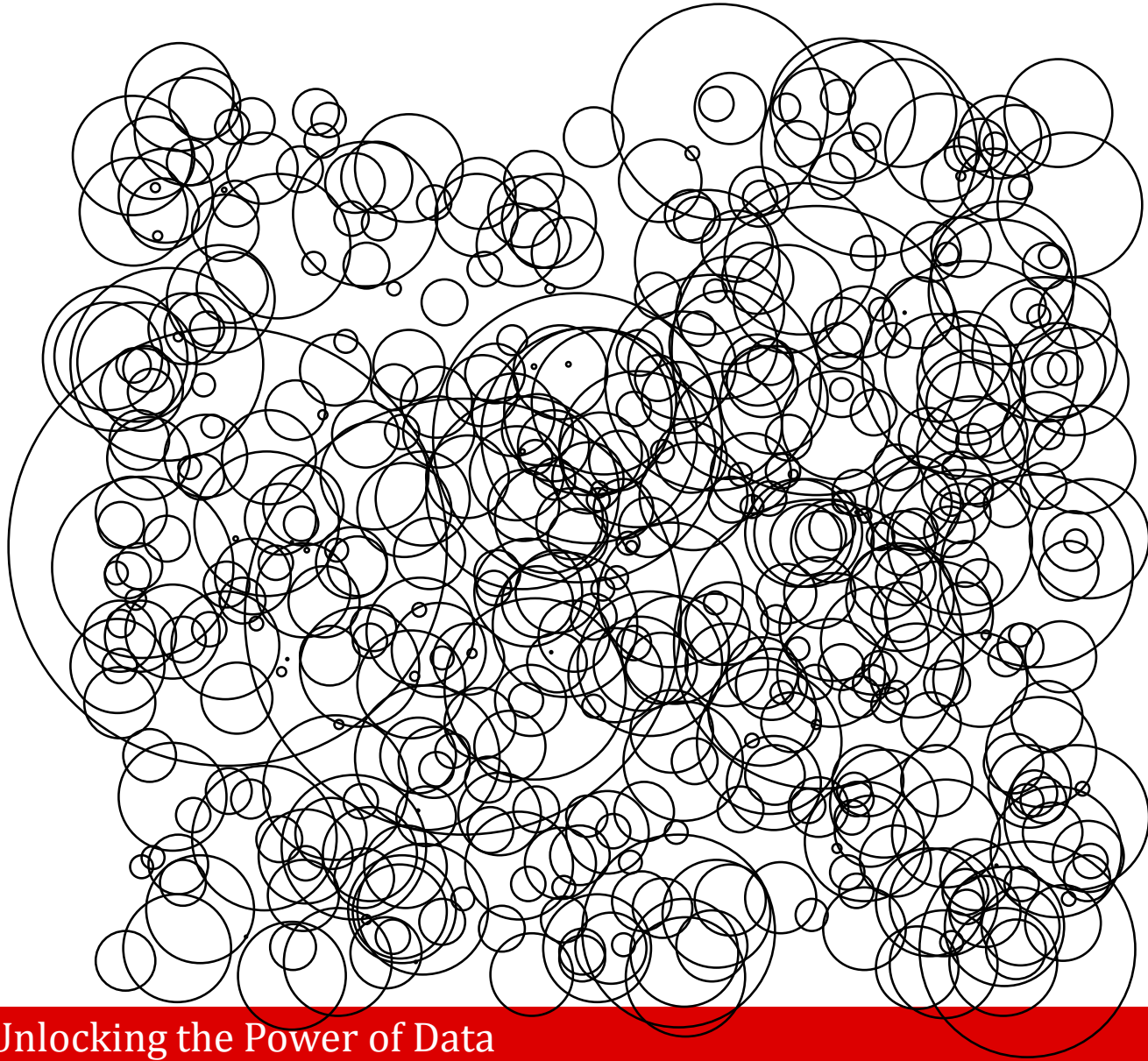
Random Sample of 500 Commutes



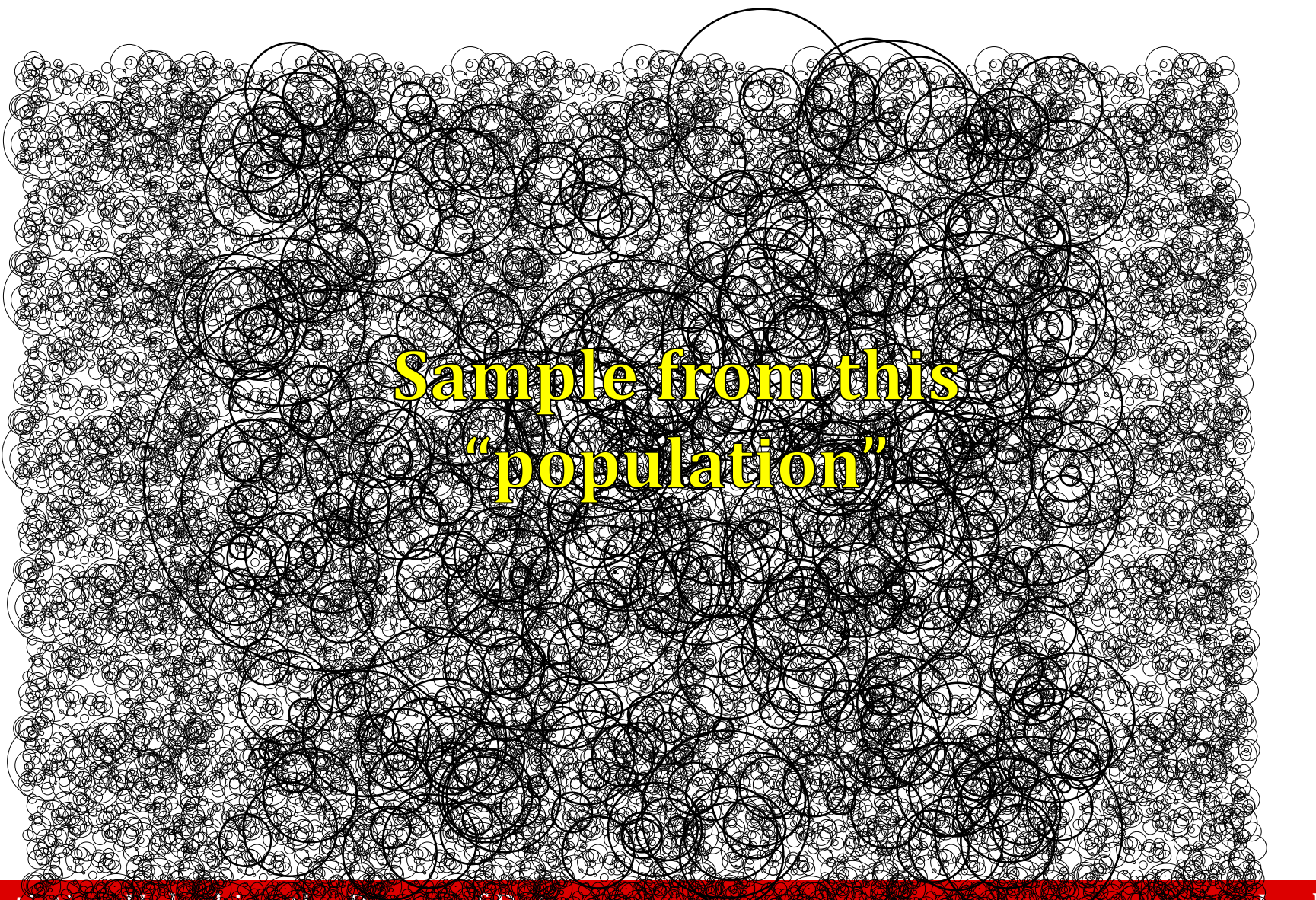
Where might the “true” μ be?

WE CAN BOOTSTRAP TO FIND OUT!!!

Original Sample

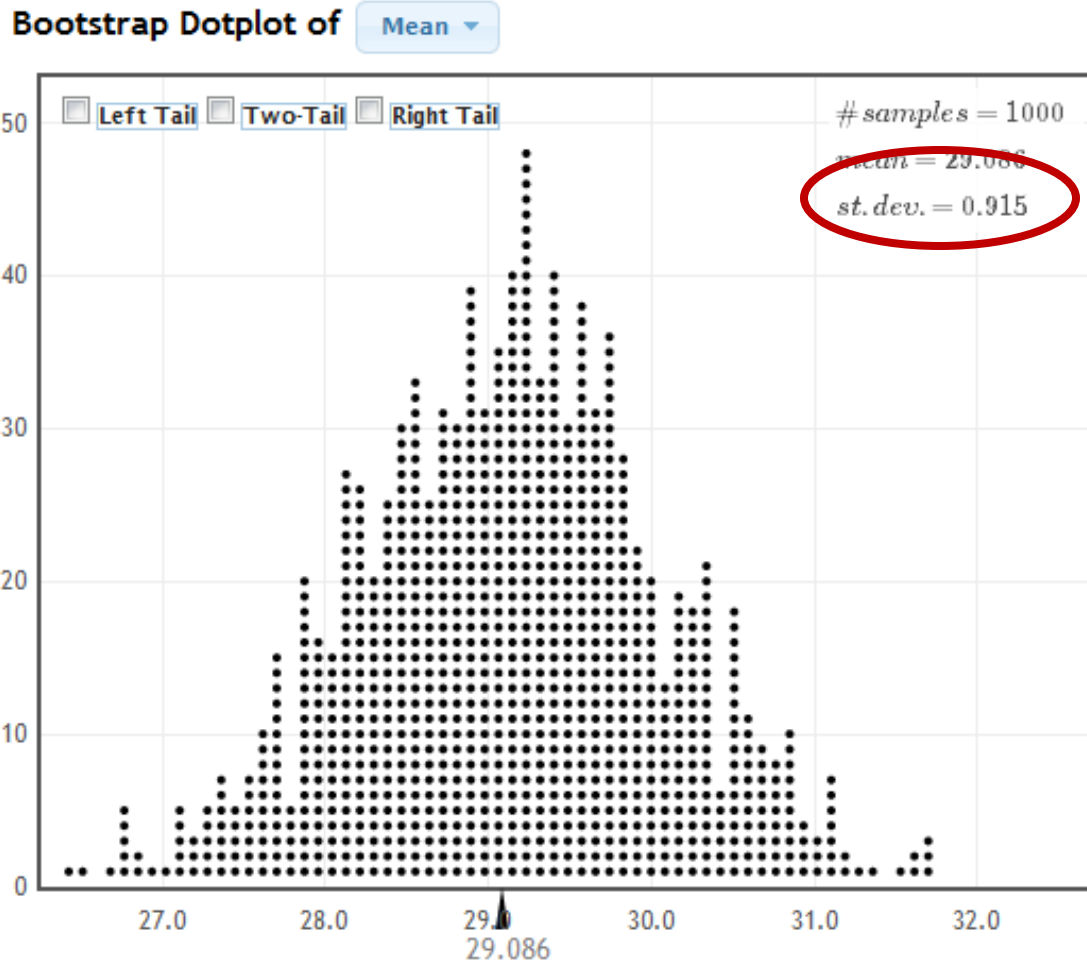


“Population” = many copies of sample



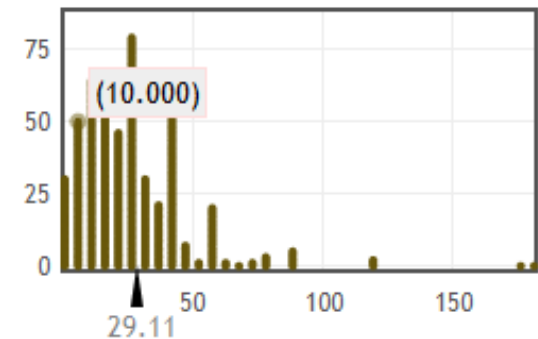
**Sample from this
“population”**

Atlanta Commutes



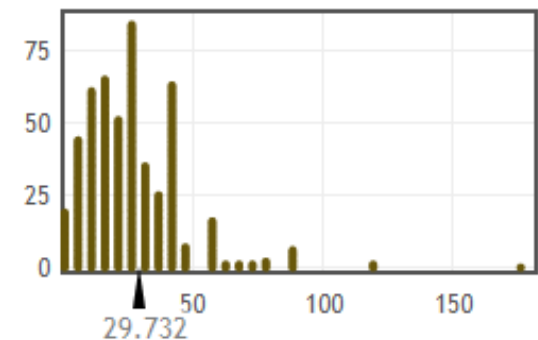
Original Sample

$n = 500$ mean = 29.11
median = 25 stdev = 20.718



Bootstrap Sample

$n = 500$ mean = 29.732
median = 30 stdev = 18.587



95% confidence interval for the average commute time for Atlantans:

$$29.11 \pm 2 \times 0.915$$

27.3 to 30.9



Global Warming

What percentage of Americans believe in global warming?

A survey on 2,251 randomly selected individuals conducted in October 2010 found that 1328 answered “Yes” to the question

“Is there solid evidence of global warming?”

Give and interpret a 95% CI for the proportion of Americans who believe there is solid evidence of global warming.

Source: “Wide Partisan Divide Over Global Warming”, Pew Research Center, 10/27/10.
<http://pewresearch.org/pubs/1780/poll-global-warming-scientists-energy-policies-offshore-drilling-tea-party>

Global Warming

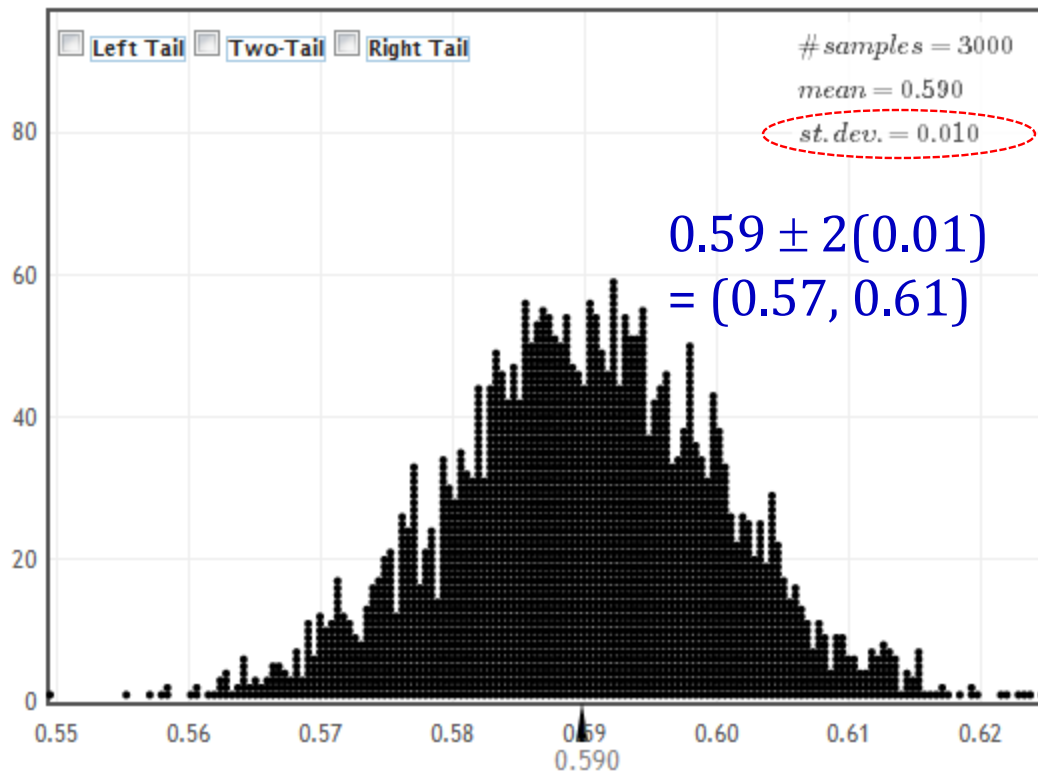
www.lock5stat.com/statkey

Bootstrap For One Categorical Variable [\[Return to StatKey Index\]](#)

Custom Data ▾ Edit Data

Generate 1 Samples Generate 10 Samples Generate 100 Samples Generate 1000 Samples Reset Plot

Bootstrap Dotplot of Proportion ▾



Original Sample

Count	n	Proportion
1328	2251	0.590

Bootstrap Sample

Count	n	Proportion
1304	2251	0.579

We are 95% sure that the true percentage of all Americans that believe there is solid evidence of global warming is between 57% and 61%



Global Warming

Does belief in global warming differ by political party?

“Is there solid evidence of global warming?”

The sample proportion answering “yes” was 79% among Democrats and 38% among Republicans.

(exact numbers for each party not given, but assume $n=1000$ for each group)

Give a 95% CI for the difference in proportions.

Source: “Wide Partisan Divide Over Global Warming”, Pew Research Center, 10/27/10.
<http://pewresearch.org/pubs/1780/poll-global-warming-scientists-energy-policies-offshore-drilling-tea-party>

Global Warming

www.lock5stat.com/statkey

Bootstrap For Two Binary Categorical Variables [\[Return to StatKey Index\]](#)

$$0.41 \pm 2(0.02) \\ = (0.37, 0.45)$$

Custom Data ▾

Edit Data

Generate 1 Samples

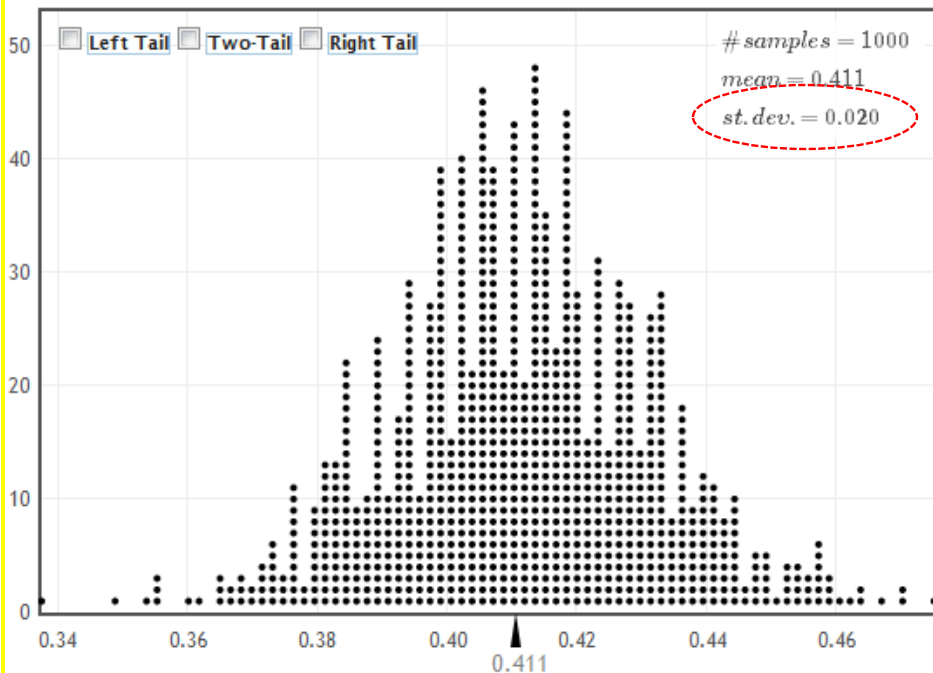
Generate 10 Samples

Generate 100 Samples

Generate 1000 Samples

Reset Plot

Bootstrap Dotplot of $proportion_1 - proportion_2$



Original Sample

Group	Count	n	Proportion
Group 1	790	1000	0.79
Group 2	380	1000	0.38
Group 1-Group 2	410	n/a	0.41

Bootstrap Sample

Group	Count	n	Proportion
Group 1	782	1000	0.78
Group 2	370	1000	0.37
Group 1-Group 2	412	n/a	0.41

We are 95% sure that the difference in the proportion of Democrats and Republicans who believe in global warming is between 0.37 and 0.45.

Global Warming

Based on the data just analyzed, can you conclude with 95% certainty that the proportion of people believing in global warming differs by political party?

Yes. We are 95% confident that the difference is between 0.37 and 0.45, and this interval does not include 0 (no difference)

Summary

- To generate a bootstrap distribution, we
 - Generate *bootstrap samples* by sampling with replacement from the original sample, using the same sample size
 - Compute the statistic of interest, a *bootstrap statistic*, for each of the bootstrap samples
 - Collect the statistics for many bootstrap samples to form a *bootstrap distribution*
- If the bootstrap distribution is symmetric and bell-shaped, a 95% CI can be estimated by $statistic \pm 2 \cdot SE$, where SE can be estimated as the standard deviation of a bootstrap distribution