# Confidence intervals, sampling distributions, and standard errors

# Overview

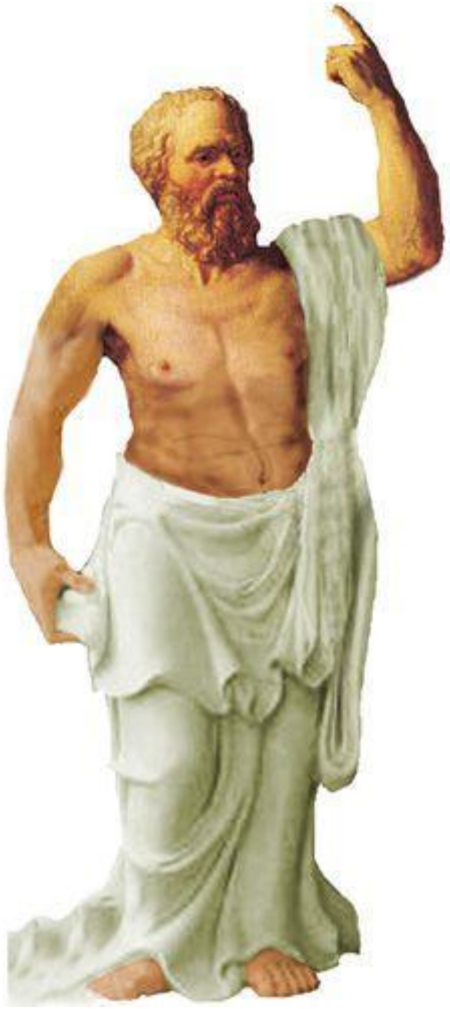
Review: confidence intervals and sampling distributions

The bootstrap

# Review of confidence intervals and sampling distributions

Question$_0$: Who is this?

- Socrates!



# 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

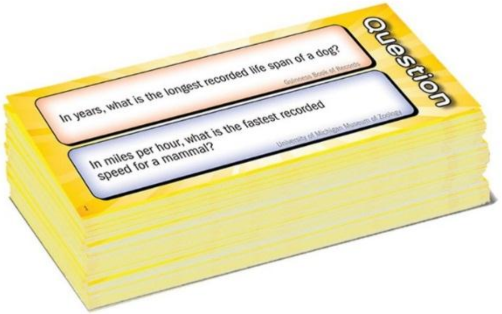

$Q_2$: What is the **confidence level**?

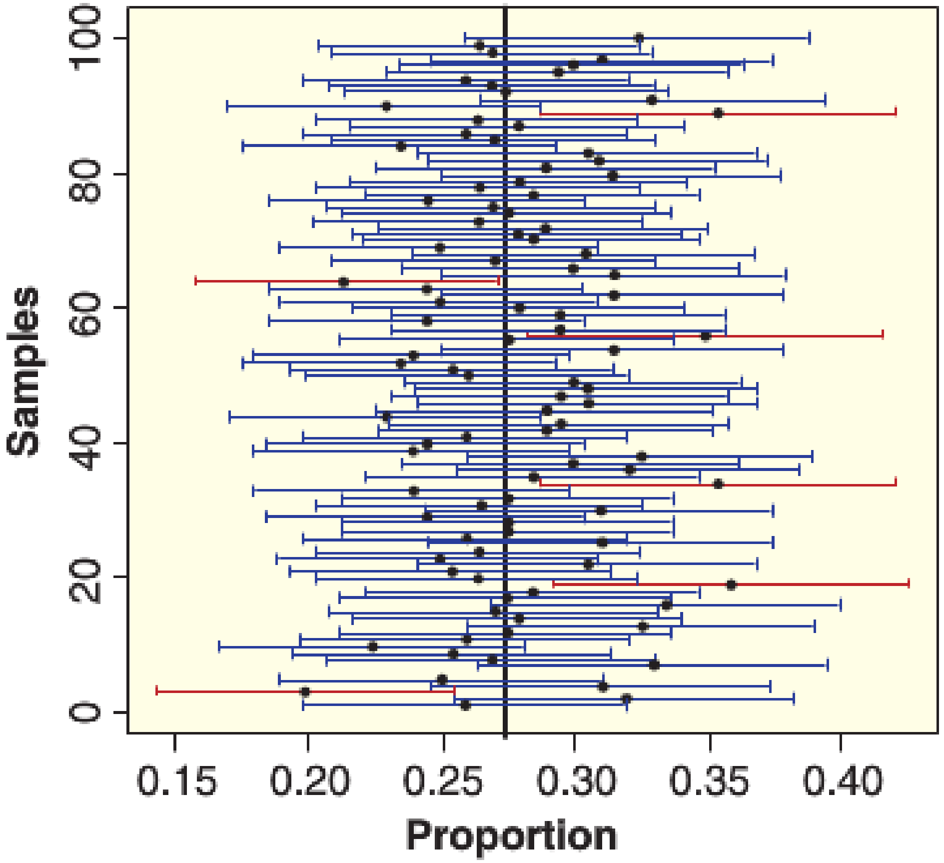
- $A_2$: The **confidence level** is the percent of all intervals that contain the parameter

# Confidence Intervals

$Q_3$: For a **confidence level** of 90%, how many of these intervals should have the parameter in them?

- A: 90%




$Q_4$: For a given confidence interval, do we know if it contains the parameter?

- A: No! ☹

Q$_5$: For the cartoon below, what is the confidence level the weatherman is using?

- A: 100%



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

# Example

130 observations of body temperature of men were made

A 95% confidence interval for the body temperatures is:
[98.123, 98.375]

How do we interpret these results?

Is this what you would expect?

[Original paper](#) [Statistic teaching resource](#)

# Confident intervals

$Q_6$: Are we feeling confident about confidence intervals?

# 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?

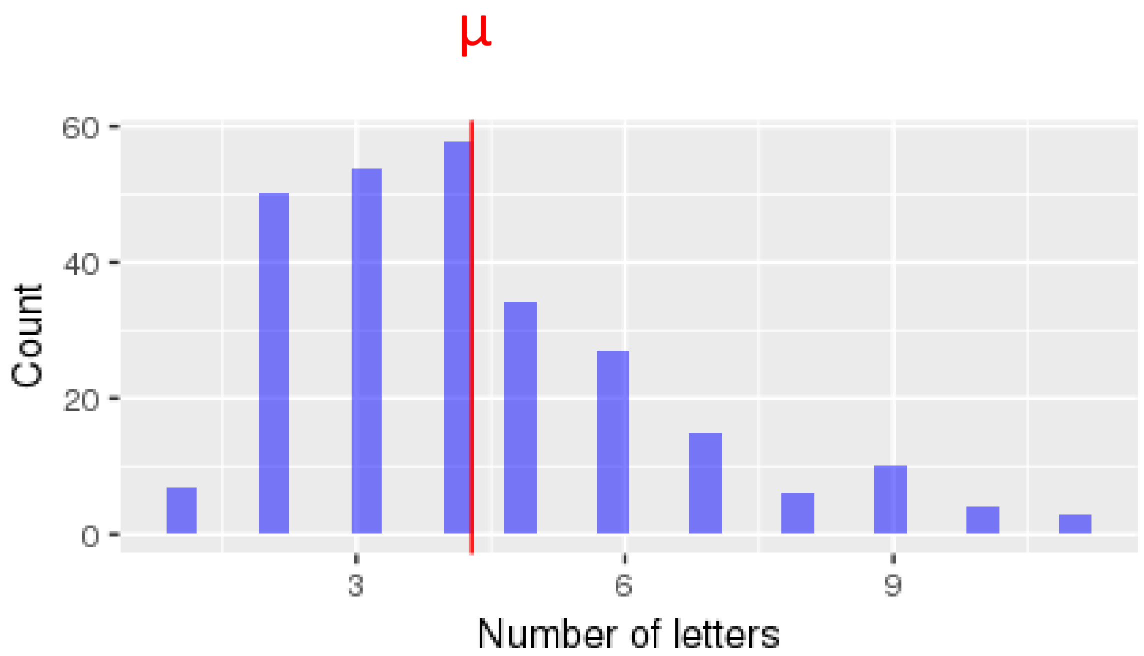
- A: A sampling distribution shows us how the sample statistic varies from sample to sample

# 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

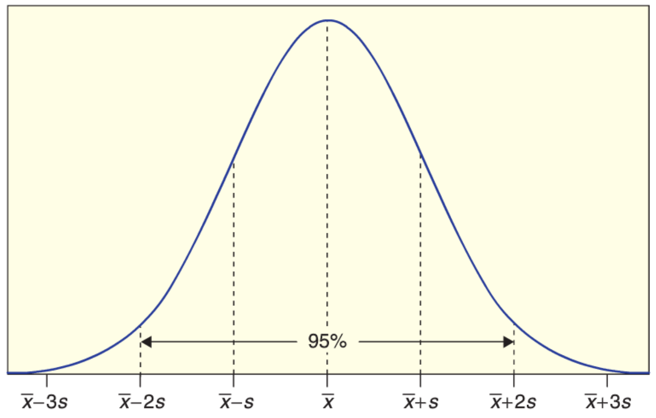
$\bar{x} = 5$

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]

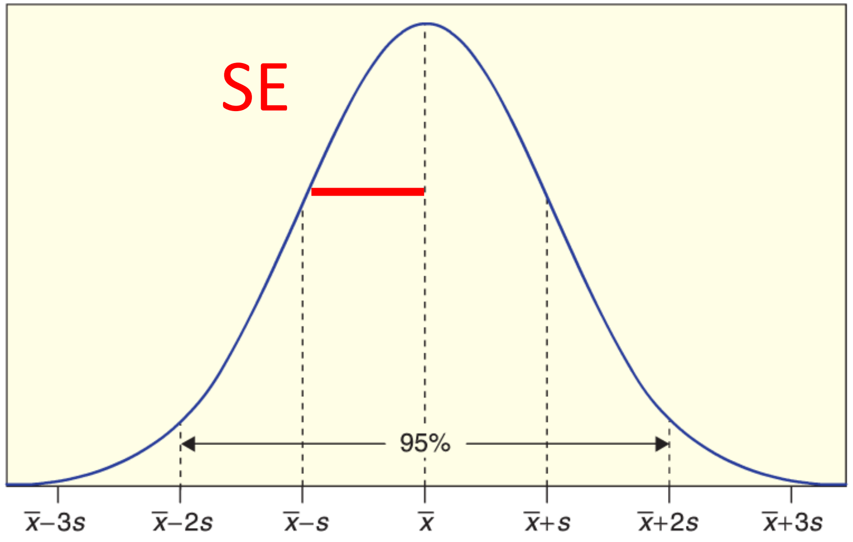
# The standard error

$Q_9$: What is the **standard error**?

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

$Q_{10}$: What symbol do we use to denote the standard error?

- SE

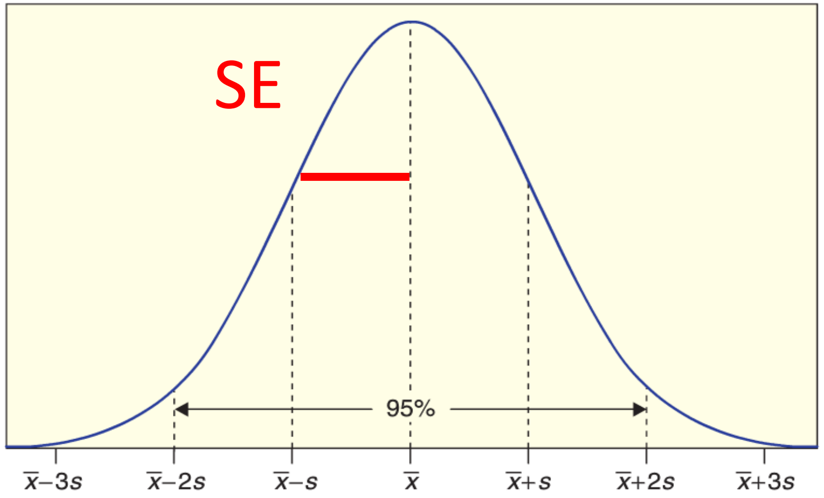


# Sampling distribution in R

$Q_{11}$: If we could easily sample infinitely many times from our population, how could we calculate the SE of the mean using R?

```
sampling_distribution_vec <- NULL

for (i in 1:100000) {

        curr_sample <- sample(population_vector, 10)
        sampling_distribution_vec[i] <-  mean(curr_sample)

}

SE_mean <- sd(sampling_distribution_vec)
```

# The standard error




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

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

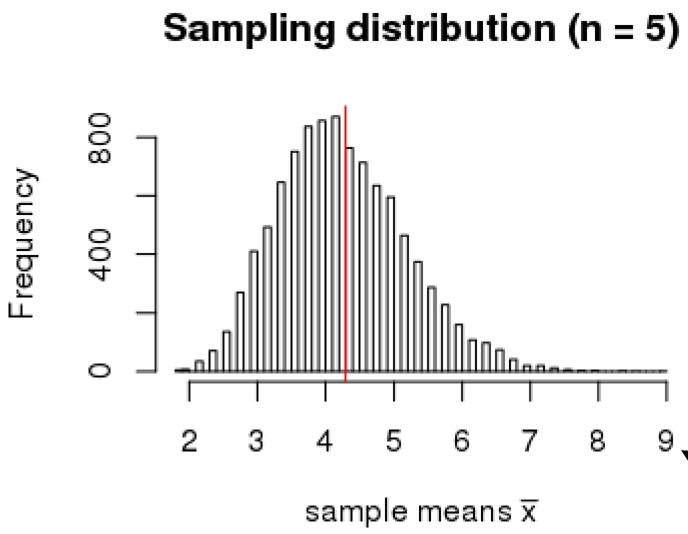
$Q_{13}$: What would be mean if there is a large SE?

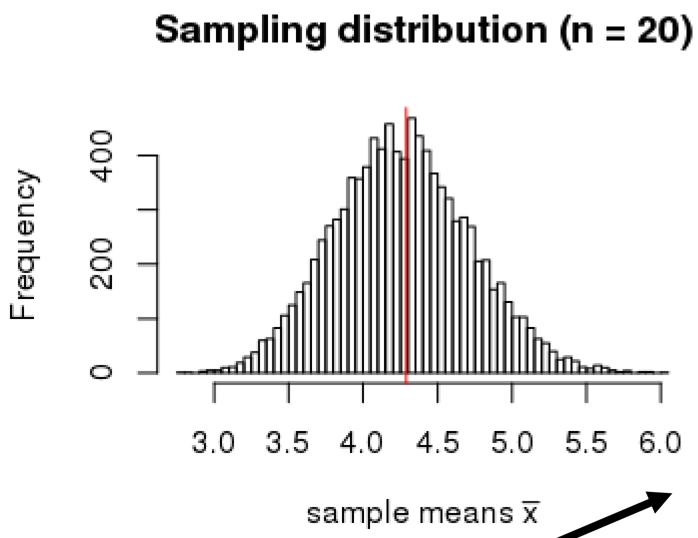
- A large SE means our statistic (point estimate) could be far from the parameter
- E.g., $\overline{x}$ could be far from $\mu$

# $Q_{14}$: How does the sampling distribution change with larger sample size n?

A: 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




x-axis range 9 vs. 6

# Shapes of sampling distributions

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

A: Normal!

# Normal distributions

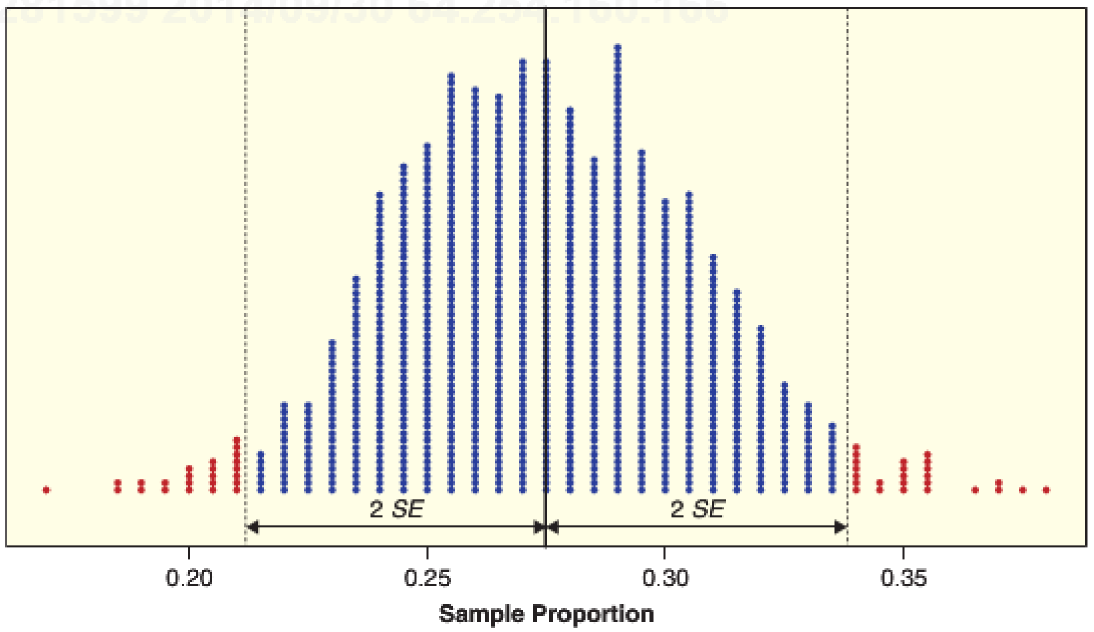
$Q_{15}$: For a normal distribution, what percentage of points lie within 2 standard deviations for the population mean?

A: 95%

# 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! (assuming the sampling distribution is normal, which it often is)

# Sampling distributions

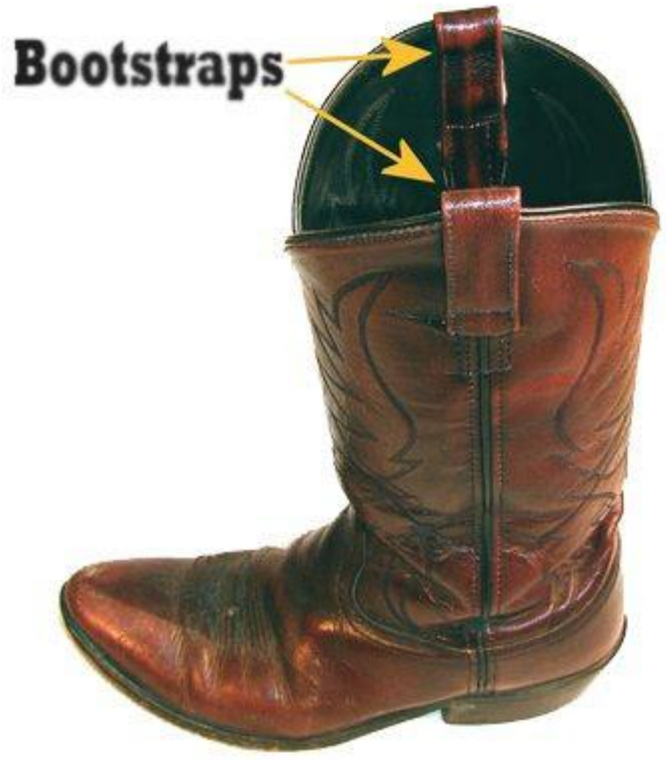
$Q_{18}$: Could we repeat the sampling process many times to create a sampling distribution and then calculate the SE?

- 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*

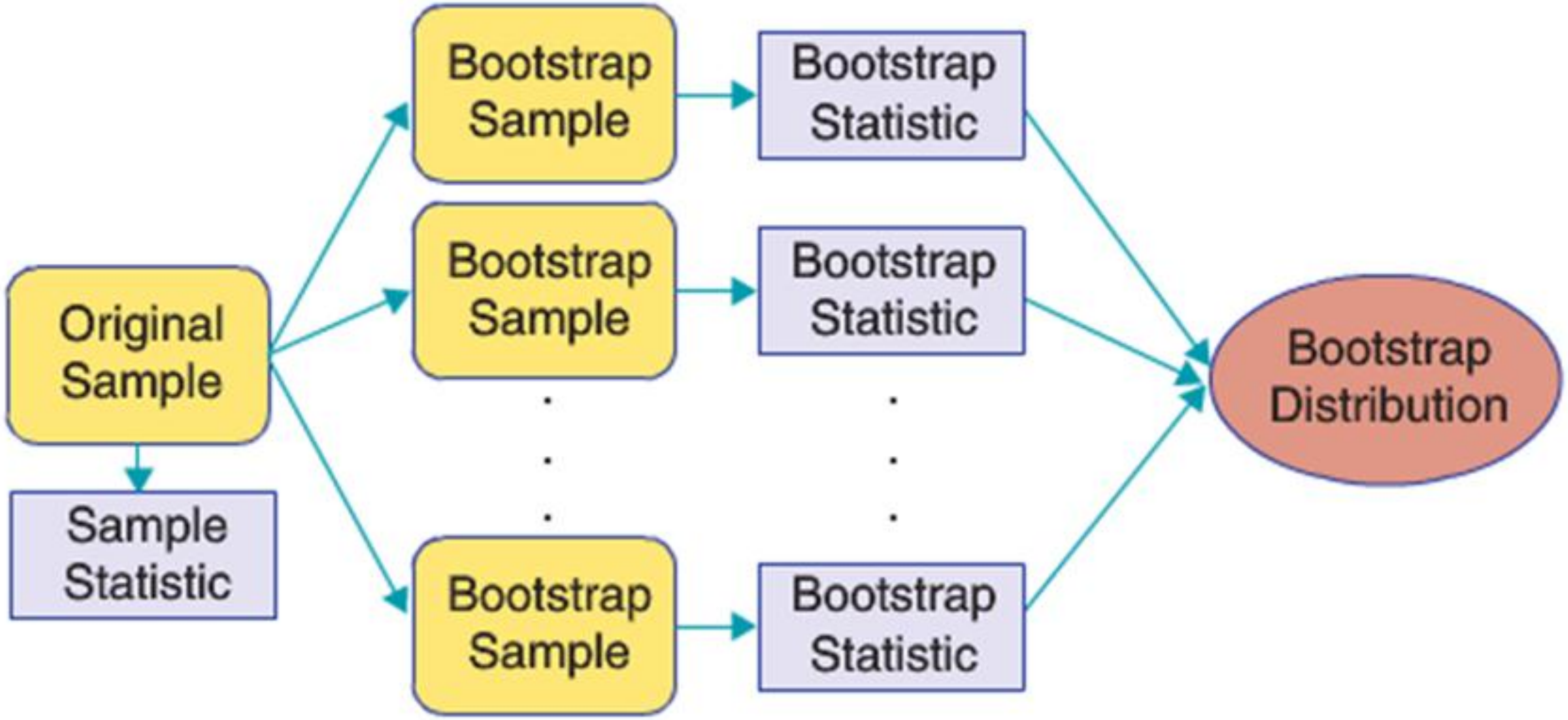
We pretend that this sample is the population (plug-in principle)

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

# Worksheet 6

Due midnight on Sunday Oct 21st

```
source('/home/shared/intro_stats/cs206_functions.R')
> get_worksheet(6)
```