Week 5 Confidence interval

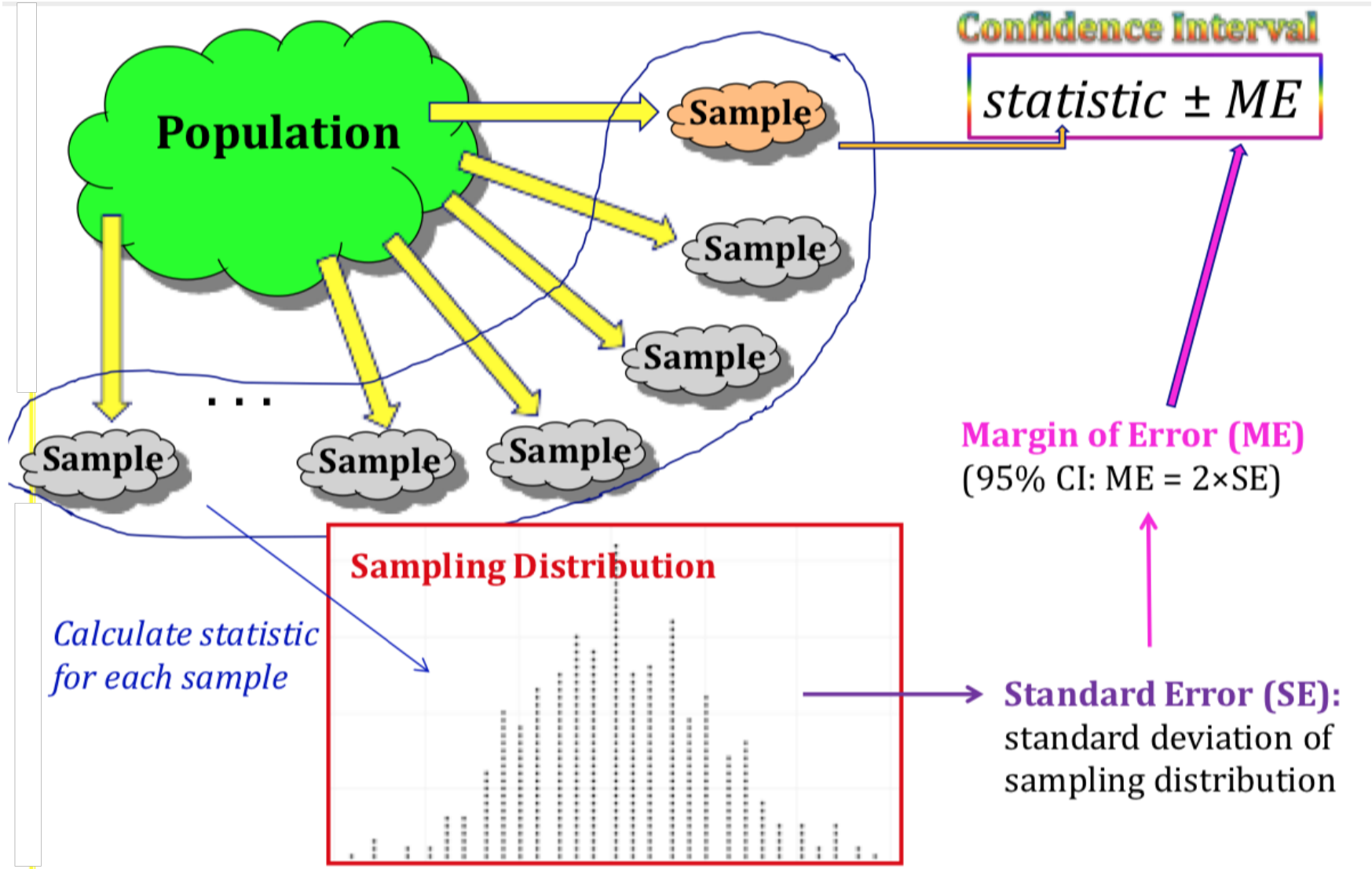
5. Bootstrapping method

Stat 140 - 04

Mount Holyoke College

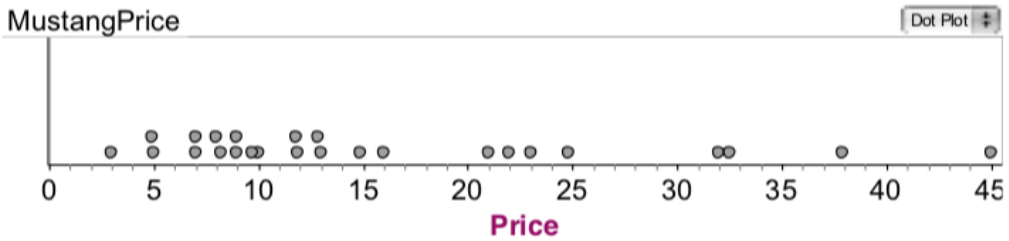
Dr. Shan Shan

Slides posted at *http://sshanshans.github.io/stat140*


Population
Sample
Sample
Sample
Sample
Sample
Sample
Confidence Interval
statistic ± ME
Margin of Error (ME)
(95% CI: ME = 2×SE)
Sampling Distribution
Calculate statistic
for each sample
Standard Error (SE):
standard deviation of
sampling distribution

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.



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

Poll question

Does CLT based method work here?

a. Yes
b. No

What is the **median** year of minting of all pennies used in the U.S. in 2019?

We collected a sample of 50 pennies from a local bank in downtown Northampton, Massachusetts, USA, and record the year of minting for each coin.

Poll question

Does CLT based method work here?

a. Yes
b. No

Some problems we might encounter when using CLT based methods,

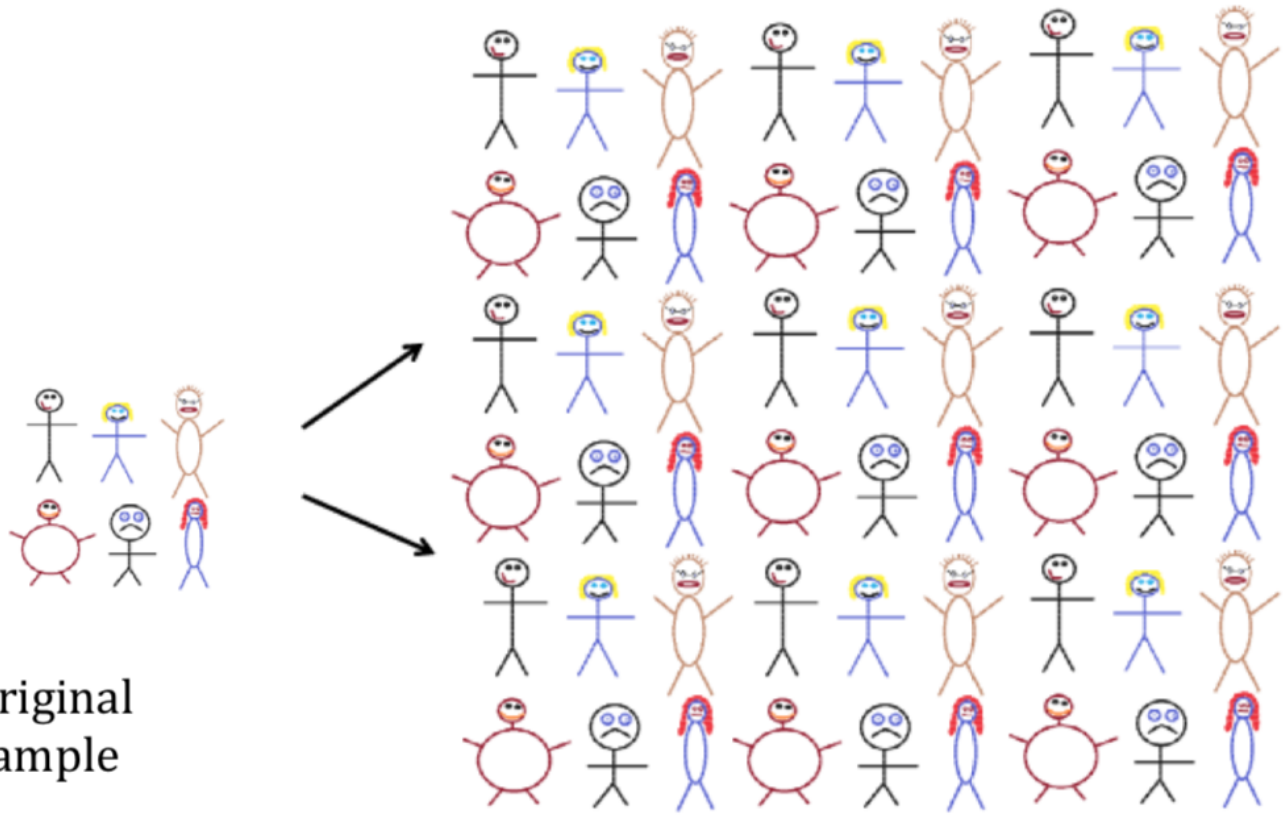
- Sample sizes might be "small"
- Assumptions we made in the population model might have been violated
- We might be interested in more complex population parameter

BOOTSTRAP!

- This term comes from the phrase "pulling oneself up by one's bootstraps", which is a metaphor for accomplishing an impossible task without any outside help.
- In this case the impossible task is estimating a population parameter, and we'll accomplish it using data from only the given sample.

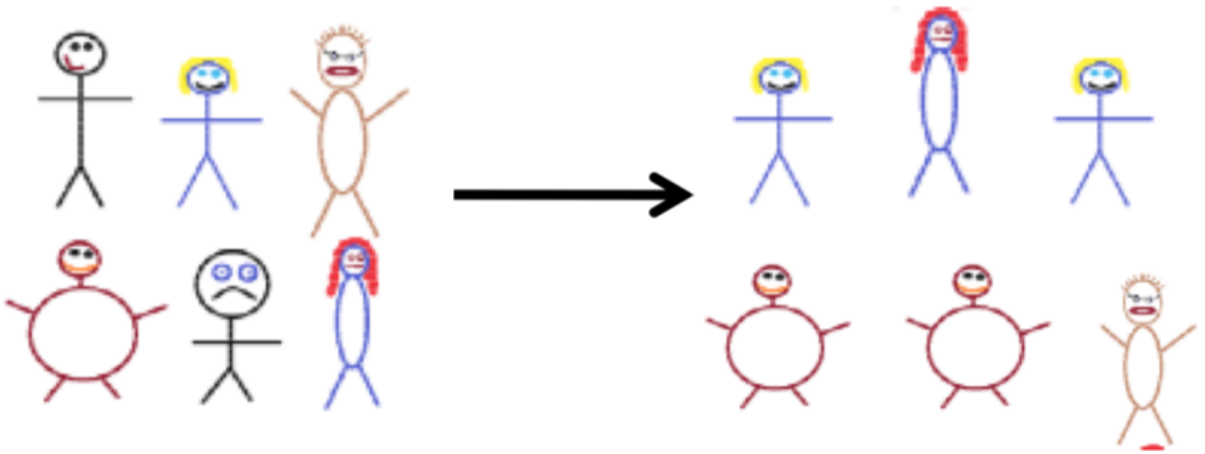
Suppose we have a random sample of 6 people

Sample with replacement from the sample we have (each unit can be selected more than once)



A simulated "population" to sample from

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



Original Sample

Bootstrap Sample

Your original sample has data values

$$18, 19, 19, 20, 21$$

Poll question

Is the following a possible bootstrap sample?

$$18, 19, 20, 21, 22$$

a. Yes
b. No

Your original sample has data values

$$18, 19, 19, 20, 21$$

Poll question

Is the following a possible bootstrap sample?

$$18, 19, 20, 21$$

a. Yes
b. No

Your original sample has data values

$$18, 19, 19, 20, 21$$

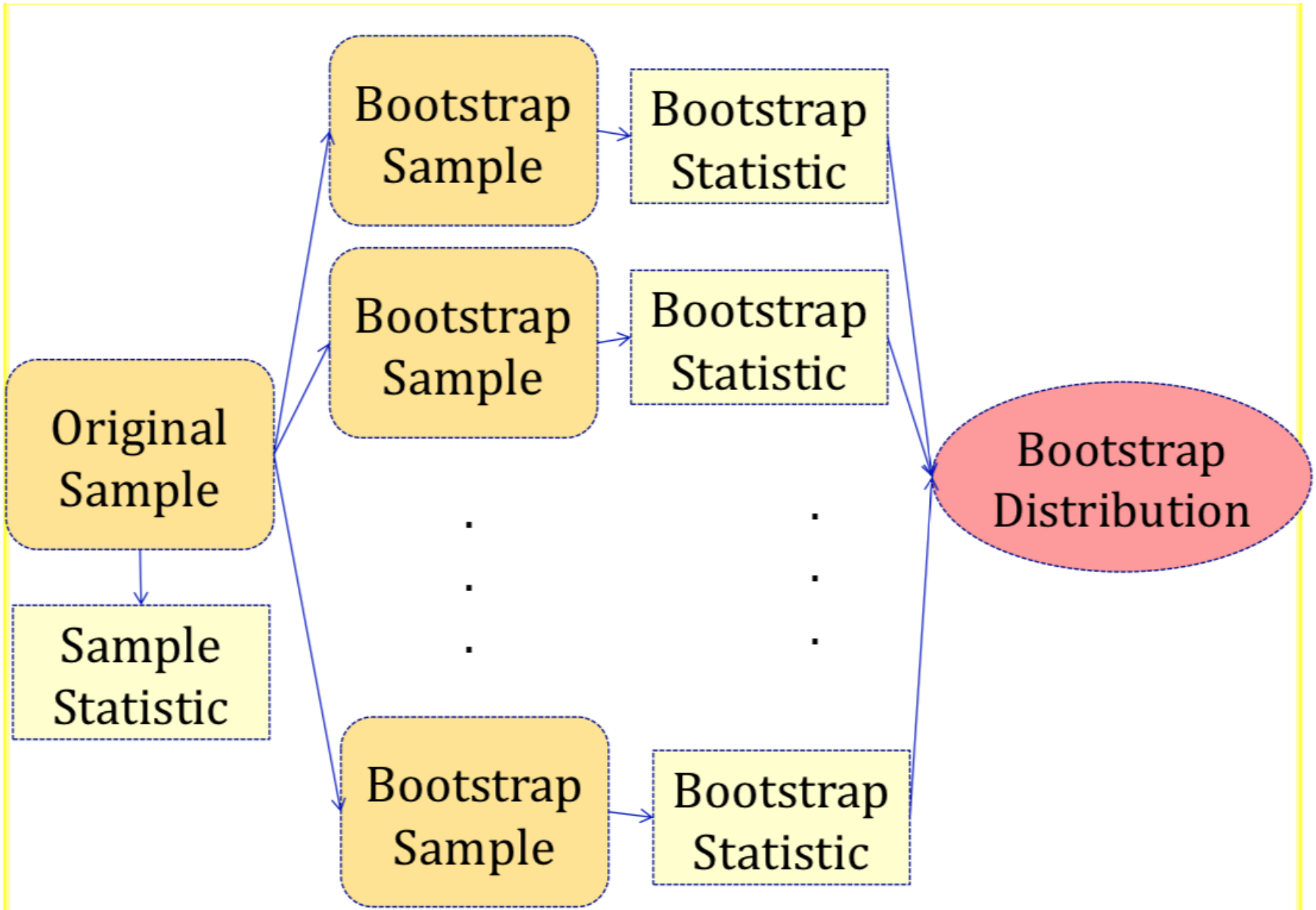
Poll question

Is the following a possible bootstrap sample?

$$18, 18, 19, 20, 21$$

a. Yes
b. No

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

You have a sample of size $n = 50$. You sample with replacement 1000 times to get 1000 bootstrap samples.

Poll question

What is the sample size of each bootstrap sample?

a. 50
b. 1000

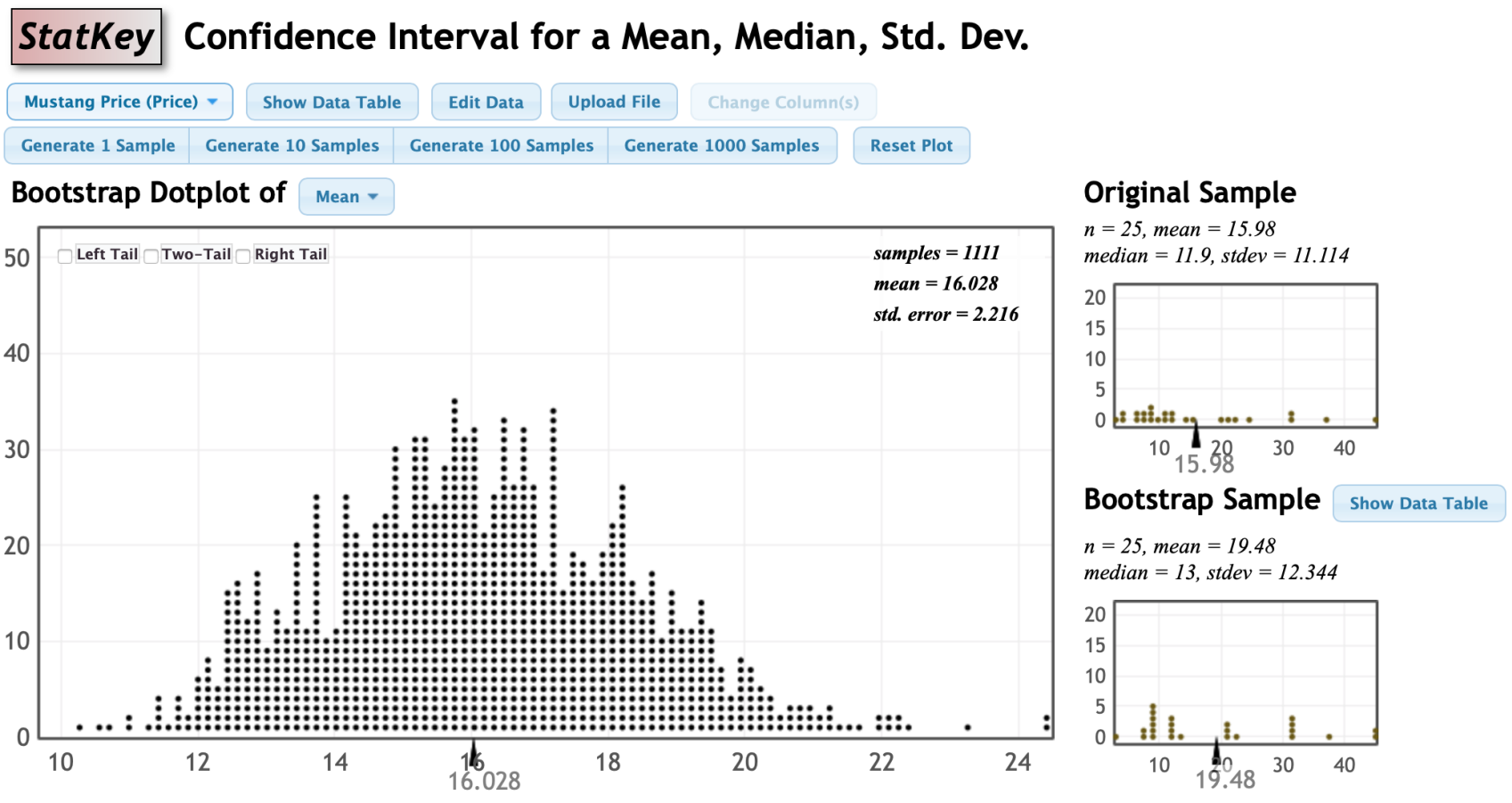
You have a sample of size $n = 50$. You sample with replacement 1000 times to get 1000 bootstrap samples.

Poll question

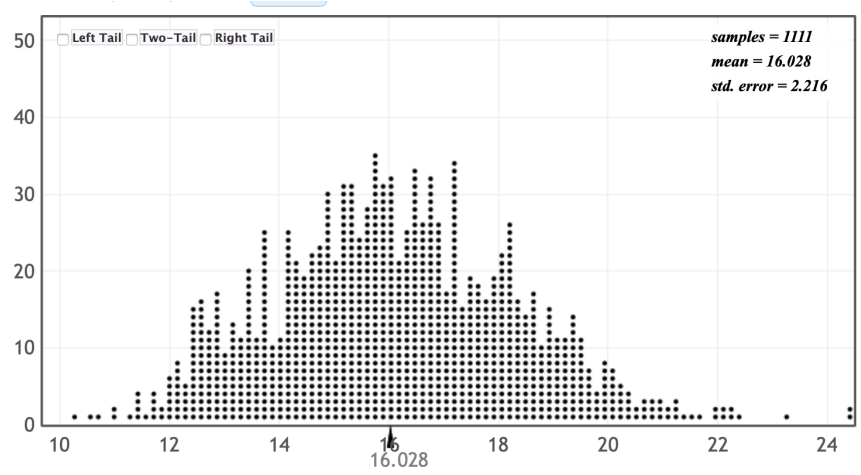
How many bootstrap statistics will you have?

a. 50
b. 1000

Example: Mustang price



The dot plot below is the bootstrap distribution of means constructed using 100 simulations.



Poll question

What does each dot on the dot plot represent?

a. Price of a Mustang
b. Mean price from one bootstrap sample
c. Mean price from one sample

We used **resampling** to mimic the **sampling variation**.

The bootstrap distribution is an *approximation* to the sampling distribution, in the sense that both distributions will have a similar shape and similar spread.

What about the the center and spread?

Poll question

- The sampling distribution is centered around the population parameter
- The bootstrap distribution is centered around the ?

a. population parameter
b. sample statistic
c. bootstrap statistic

Poll question

- The sampling distribution is centered around the population parameter
- The bootstrap distribution is centered around the ?

a. population parameter
b. sample statistic
c. bootstrap statistic

Luckily, we don't care about the center... we care about the **variability**!

- 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!

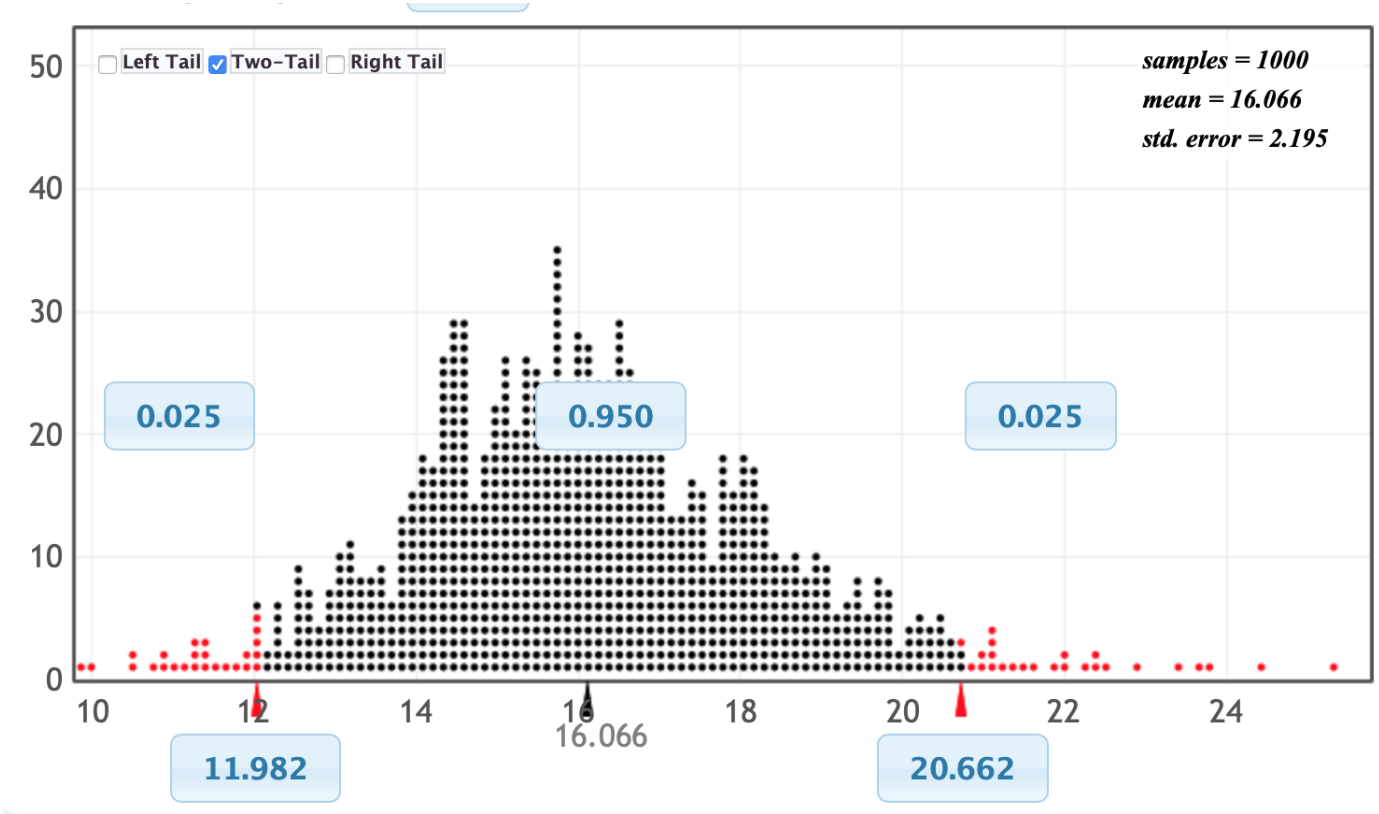
95% CI with the SE Method

Used Mustangs

$$\text{statistics} \pm 2 \times \text{SE}$$
$$\$15,980 \pm 2 \times \$2,178$$
$$(\$11,624, \$20,336)$$

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

95% CI with the Quartile Method

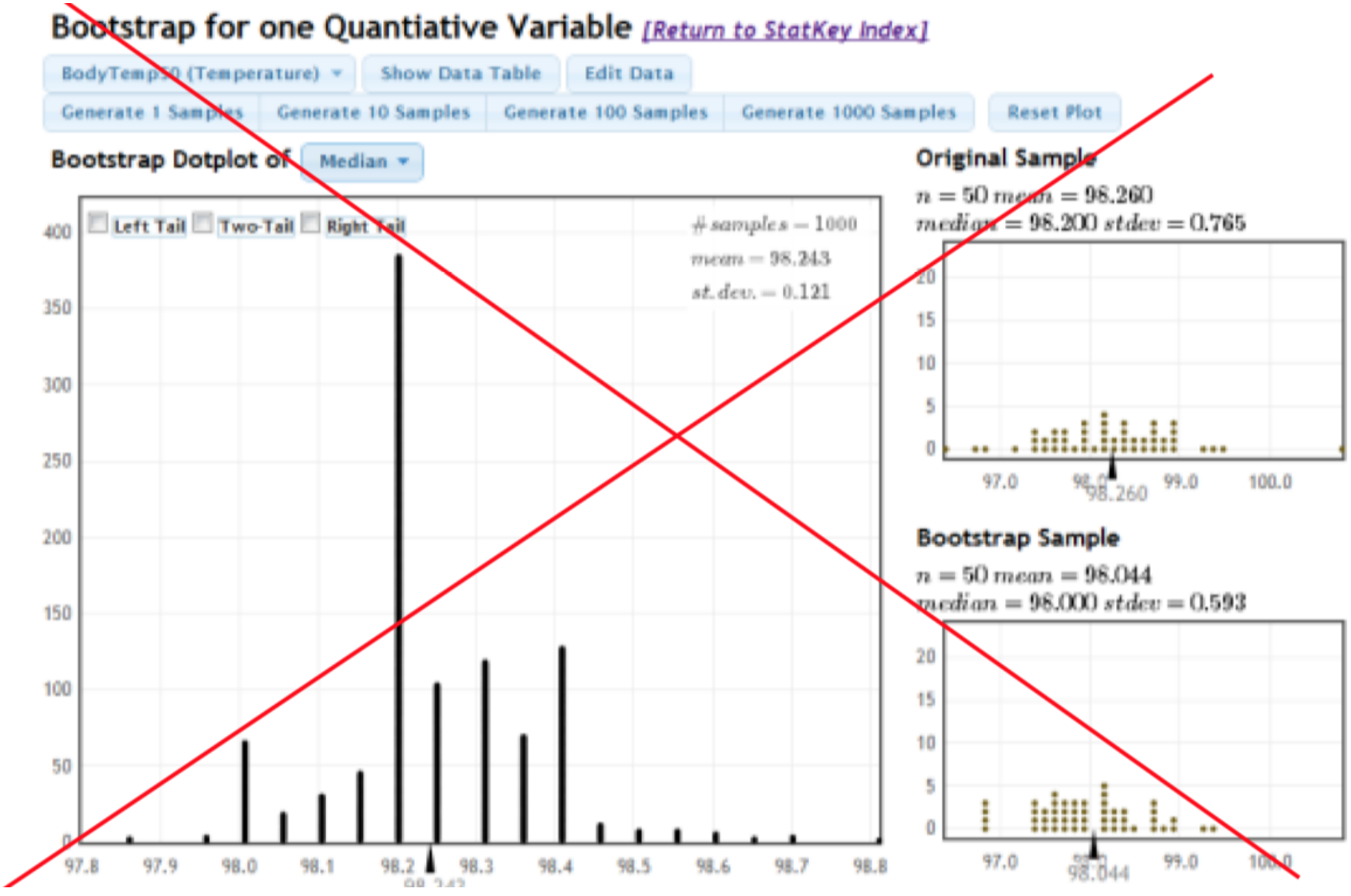


We are 95% confident that the average price of a used Mustang on autotrader.com is between \$11,982 and \$20,662.

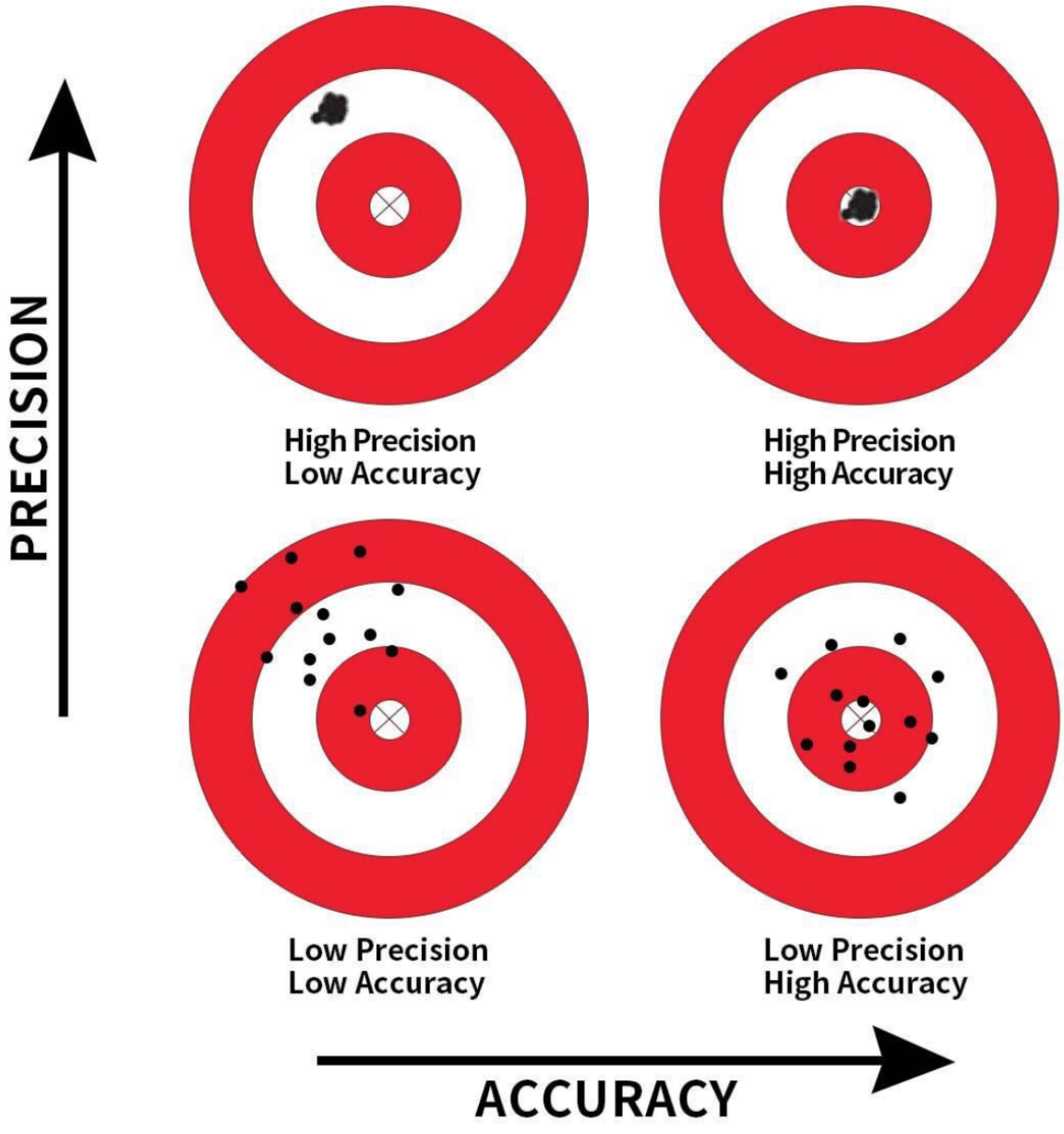
Which method should we use?

- For a symmetric, bell-shaped bootstrap distribution, using either the standard error method or the percentile method will given similar 95% confidence intervals
- If the bootstrap distribution is not bell-shaped or if a level of confidence other than 95% is desired, use the percentile method

- These methods for creating a confidence interval only work if the bootstrap distribution is smooth and symmetric
- ALWAYS look at a plot of the bootstrap distribution!
- If the bootstrap distribution is highly skewed or looks "spiky" with gaps, you will need to go beyond intro stat to create a confidence interval


Bootstrap for one Quantiative Variable [Return to StatKey Index]
BodyTemps50 (Temperature)
Show Data Table
Edit Data
Generate 1 Samples
Generate 10 Samples
Generate 100 Samples
Generate 1000 Samples
Reset Plot
Bootstrap Dotplot of
Median
Left Tail
Two-Tail
Right Tail
# samples = 1000
mean = 98.243
st. dev. = 0.121
Original Sample
n = 50 mean = 98.260
median = 98.200 stdev = 0.765
Bootstrap Sample
n = 50 mean = 98.044
median = 98.000 stdev = 0.593

- When using bootstrapping, you may get a slightly different confidence interval each time. This is fine!
- The more bootstrap samples you use, the more precise your answer will be
- Increasing the number of bootstrap samples will not change the SE or interval (except for random fluctuation)
- For the purposes of this class, 1000 bootstrap samples is fine. In real life, you probably want to take 10,000 or even 100,000 bootstrap samples


PRECISION
High Precision
Low Accuracy
High Precision
High Accuracy
Low Precision
Low Accuracy
Low Precision
High Accuracy
ACCURACY

1. Bootstrapping works as follows:
   - Take a bootstrap sample – a random sample taken with replacement from the original sample, of the same size as the original sample
   - Calculate the bootstrap statistic – a statistic such as mean, median, proportion, etc. computed on the bootstrap samples
   - Repeat the previous steps many times to create a bootstrap distribution – a distribution of bootstrap statistics
2. The XX% bootstrap confidence interval
   - can be estimated by the cutoff values for the middle XX% of the bootstrap distribution
   - can be estimated by sample statistic $\pm 2\times$ SE