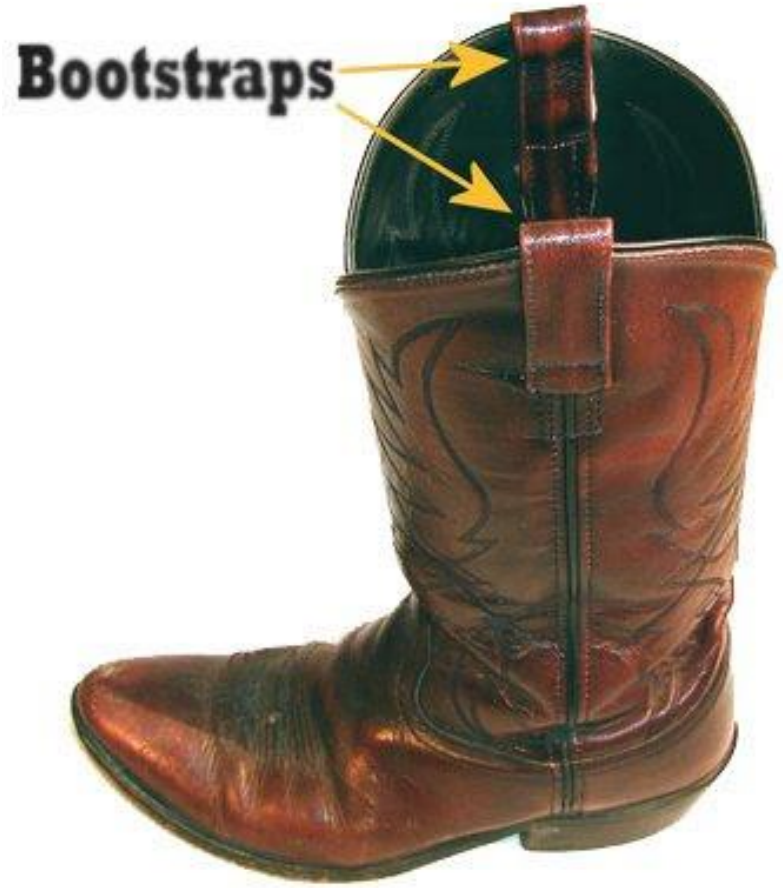
S&DS 101
Intro Statistics: Life Sciences

Overview

The bootstrap

If there is time: theories of hypothesis testing

The bootstrap for calculating confidence intervals



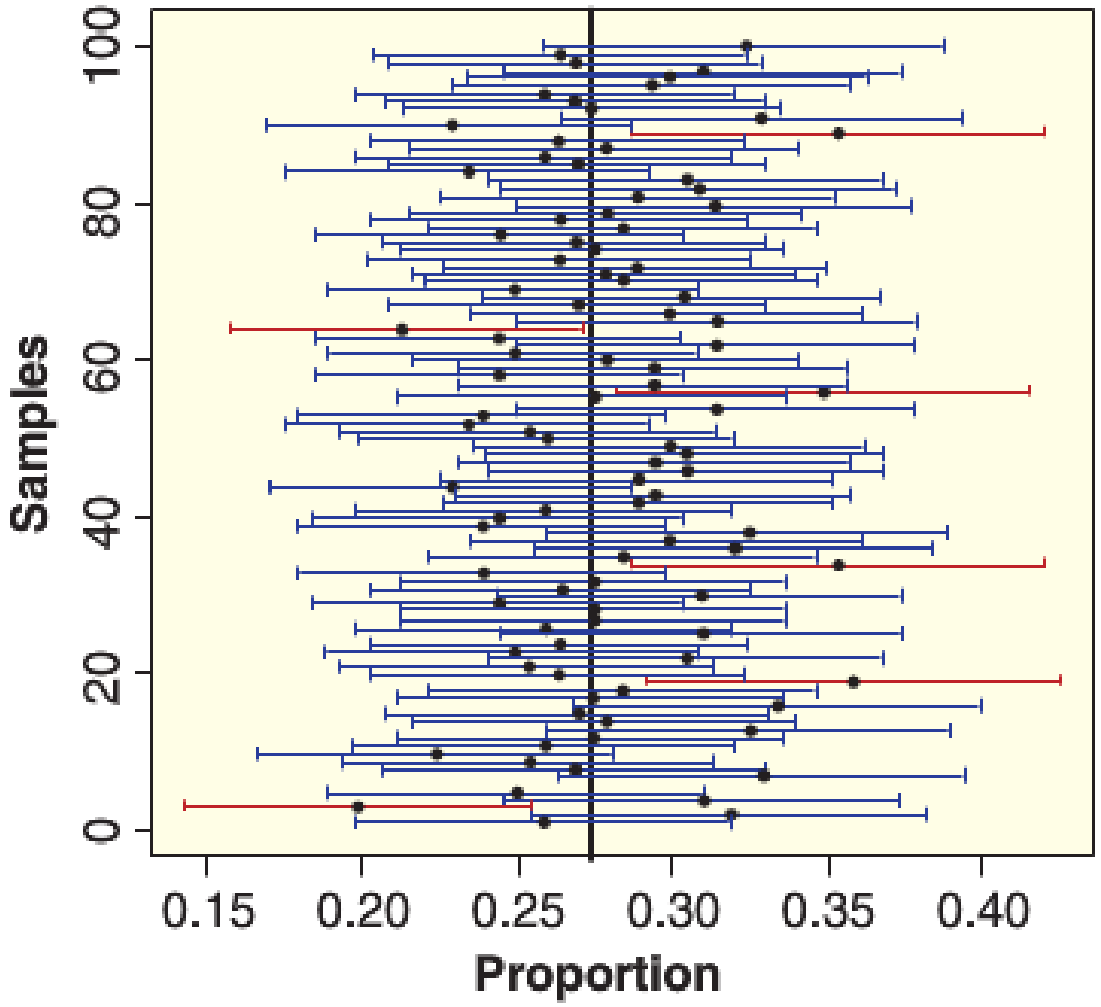
Confidence Intervals

For a **confidence level** of 95%...

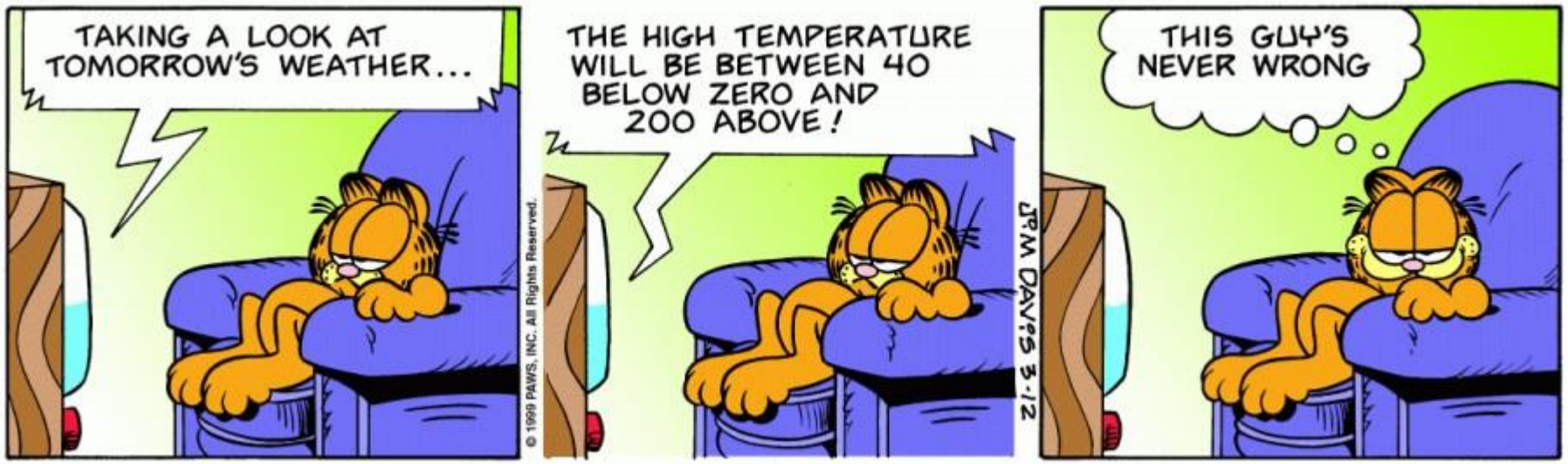
95% of the **confidence intervals** will have the ***parameter*** in them

Common form of a confidence interval:

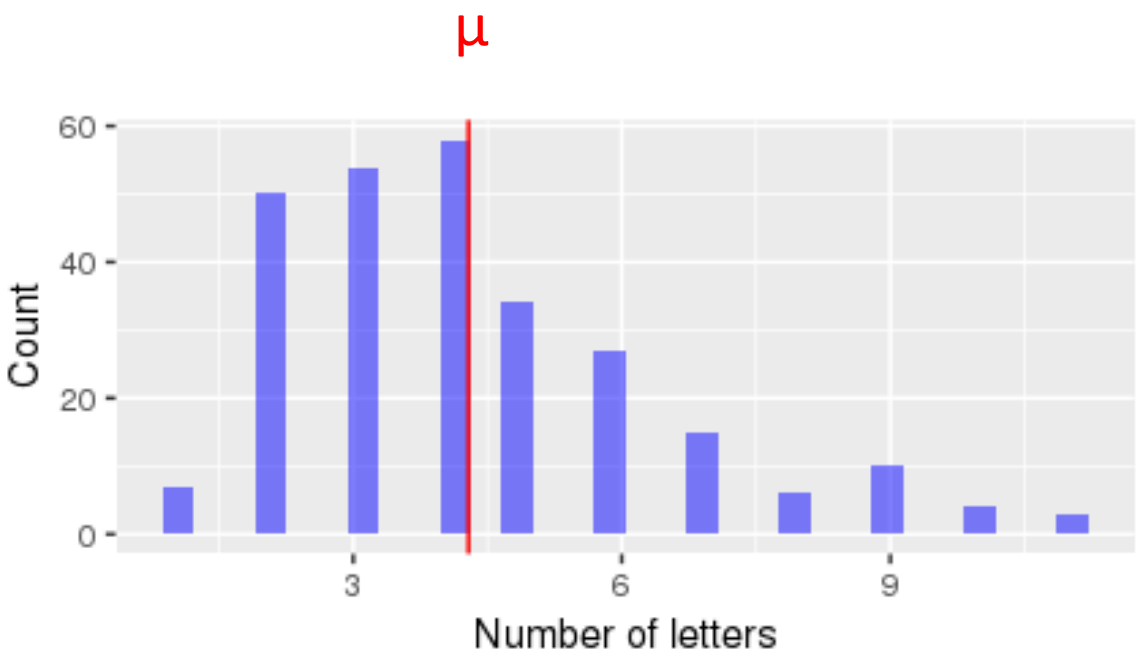
statistic ± q* · $\hat{SE}$



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



Review: sampling distribution illustration



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

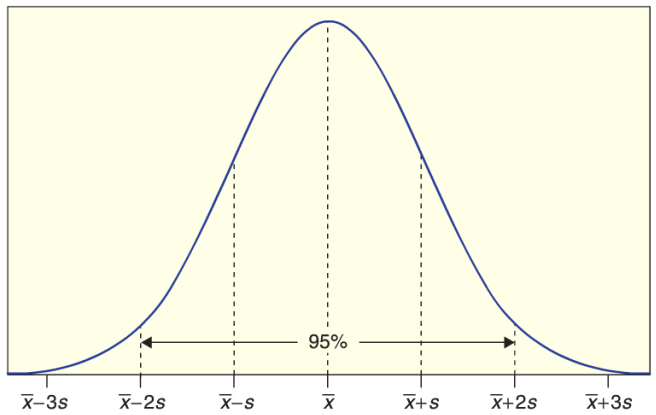
$\bar{x} = 5$

2, 6, 2, 6, 6,
2, 5, 3, 2, 9

$\bar{x} = 4.3$

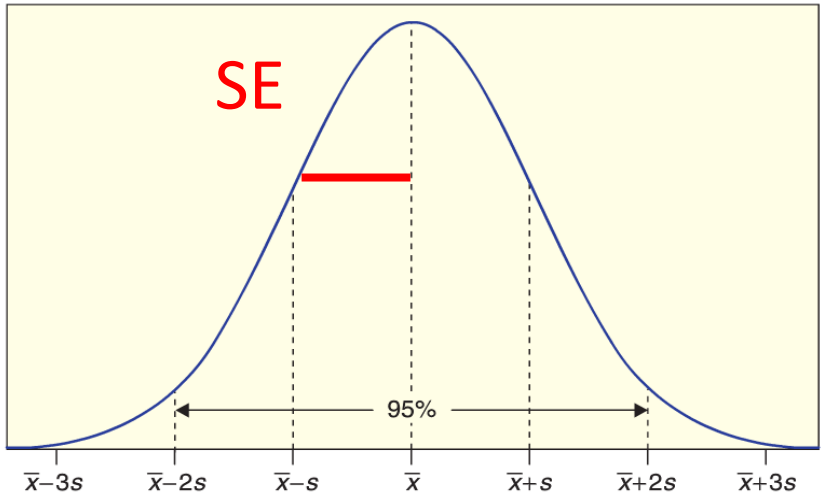
3, 9, 3, 4, 4,
3, 6, 6, 2, 2

$\bar{x} = 4.2$



Sampling distribution!

The standard error



Q: What does the size of the standard error tell us?

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

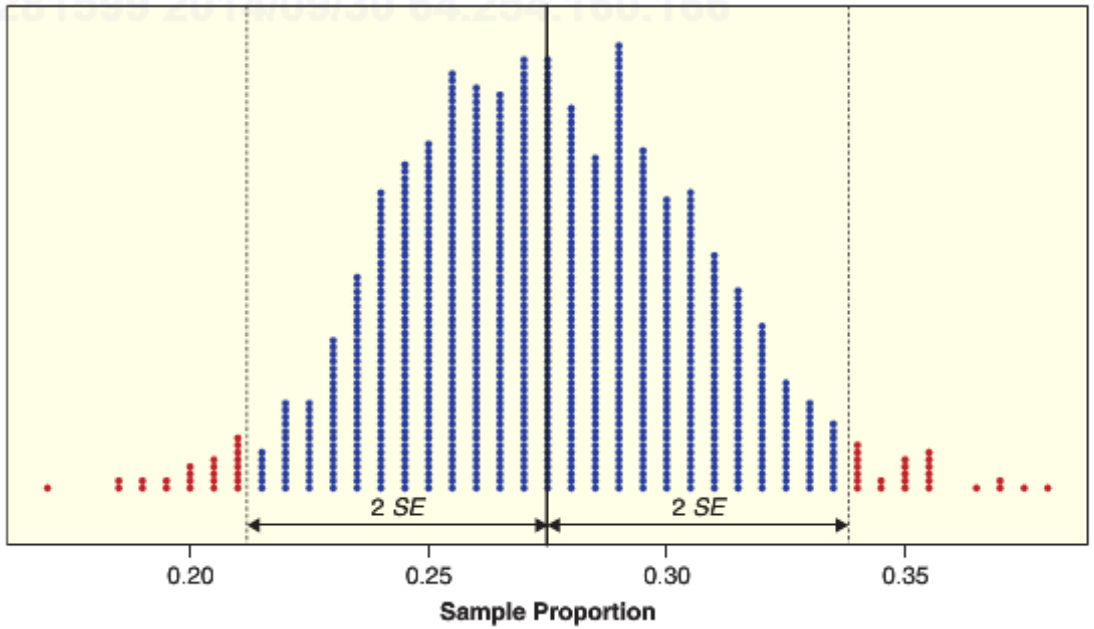
Q: What would be mean if there is a large SE?

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

Sampling distributions

Q: 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: If we had a statistic 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). CI = stat +/- q* SE

Formulas for the standard error of the mean

As you learned in intro statistics class, there is formula the **standard error of the mean (SEM)** which is:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad\qquad s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

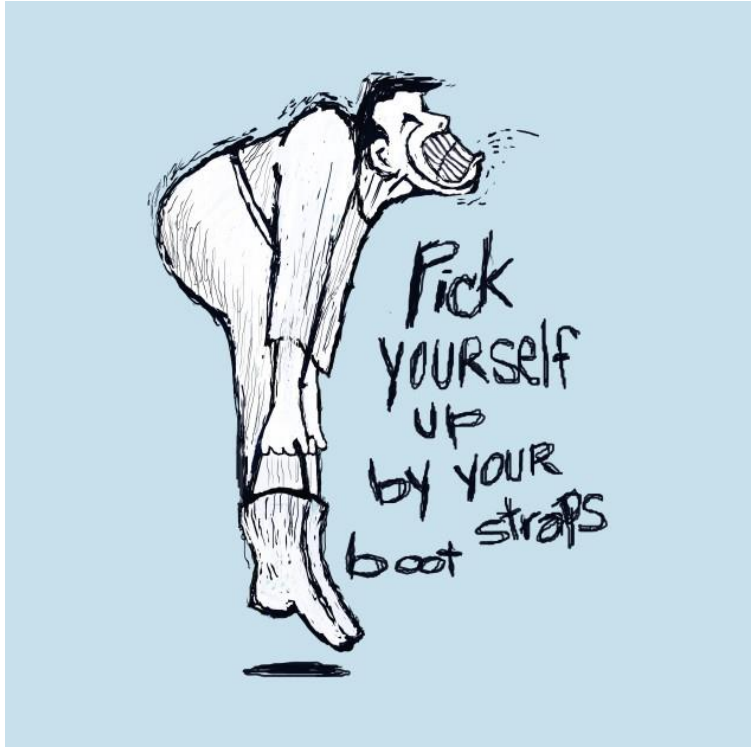
Where:

- σ is population standard deviation parameter
- n is the sample size
- s is the sample standard deviation

Sampling distributions

If we can't calculate the the standard error (SE) from a formula, we can use the bootstrap to get an estimate of the (SE)

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

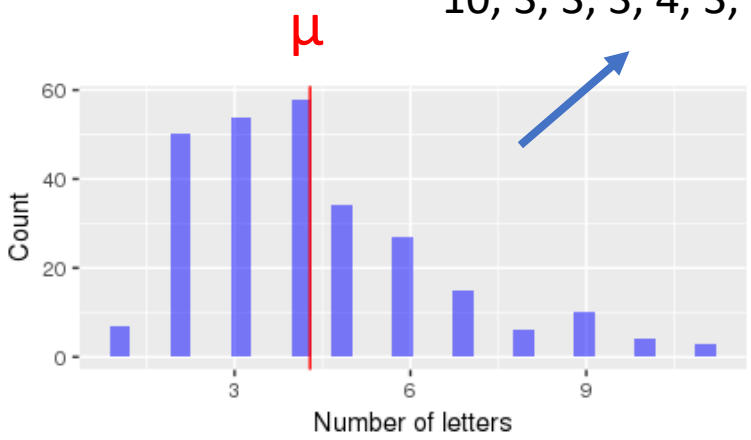
2. We repeat this process 1000's of times and get a *bootstrap* sample distribution

3. The standard deviation of this bootstrap distribution (SE* bootstrap) is a good approximate for standard error SE from the real sampling distribution

Gettysburg address word length bootstrap distribution

The sample (n = 10)

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



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

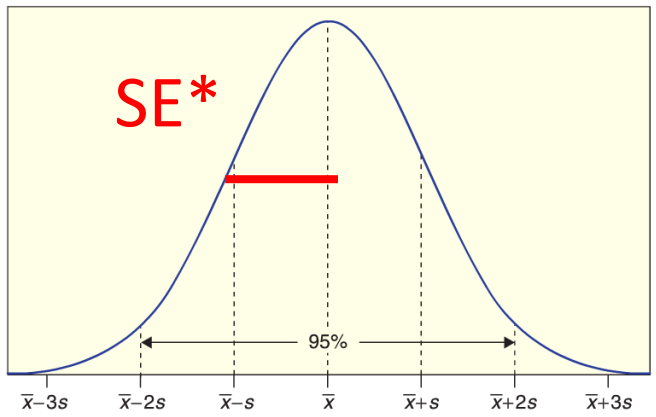
$\bar{x}^* = 4$

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

$\bar{x}^* = 4.1$

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

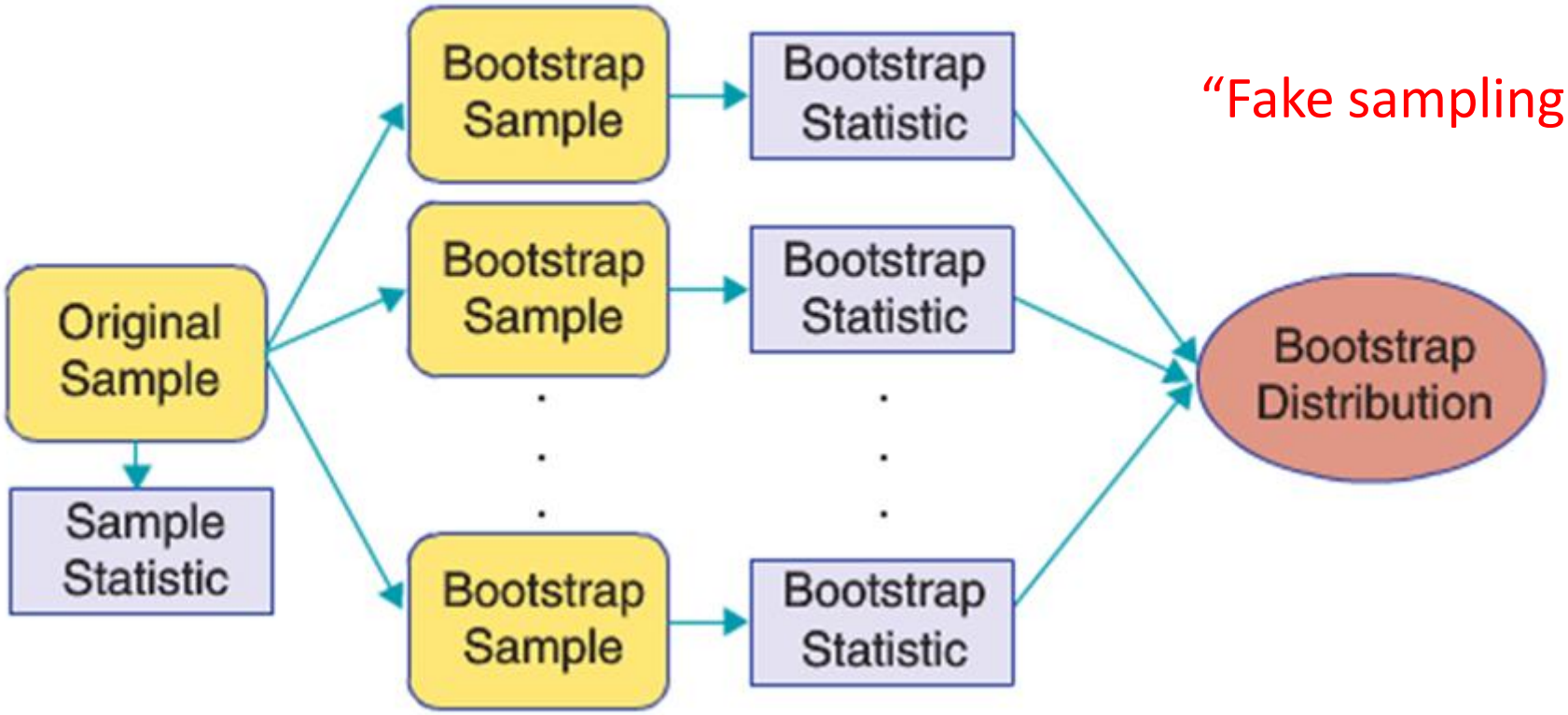
$\bar{x}^* = 3.9$



Bootstrap distribution!

Notice there is no 9's in the bootstrap samples

Bootstrap process



"Fake sampling distribution"

95% Confidence Intervals

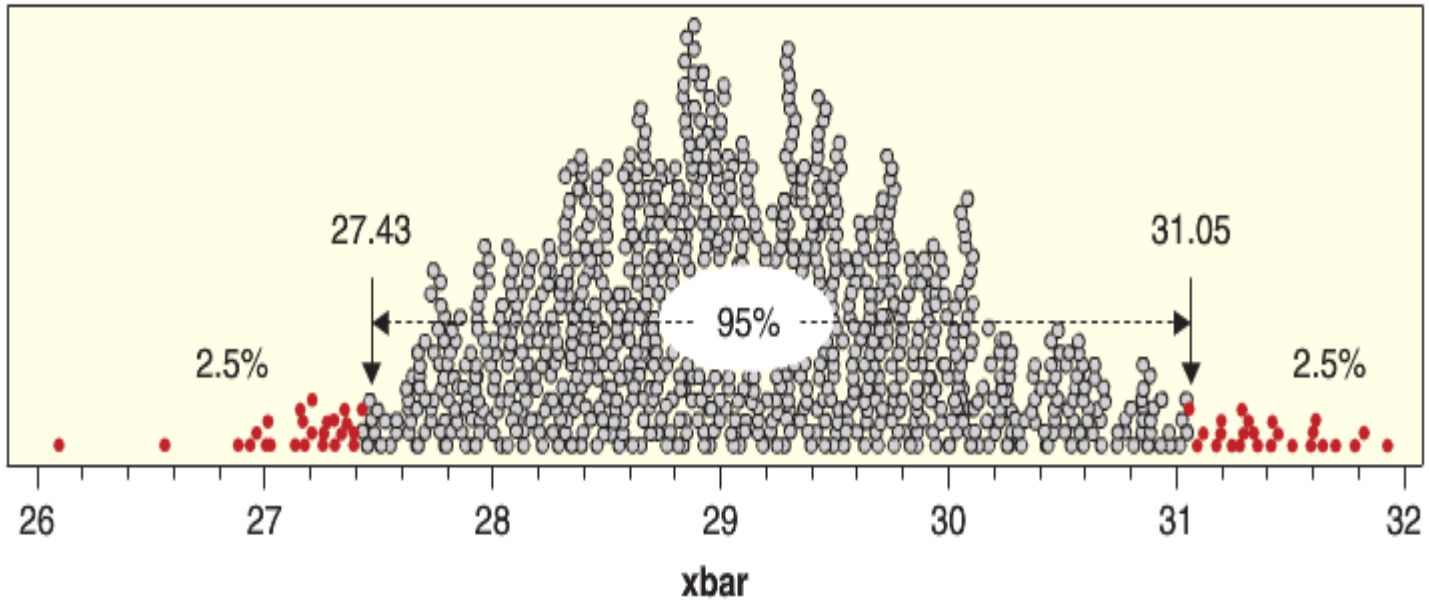
When a bootstrap distribution for a sample statistic is approximately normal, we can estimate a 95% confidence interval using:

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

Where SE* is the standard error estimated using the bootstrap

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!

Let’s try it in R…