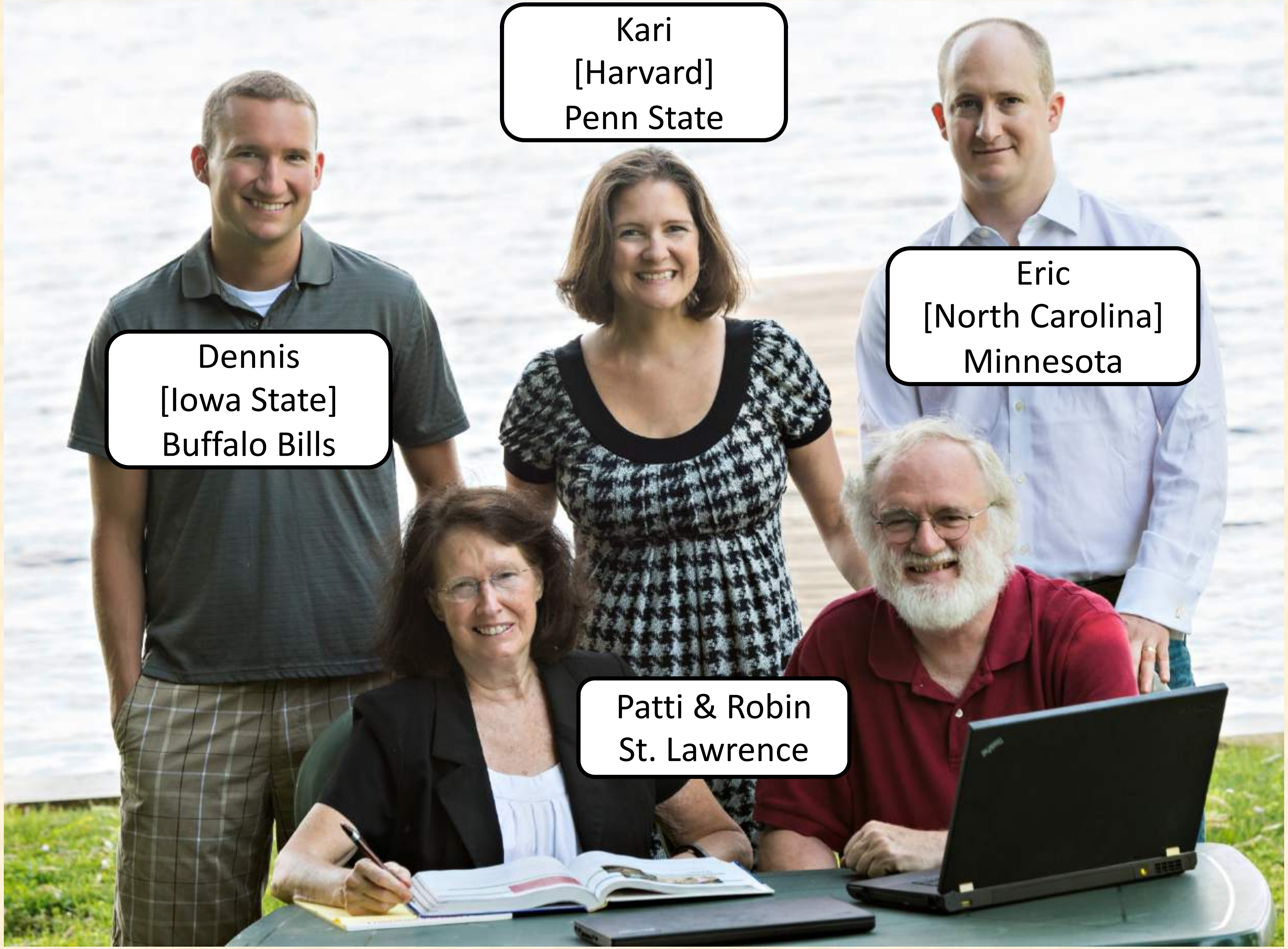
Teaching with Simulation-Based Inference, for Beginners

Robin H Lock, St. Lawrence University
Patti Frazer Lock, St. Lawrence University
Kari Lock Morgan, Penn State University

USCOTS
May 2019

The Lock^5 Team



Introductions:

Name
Institution
Experience

GOAL

Use

Simulation Methods

to

- increase understanding
- reduce prerequisites
- Increase student success

Key Concept: Variability in Sample Statistics

<u>Example</u>: We wish to estimate p = the proportion of Reese's Pieces that are orange.

How much variation is there in the sample statistics $\hat{p}$ if n = 56?

Find the proportion in <u>your</u> sample.
(Then feel free to eat the evidence!)

How much variation is there in the sample statistics $\hat{p}$ if n = 56?

We get a better sense of the variation if:

We have thousands of samples!

We need technology!!

StatKey!!

www.lock5stat.com/statkey

Free, online, works on all platforms, easy to use

Sampling Distribution

- Created from the population
- Centered at the population parameter
- Bell-shaped
- Shows variability in sample statistics

  (Standard Error = standard deviation of sample statistics given a fixed sample size)

Sampling Distribution:
Big Problem!!

In order to create a sampling distribution, we need to already know the population parameter or be able to take thousands of samples!

Not helpful in real life!!

Key Concept:
Variability in Sample Statistics

Overview for Today:

Two simulated distributions for a statistic

- Emphasize the key concept of variability
- Are extremely useful in doing statistics

Simulated Distribution #1

- Created using only the sample data
- Centered at the sample statistic
- Bell-shaped
- Same Variability/Standard Error !!!

Bootstrap Distribution
(for Confidence Intervals)

Simulated Distribution #2

- Created assuming a null hypothesis is true
- Centered at the null hypothesis value
- Bell-shaped
- Same Variability/Standard Error !!!
  (assuming the null hypothesis is true)

Randomization Distribution
(for Hypothesis Tests)

Key Concept:
Variability in Sample Statistics

Sampling Distribution

- Created from the population
- Centered at population parameter
- Bell-shaped
- Gives variability/standard error

Simulated Distribution #1

- Created from the sample
- Centered at sample statistic
- Bell-shaped
- Gives variability/standard error

Simulated Distribution #2

- Created assuming null is true
- Centered at null value
- Bell-shaped
- Gives variability/standard error

Let’s get started…

This material can come very early in a course.
It requires only basic knowledge of summary statistics and sampling.

Bootstrap
Confidence Intervals

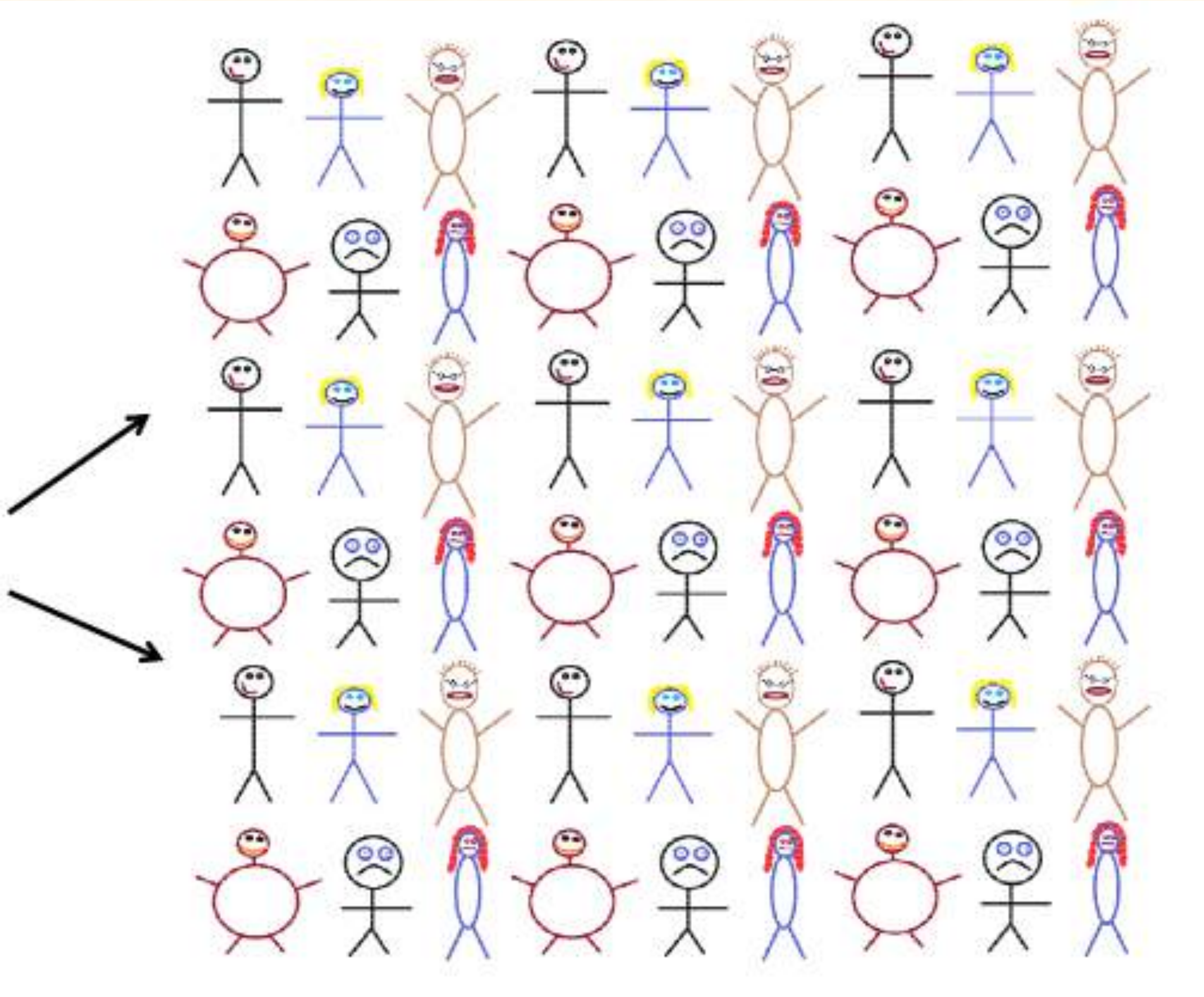
How can we estimate the variability of a statistic when we only have one sample?

Assessing Uncertainty

- Key idea: how much do statistics vary from sample to sample?
- Problem?
  - *We can't take lots of samples from the population!*
- Solution?
  - *(re)sample from our best guess at the population – the sample itself!*

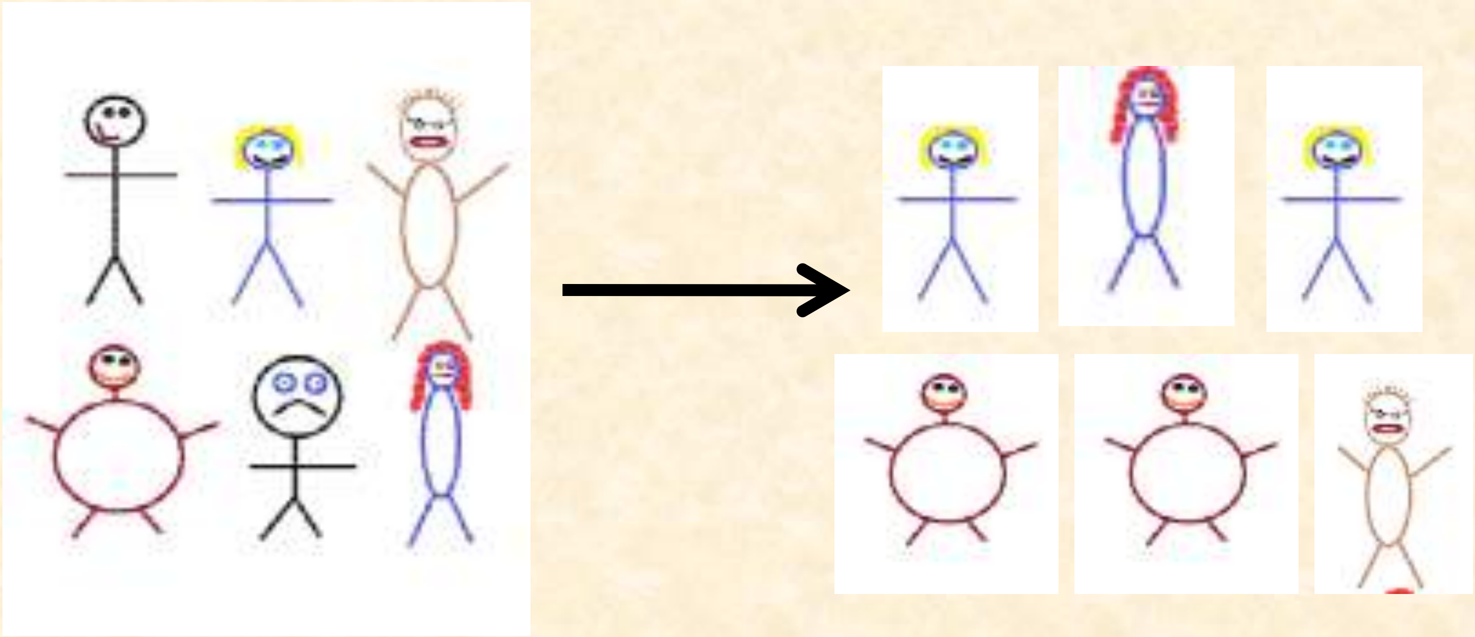
Suppose we have a random sample of 6 people:

Original Sample

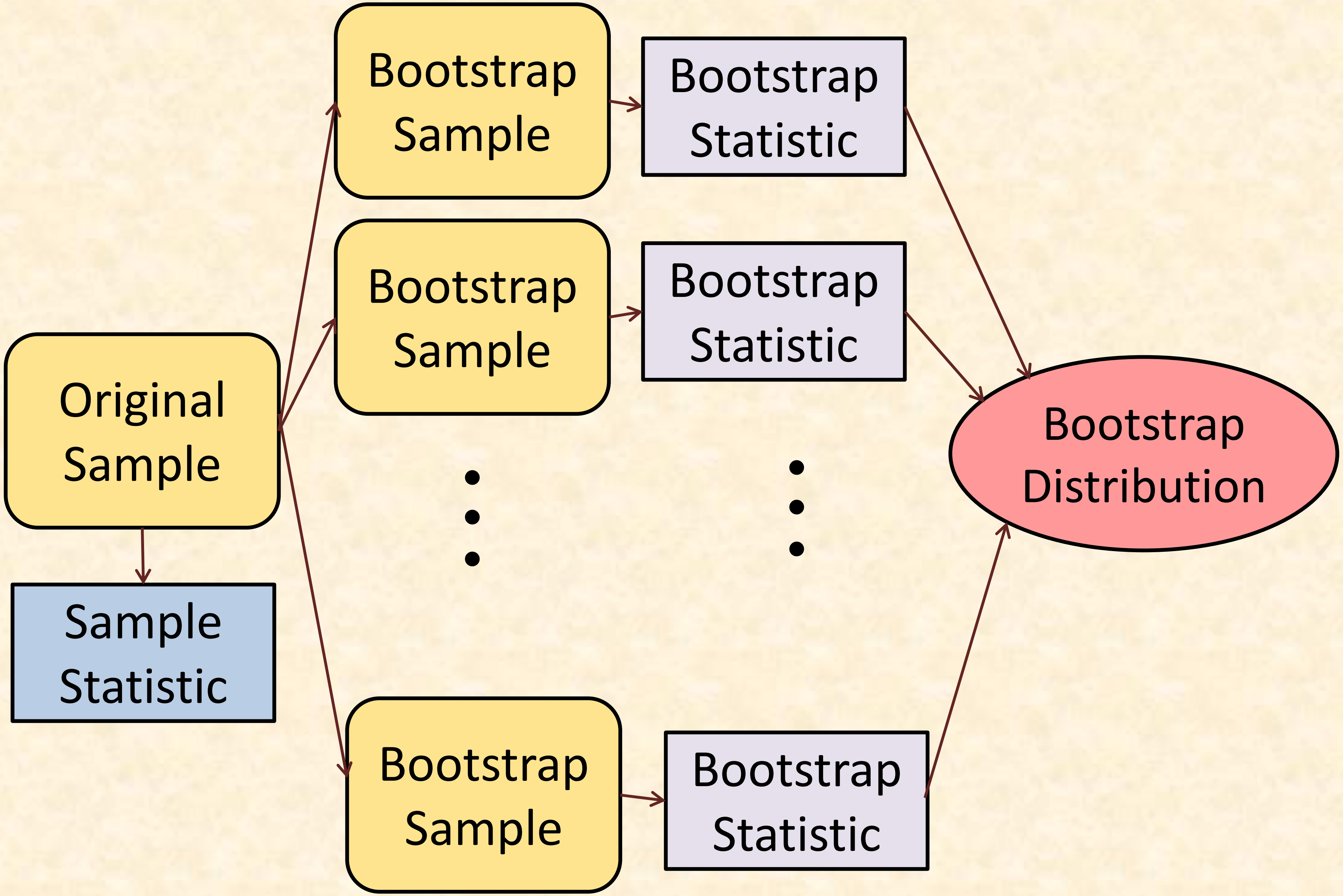


A simulated “population” to sample from

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

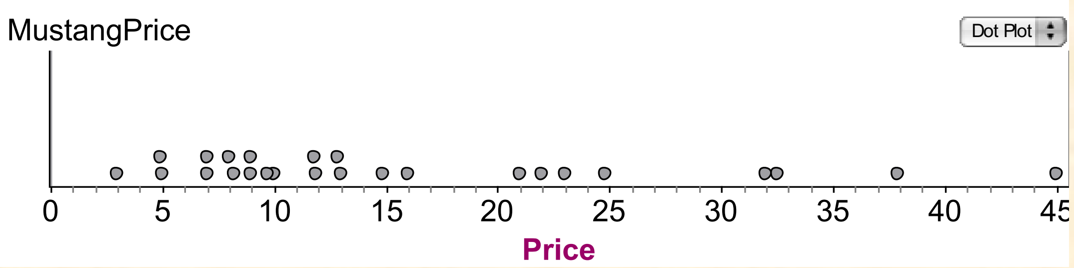


Original Sample   Bootstrap Sample


Original Sample
Sample Statistic
Bootstrap Sample
Bootstrap Sample
Bootstrap Sample
Bootstrap Statistic
Bootstrap Statistic
Bootstrap Statistic
Bootstrap Distribution

Example: Mustang Prices

Start with a random sample of 25 prices (in $1,000's)



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

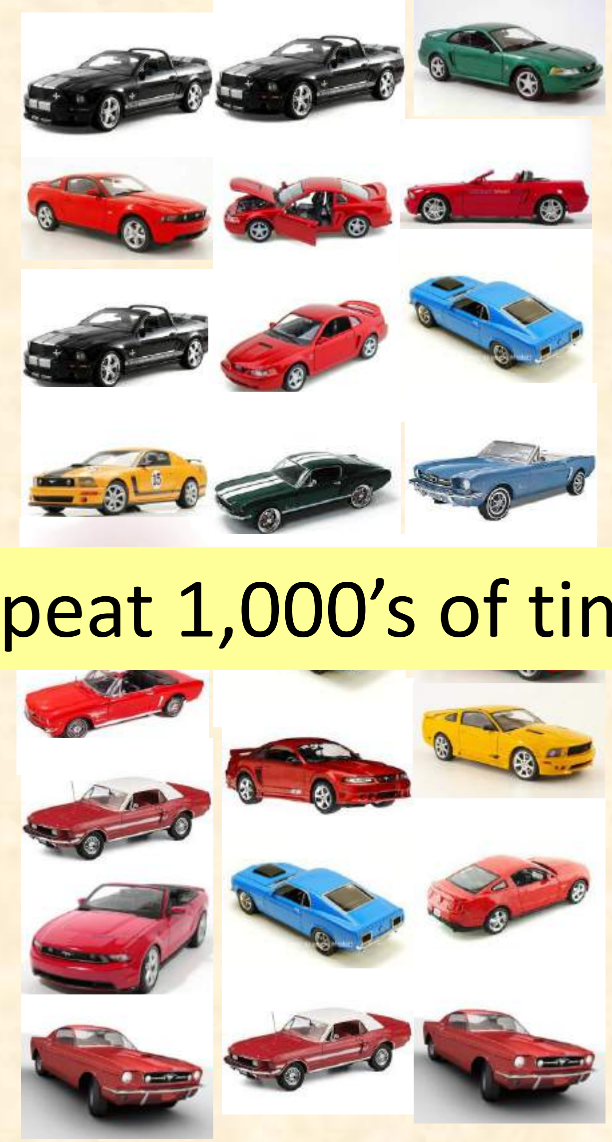
Goal: Find an interval that is likely to contain the mean price for all Mustangs

Key concept: How much can we expect means for samples of size 25 to vary just by random chance?

Original Sample

$\bar{x} = 15.98$

Bootstrap Sample



Repeat 1,000's of times!

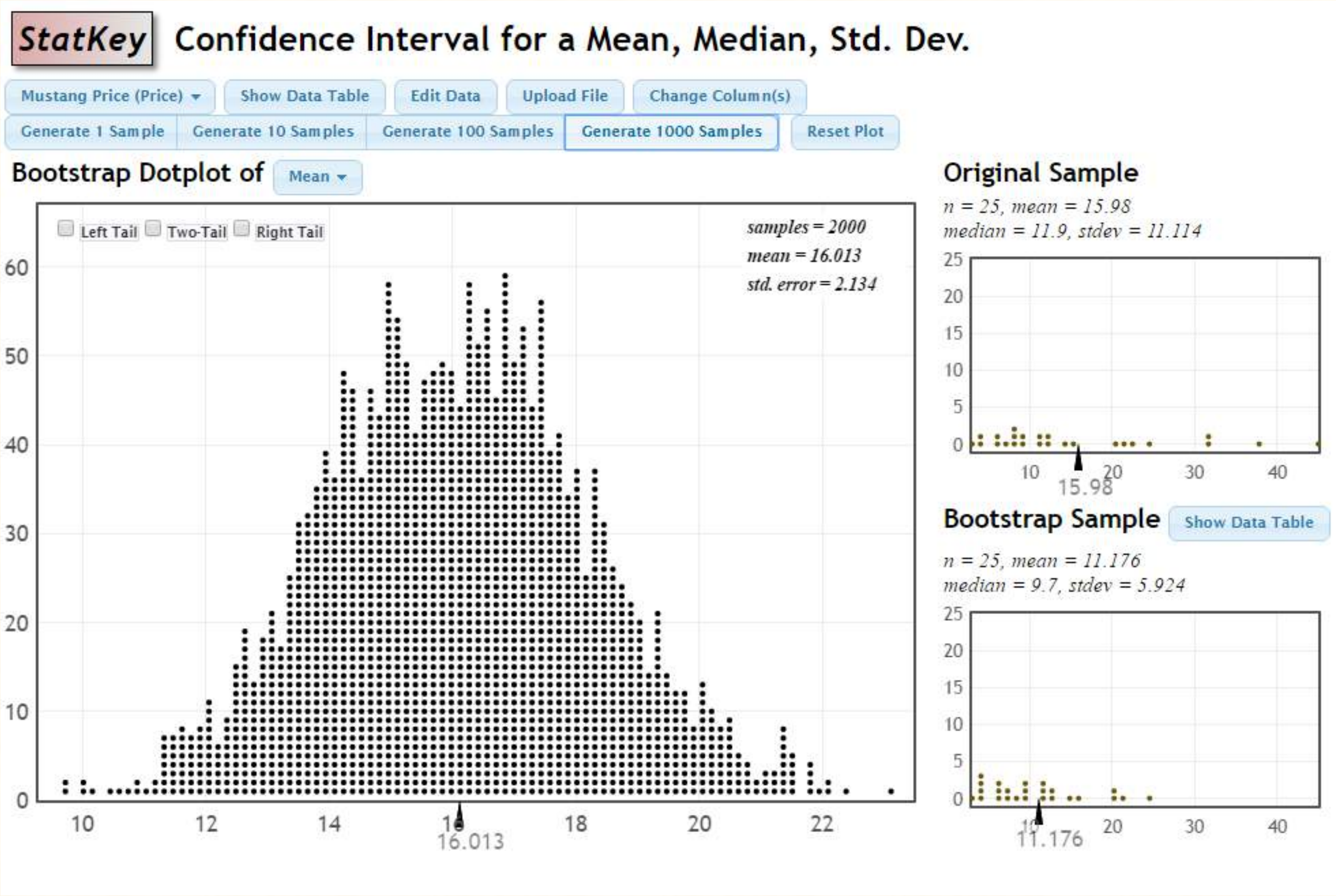
$\bar{x} = 17.51$

We need technology!!

StatKey!!

www.lock5stat.com/statkey

Bootstrap Distribution for Mustang Price Means

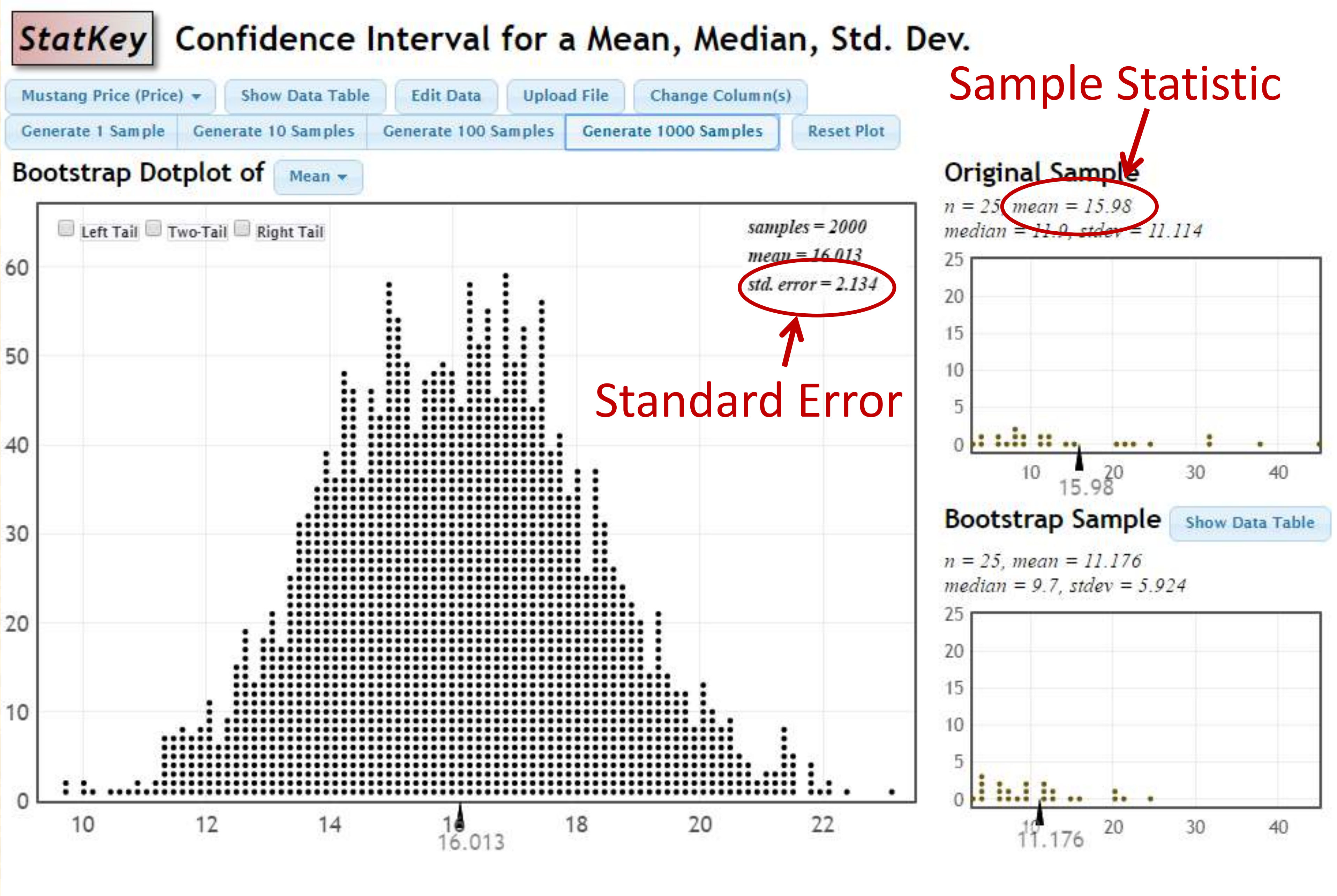


How do we get a CI from the bootstrap distribution?

Method #1: Standard Error

- Find the standard error (SE) as the standard deviation of the bootstrap statistics
- Find an interval with

$$Original\ Statistic \pm 2 \cdot SE$$



$$Statistic \pm 2 \cdot SE = 15.98 \pm 2 \cdot (2.134) = (11.71, 20.25)$$

How do we get a CI from the bootstrap distribution?

Method #1: Standard Error

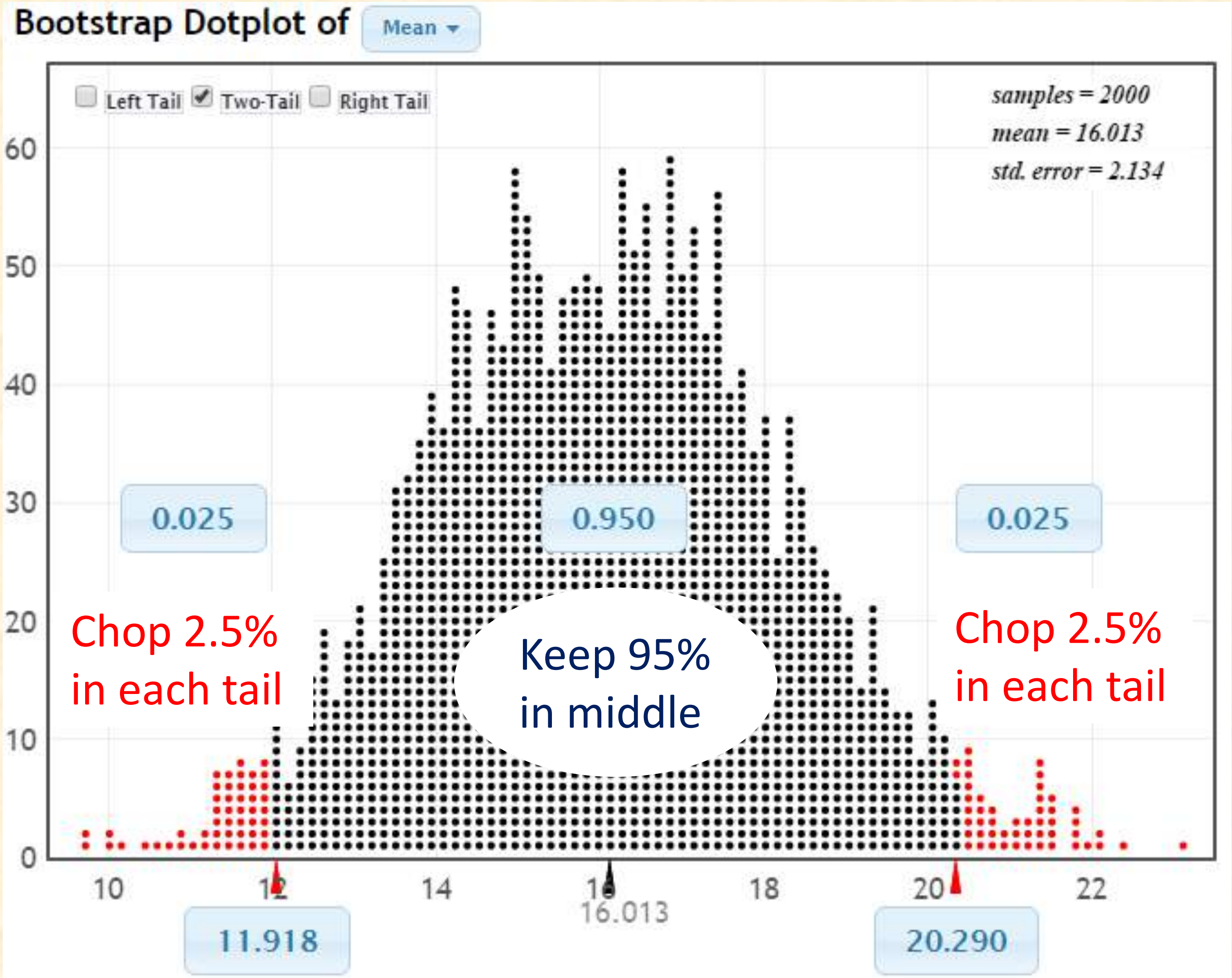
- Find the standard error (SE) as the standard deviation of the bootstrap statistics
- Find an interval with

$$Original\ Statistic \pm 2 \cdot SE$$

Method #2: Percentile Interval

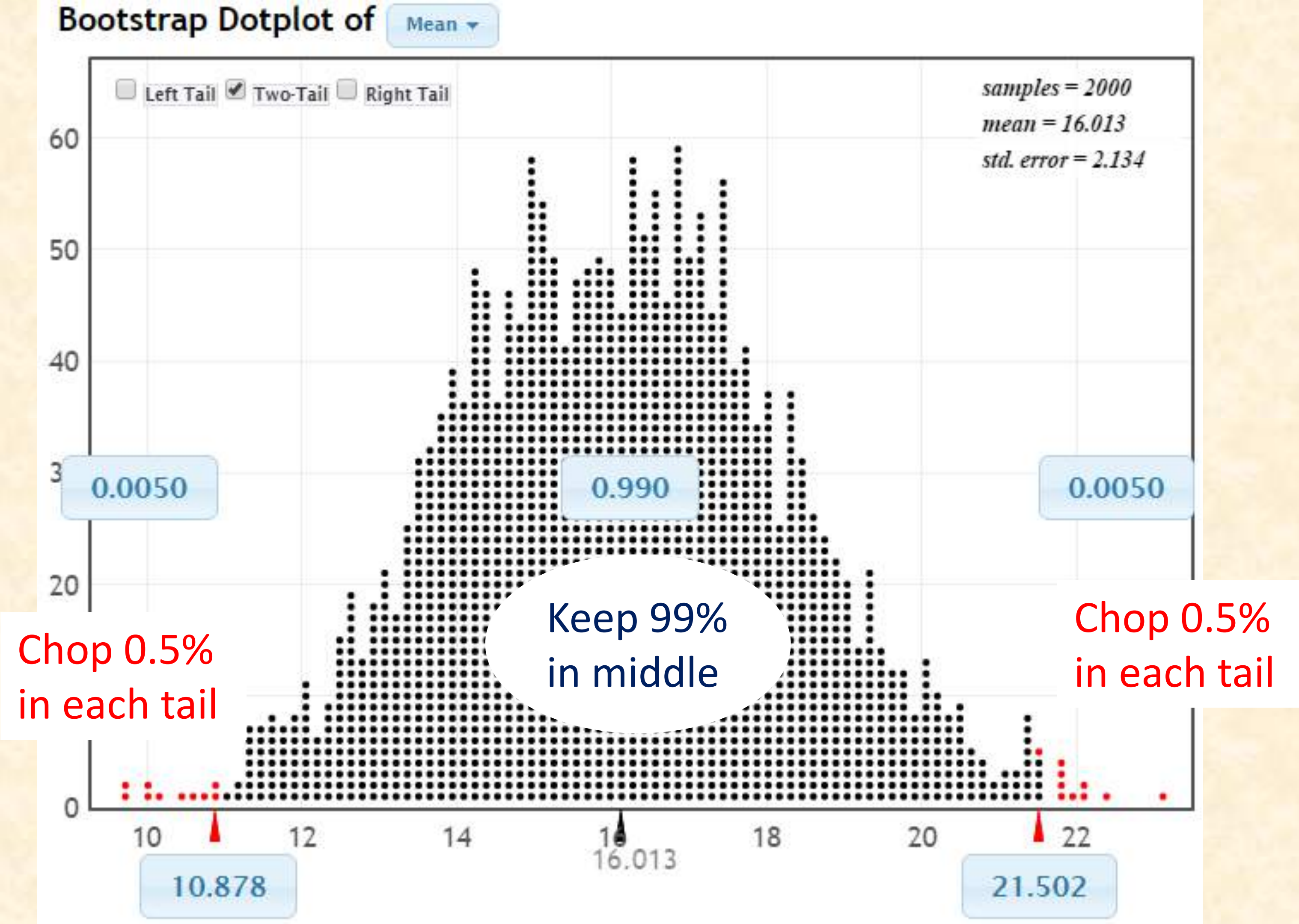
- For a 95% interval, find the endpoints that cut off 2.5% of the bootstrap means from each tail, leaving 95% in the middle

95% CI via Percentiles



We are 95% sure that the mean price for Mustangs is between $11,918 and $20,290

99% CI via Percentiles



We are 99% sure that the mean price for Mustangs is between $10,878 and $21,502

Bootstrap Confidence Intervals

<u>Version 1 (Statistic $\pm$ 2 SE)</u>:
Great preparation for moving to traditional methods

<u>Version 2 (Percentiles)</u>:
Great at building understanding of confidence level

Bootstrap Approach

- Create a **bootstrap distribution** by simulating many samples from the original data, with replacement, and calculating the sample statistic for each new sample.
- Estimate **confidence interval** using either statistic $\pm$ 2 SE or the middle 95% of the bootstrap distribution.

Same process works for different parameters!

Your Turn!

Three more examples.

Example #1: Atlanta Commutes

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

Data: The American Housing Survey (AHS) collected data from Atlanta. We have a microdata sample of 500 commuters from that sample.

Find a 95% confidence interval for the mean commute time for all Atlanta commuters.

Example #2: News Sources

How do people like to get their news?

A recent Pew Research poll asked “ Do you prefer to get your news by watching it, reading it, or listening to it?”

They found that 1,164 of the 3,425 US adults sampled said they preferred reading the news.

Use this information to find a 90% confidence interval for the proportion of all US adults who prefer to get news by reading it.

https://www.journalism.org/2018/12/03/americans-still-prefer-watching-to-reading-the-news-and-mostly-still-through-television/

Example #3: Diet Cola and Calcium

How much does diet cola affect calcium excretion?

In an experiment, 16 healthy women were randomly assigned to drink 24 ounces of either diet cola or water.

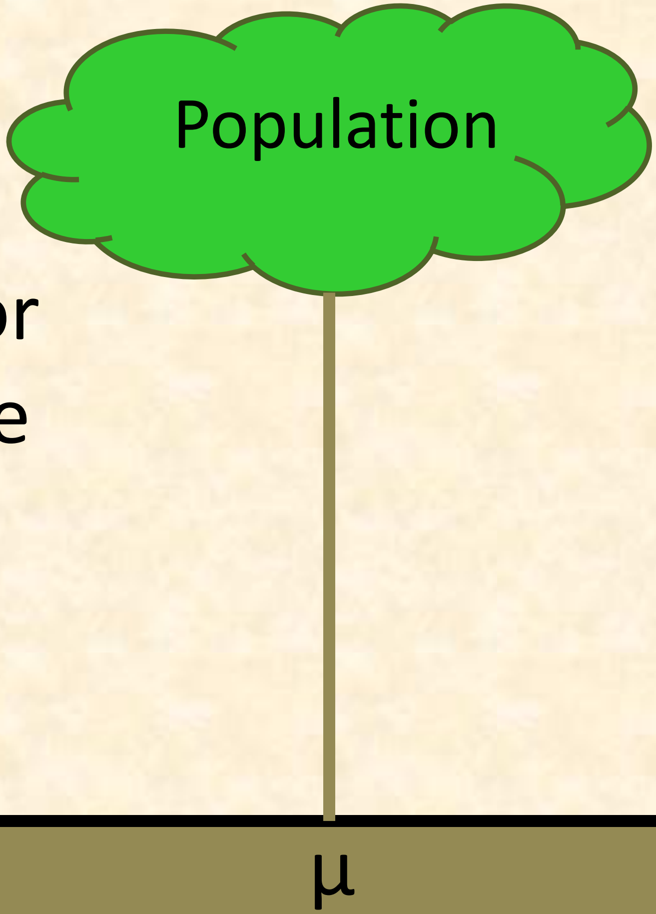
Calcium loss (in mg) was measured in urine over the next few hours.

Use this information to find a 95% confidence interval for the difference in mean calcium loss between diet cola and water drinkers.

Larson, N.S. et al, 'Effect of Diet Cola on Urine Calcium Excretion", Endocrine Reviews, 2010.

Sampling Distribution

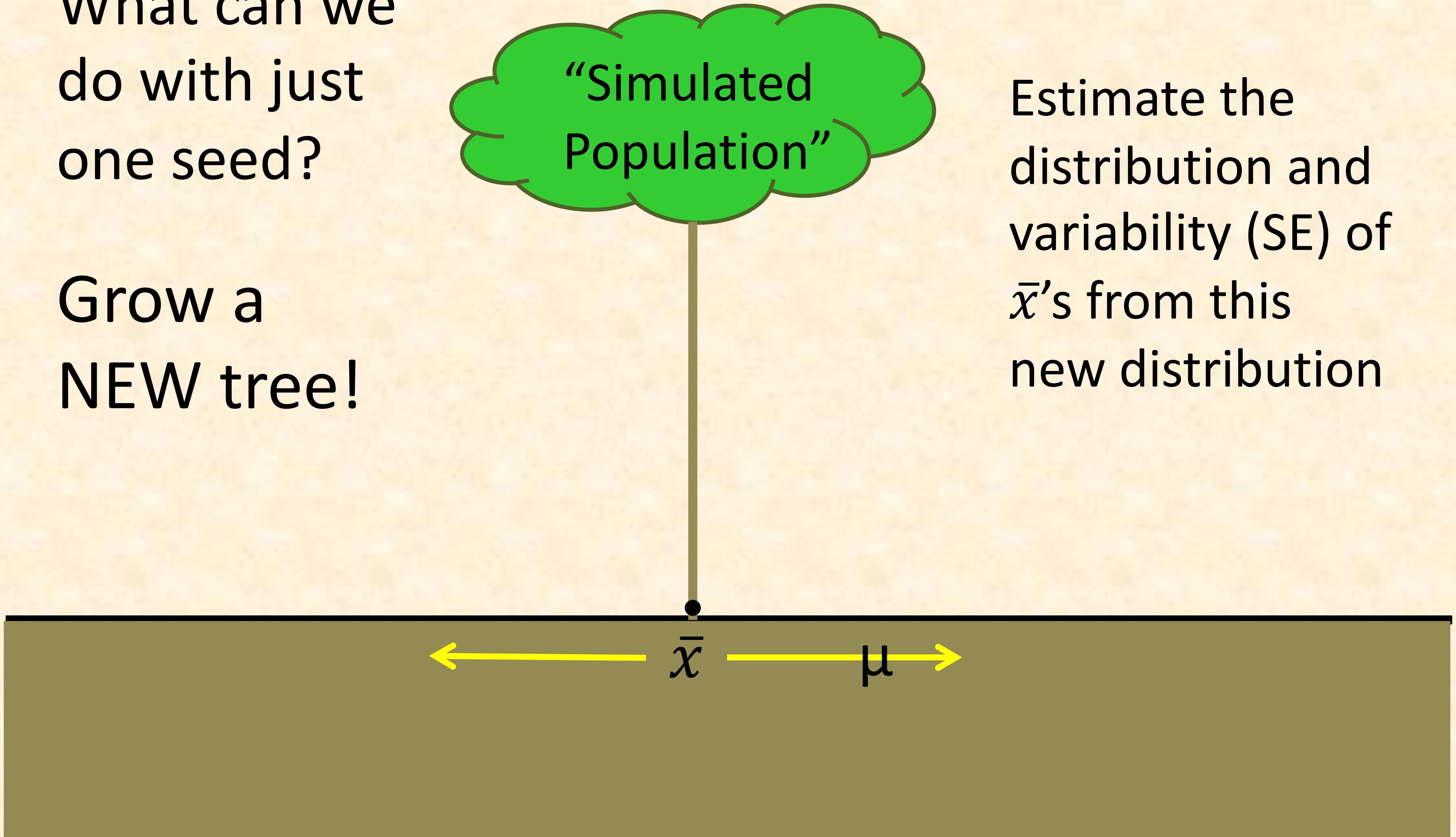
BUT, in practice we don’t see the “tree” or all of the “seeds” – we only have ONE seed



Using only the Sample Data

What can we do with just one seed?

Grow a NEW tree!



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

Randomization Hypothesis Tests

Key Concept: *How do we measure strength of evidence?*

Example: Beer & Mosquitoes

Question: Does consuming beer attract mosquitoes?

Experiment:
25 volunteers drank a liter of beer,
18 volunteers drank a liter of water
Randomly assigned!
Mosquitoes were caught in traps as they approached the volunteers.[1]

[1] Lefvre, T., et. al., "Beer Consumption Increases Human Attractiveness to Malaria Mosquitoes," *PLoS ONE,* 2010; 5(3): e9546.

Beer and Mosquitoes

Number of Mosquitoes

Beer	Water
27	21
20	22
21	15
26	12
27	21
31	16
24	19
19	15
23	24
24	19
28	23
19	13
24	22
29	20
20	24
17	18
31	20
20	22
25	
28	
21	
27	
21	
18	
20	

Beer mean = 23.6

Water mean = 19.22

Beer mean – Water mean = 4.38

Does drinking beer actually attract mosquitoes or is the difference just due to random chance?

Beer and Mosquitoes

Number of Mosquitoes

Beer	Water
27	21
20	22
21	15
26	12
27	21
31	16
24	19
19	15
23	24
24	19
28	23
19	13
24	22
29	20
20	24
17	18
31	20
20	22
25	
28	
21	
27	
21	
18	
20	

What kinds of results would we see, just by random chance, if there were no difference between beer and water?

Beer and Mosquitoes

Number of Mosquitoes

Beer					Water
27	27	19	21	24	21
20	20	24	18	19	22
21	21	29	20	23	15
26	26	20	21	13	12
27	27	27	22	22	21
31	31	31	15	20	16
24	24	20	12	24	19
19	19	25	21	18	15
23	23	28	16	20	24
24	24	21	19	22	19
28	28	27	15		23
19					13
24					22
29					20
20					24
17					18
31					20
20					22
25					
28					
21					
27					
21					
18					
20					

What kinds of results would we see, just by random chance, if there were no difference between beer and water?

We can find out!! Just re-randomize the 43 values into one pile of 25 and one of 18, simulating the original random assignment.

Beer and Mosquitoes

Number of Mosquitoes

Beer | Water

27	19	21	24
[illegible]	24	[illegible]	19
[illegible]	29	[illegible]	23
[illegible]	20	[illegible]	13
[illegible]	27	[illegible]	22
[illegible]	31	[illegible]	20
[illegible]	20	[illegible]	24
[illegible]	25	[illegible]	18
[illegible]	28	[illegible]	20
[illegible]	21	[illegible]	22
[illegible]	27	[illegible]	
21		20	
18		27	
15		21	
21		17	
16		24	
28		28	
22			
19			
27			
20			
23			
22			
21			

What kinds of results would we see, just by random chance, if there were no difference between beer and water?

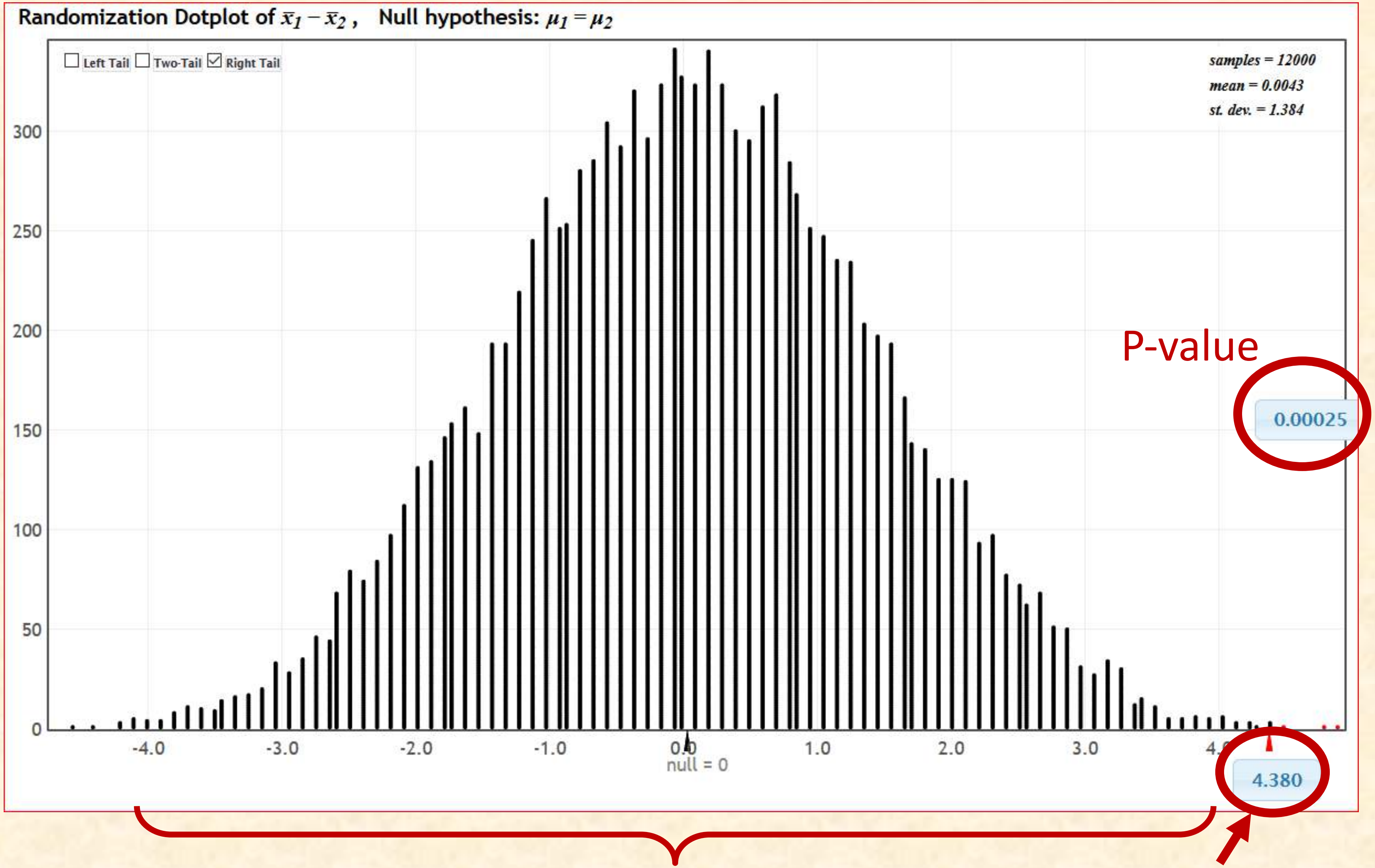
Compute the beer mean minus water mean of this simulated sample.

Do this thousands of times!

We need technology!!

StatKey!!

www.lock5stat.com/statkey



This is what we are likely to see just by random chance if beer/water doesn't matter.

This is what we saw in the experiment.

Beer and Mosquitoes
The Conclusion!

The results seen in the experiment are very unlikely to happen just by random chance (less than 1 out of 1000!)

We have strong evidence that drinking beer does attract mosquitoes!

What about the traditional approach?

“Students’ approach to p-values … was procedural … and [they] did not attach much meaning to p-values”

-- Aquilonius and Brenner, “Students’ Reasoning about P-Values”, SERJ, November 2015

Another Look at Beer/Mosquitoes

1. Check conditions

2. Which formula?

$$t = \frac{\bar{x}_B - \bar{x}_W}{\sqrt{\frac{s_B^2}{n_B} + \frac{s_W^2}{n_W}}}$$

3. Calculate numbers and plug into formula

$$t = \frac{23.6 - 19.22}{\sqrt{\frac{4.1^2}{25} + \frac{3.7^2}{18}}}$$

4. Chug with calculator

$$t = 3.68$$

0.0005 < p-value < 0.001

5. Which theoretical distribution?

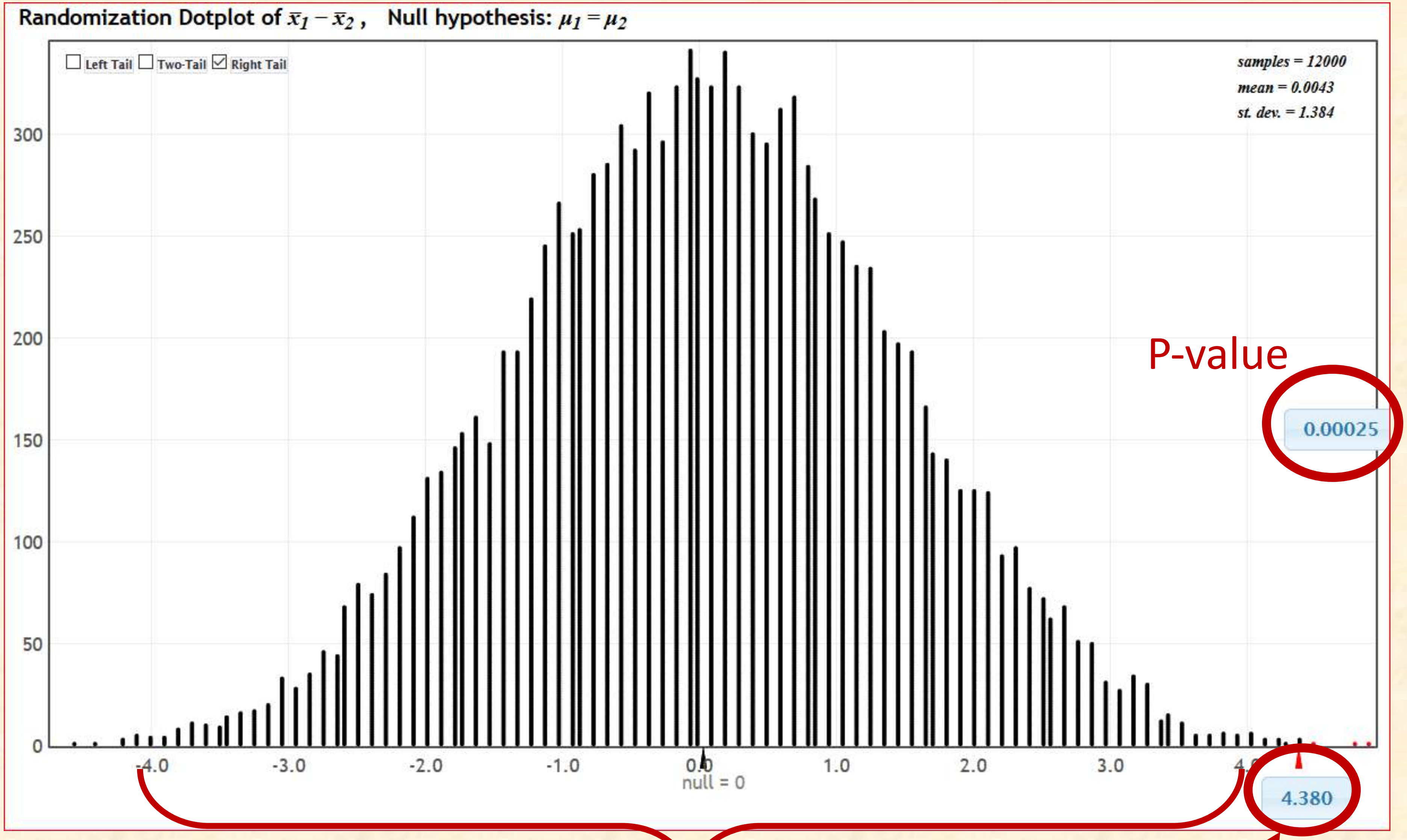
6. df?

7. Find p-value

8. Interpret a decision

TABLE B: *t*-DISTRIBUTION CRITICAL VALUES

	Tail probability *p*											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	.685	.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	.679	.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	.679	.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	.678	.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	.677	.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	.675	.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
∞	.674	.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level *C*											



This is what we are likely to see just by random variation if the null hypothesis is true.

This is what we saw in the study.

Conclusions are the same, but the process is very different!

Your Turn!

Three more examples.

Example #1: Light at Night

Does having a light on at night increase weight gain? (in mice)

Experiment:
10 mice had a light on at night
8 mice had darkness at night
Randomly assigned!
Weight gain (in grams) was recorded after three weeks.

Do the data provide convincing evidence that the mean weight gain is higher with a light at night?

[1] Fonken, L., et. al., "Light at night increases body mass by shifting time of food intake," *Proceedings of the National Academy of Sciences,* 2010; 107(43).

Example #2: Malevolent Uniforms

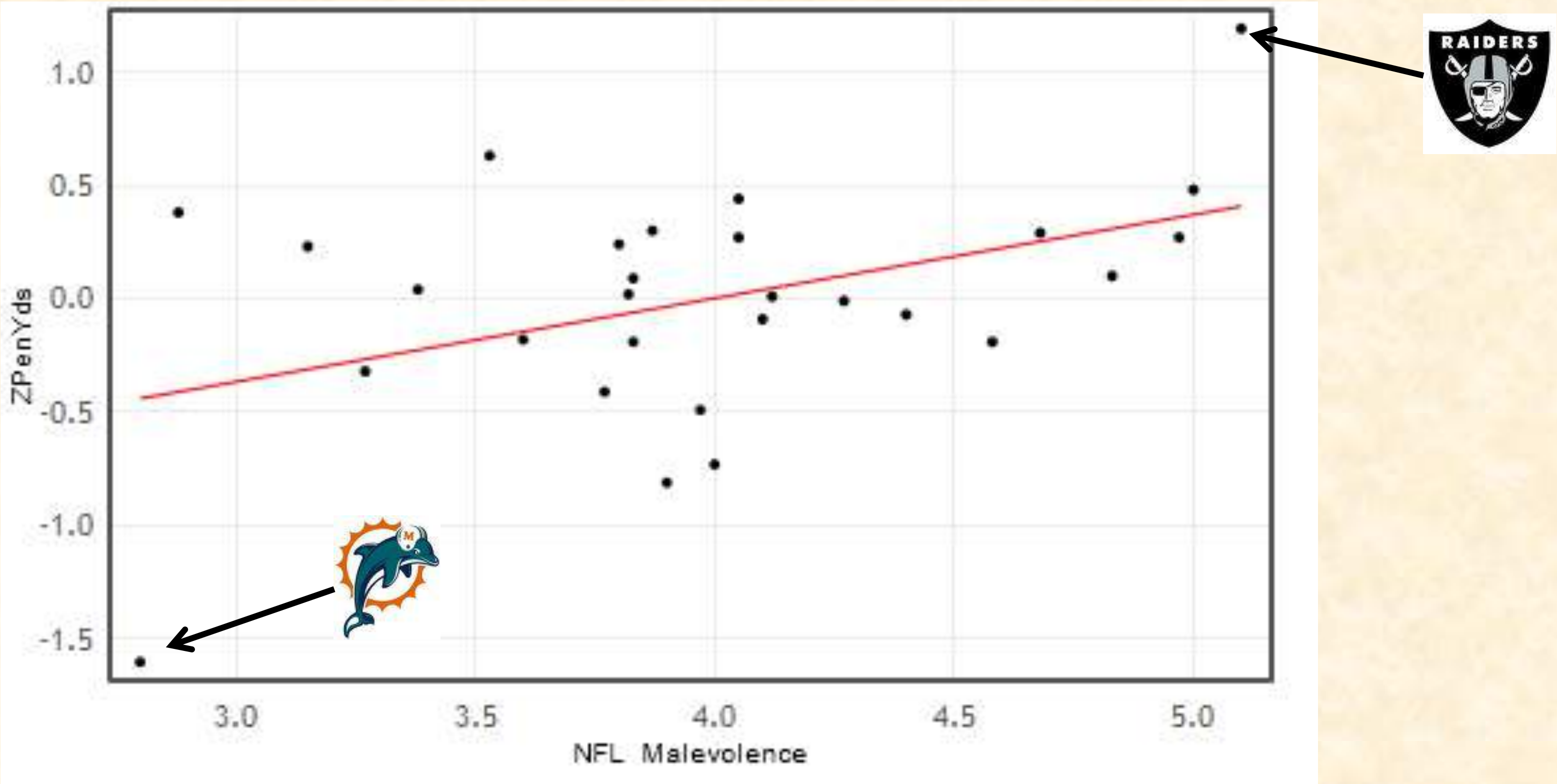
Do football teams with more malevolent uniforms and logos tend to get more penalties?

NFL_Malevolence is scored so that larger values indicate more malevolence.
ZPenYds is a z-score for penalty yards over a season.

Is there convincing evidence of a positive correlation between malevolence and penalty yards?

Ffrank, M.G. and Gilovich, T. “The Dark Side of Self- and Social Perception: Black Uniforms and Aggression in Professional Sports”, Journal of Personality and Social Psychology (1988)

Example #2: Malevolent Uniforms



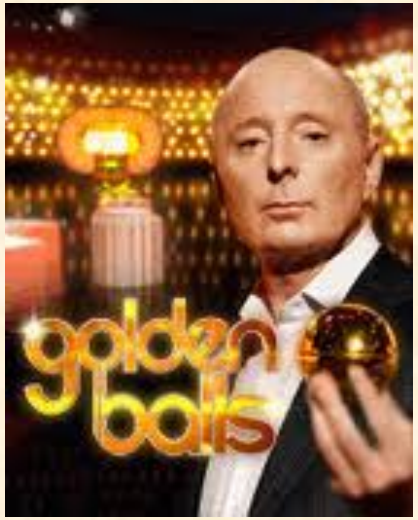
Is there convincing evidence of a positive correlation between malevolence and penalty yards?

Frank, M.G. and Gilovich, T. “The Dark Side of Self- and Social Perception: Black Uniforms and Aggression in Professional Sports”, Journal of Personality and Social Psychology (1988)

Example #3: Split or Steal?

http://www.youtube.com/watch?v=p3Uos2fzIJ0

Are younger players less likely to split than older players?



	Split	Steal	Total
Under 40	187	195	382
40 & over	116	76	192
Total	303	271	n=574

Is there convincing evidence that the proportion who split is lower for players under 40 than 40 & over?

Van den Assem, M., Van Dolder, D., and Thaler, R., "Split or Steal? Cooperative Behavior When the Stakes Are Large," 2/19/11.

Simulation Methods

Visual!

Intuitive!

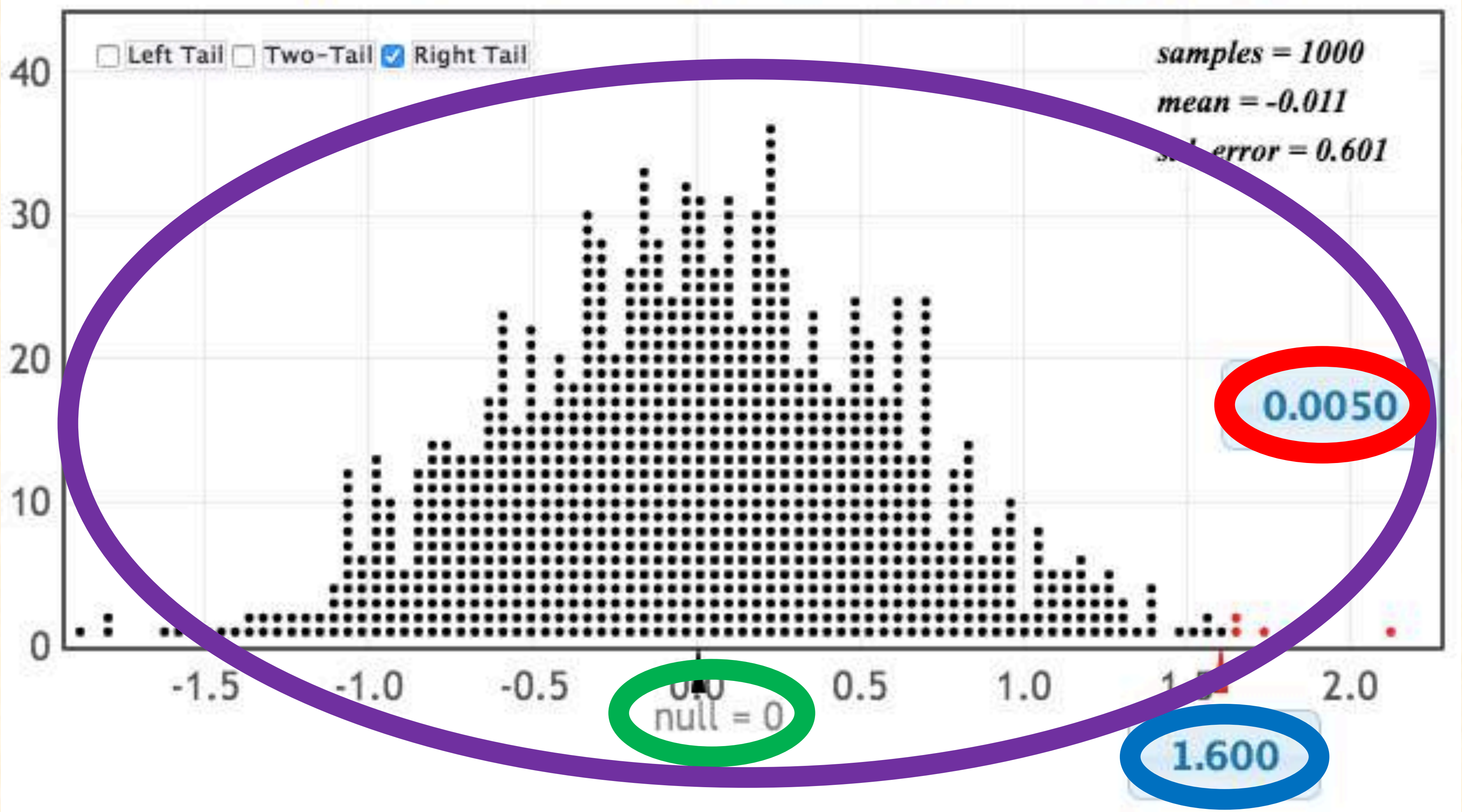
Easily incorporates active learning!

Ties directly to key idea of strength of evidence!

Ties directly to key idea of random variation!

(Note: No theoretical distributions. No formulas. No algebra. Same process for all parameters.)

p-value: The chance of obtaining a statistic as extreme as that observed, just by random chance, if the null hypothesis is true



What about Traditional Inference?

After seeing simulation-based inference:

- Students have seen **lots** of "bell-shaped" distributions and dealt often with finding "proportions in tails".
- Students have seen standard error as a measure of variability of sample statistics
- Students have seen how to understand and interpret confidence intervals
- Students have seen the conceptual underpinnings of hypothesis tests, and how to interpret a p-value.

What about Traditional Inference?

After seeing simulation-based inference:

- Students are ready to learn about theoretical distributions such as the normal distribution.
- Students are ready to learn how to calculate standard error from formulas and summary statistics.
- They already understand the basic ideas of inference!

Transitioning to Traditional Inference

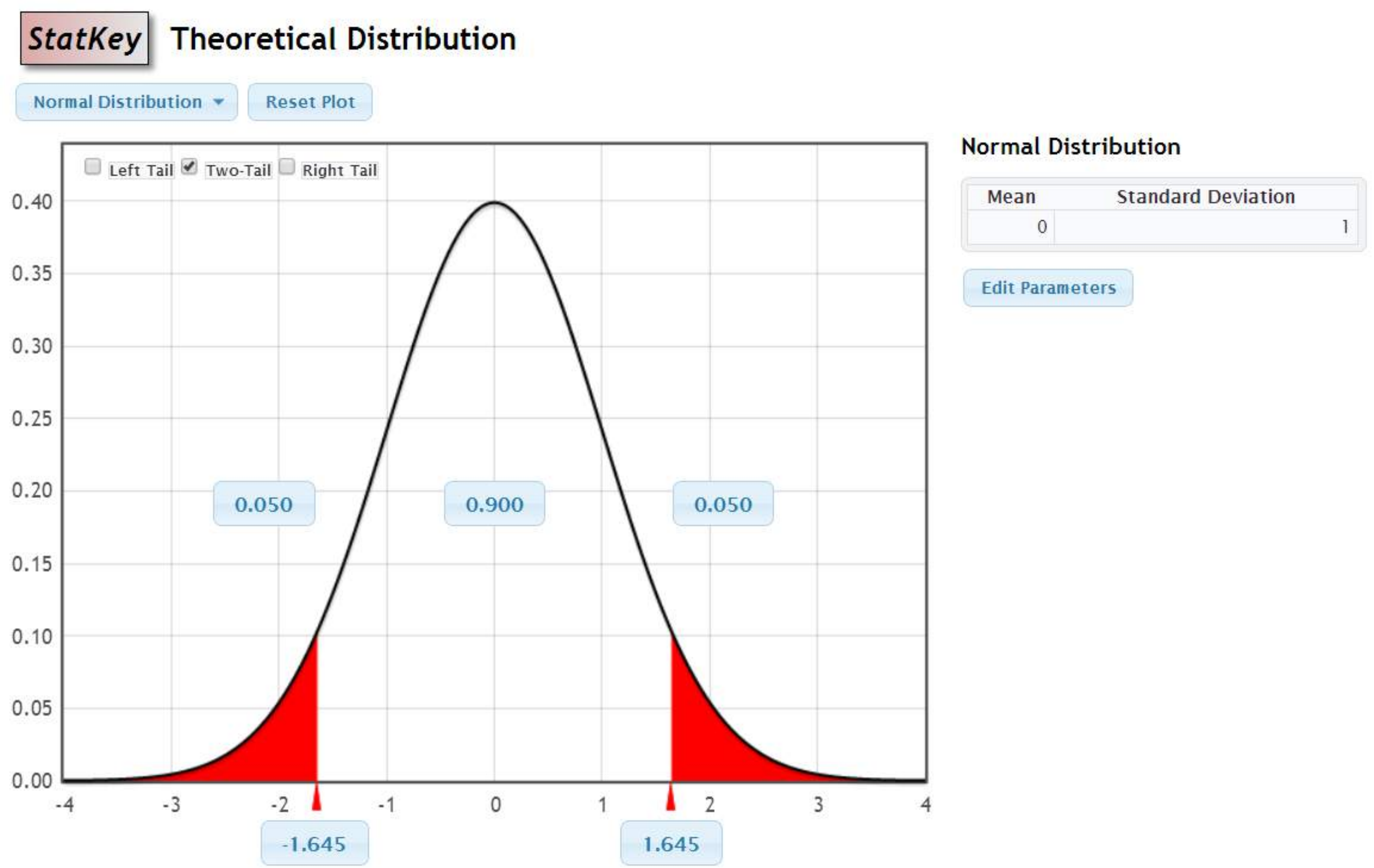
Confidence Interval:

$$Sample\ Statistic\ \pm z^* \cdot SE$$

Hypothesis Test:

$$z = \frac{Sample\ Statisic - Null\ Parameter}{SE}$$

StatKey – Theoretical Distributions



Motivation

"... Before computers statisticians had no choice. These days we have no excuse. Randomization-based inference makes a direct connection between data production and the logic of inference that deserves to be at the core of every introductory course."

-- Professor George Cobb, 2007

(TISE article at *http://escholarship.org/uc/item/6hb3k0nz)*

"Actually, the statistician does not carry out this very simple and very tedious process, but his conclusions have no justification beyond the fact that they agree with those which could have been arrived at by this elementary method."

-- Sir R. A. Fisher, 1936

GOAL

Use

Simulation Methods

to

- increase understanding
- reduce prerequisites
- Increase student success

Questions?

rlock@stlawu.edu
plock@stlawu.edu
klm47@psu.edu

www.lock5stat.com