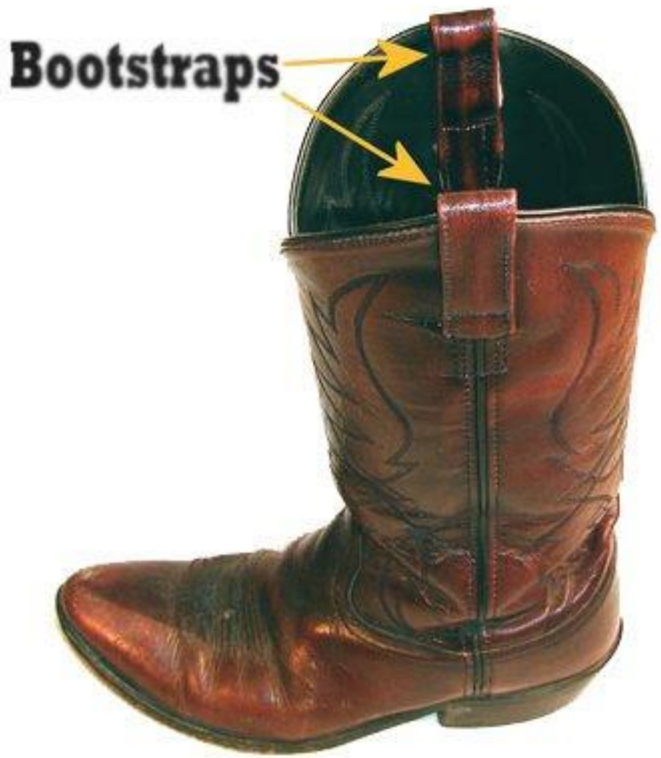
The bootstrap



Overview

Review: confidence intervals and sampling distributions

The bootstrap with code

Using a web app to better understand confidence intervals, sampling and bootstrap distributions

Confidence Intervals

Q: What is a **confidence interval**?

- A: a **confidence interval** is an interval computed by a method that will contain the *parameter* a specified percent of times

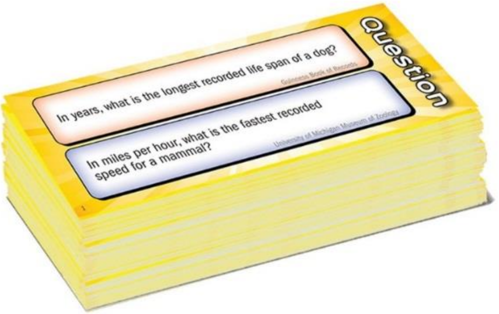
Q: What is the **confidence level**?

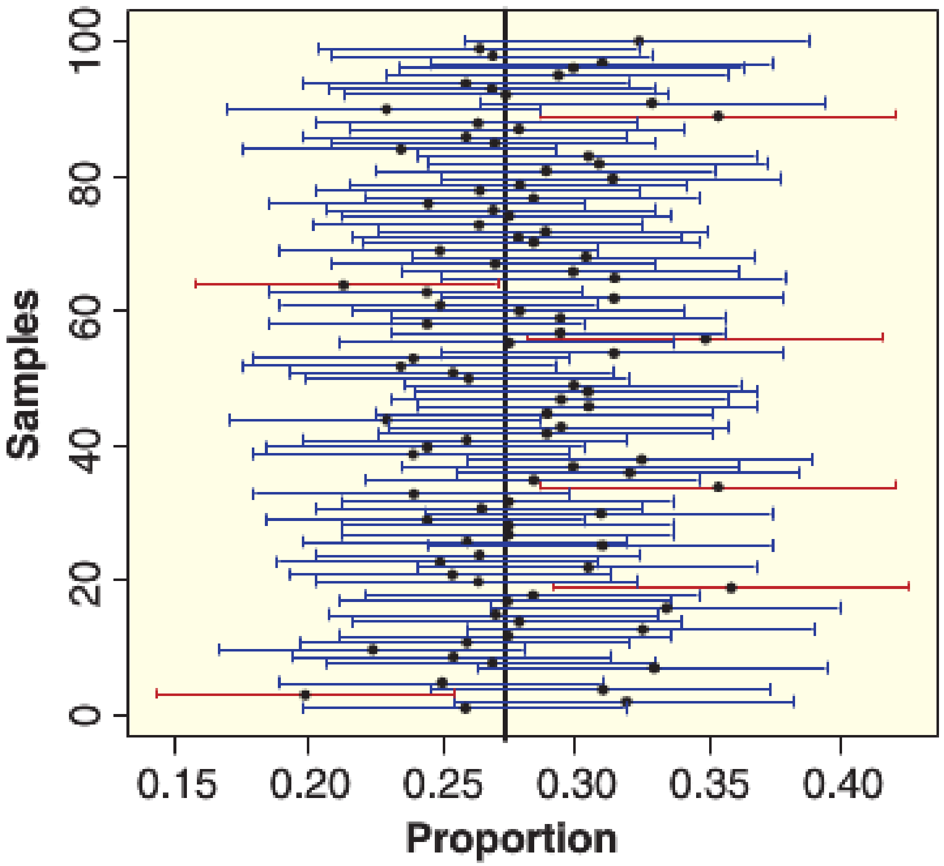
- A: The **confidence level** is the percent of all intervals that contain the parameter

Confidence Intervals

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

- A: 90%



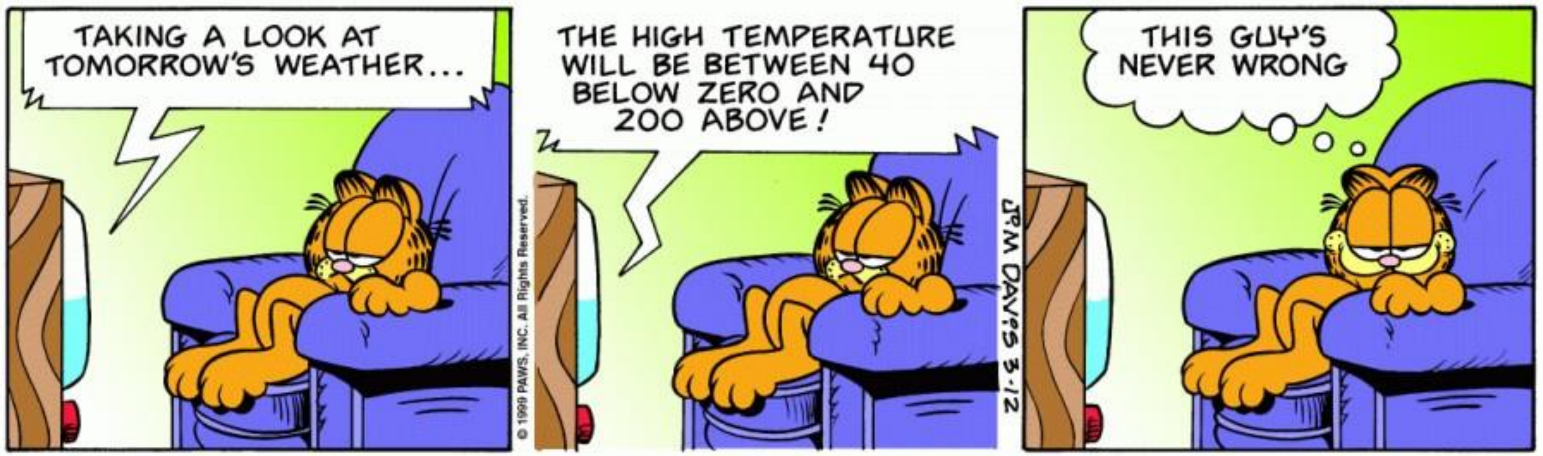


Q: For a given confidence interval, do we know if it contains the parameter?

- A: No! ☹

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

- A: 100%



There is a tradeoff between:

- The **confidence level** (percent of times we capture the parameter)
- The **confidence interval size**

Example

130 observations of body temperature of men were made

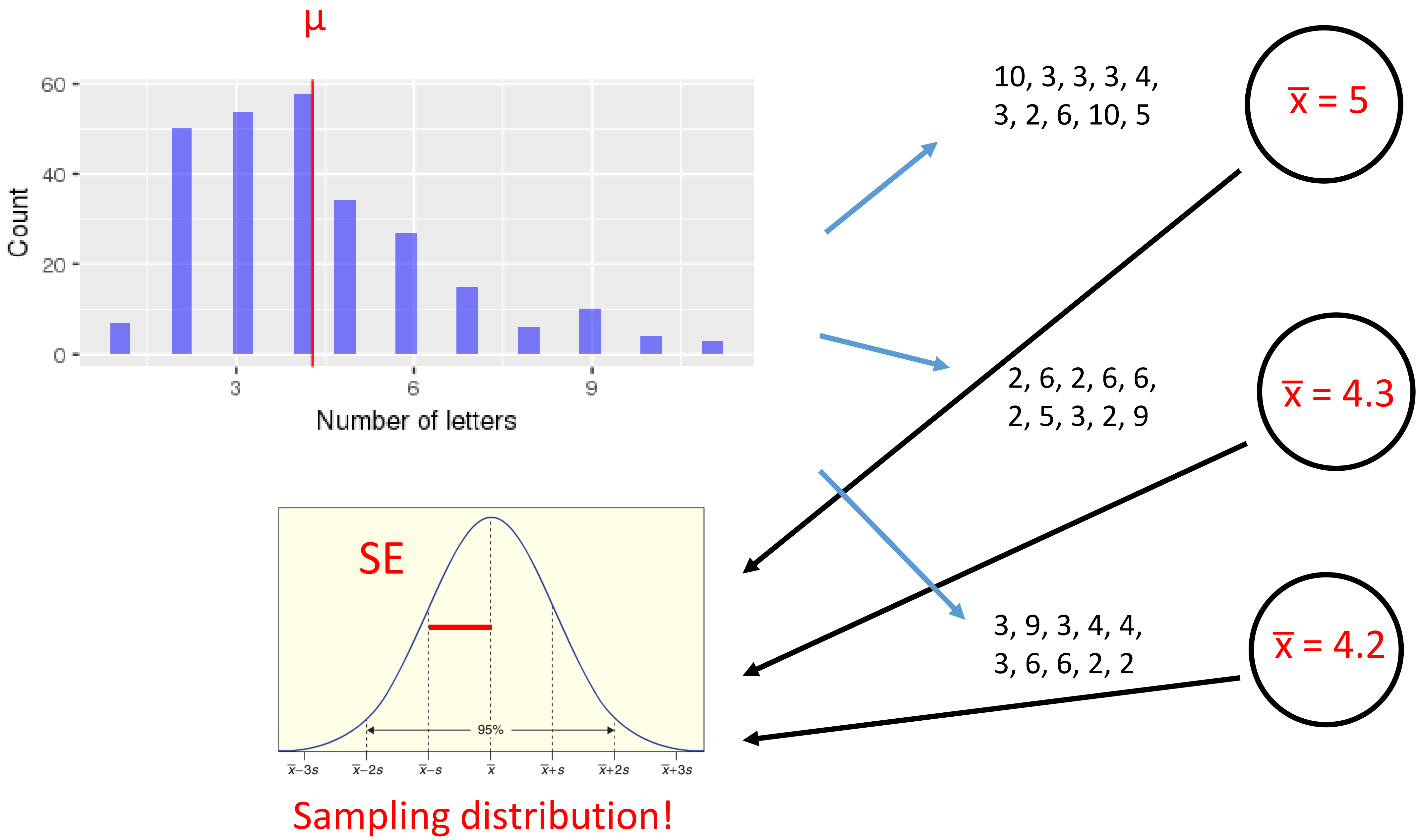
A 95% confidence interval for the body temperatures is:
[98.123, 98.375]

How do we interpret these results?

Is this what you would expect?

Original paper

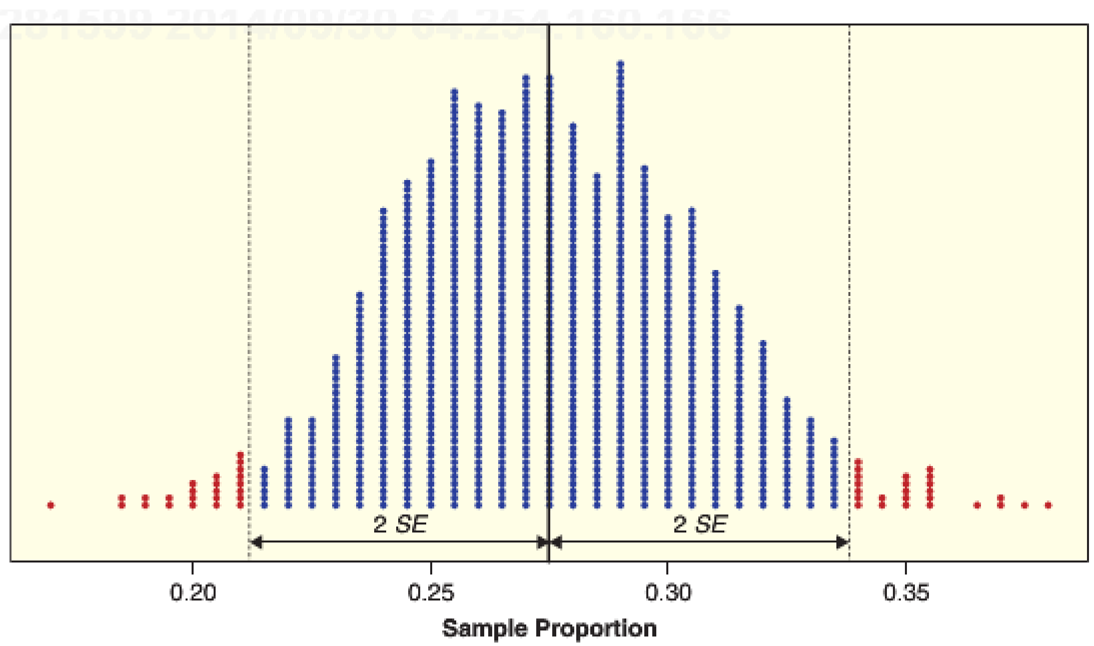
Review: sampling distribution illustration



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%



If we had:

- A statistics value
- The SE

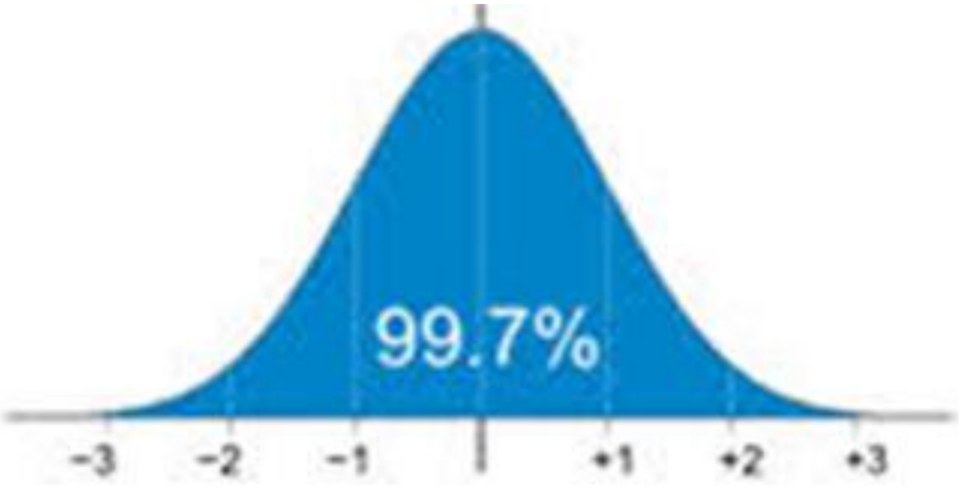
We could compute a 95% confidence interval!

$$CI_{95} = \bar{x} \pm 2 \cdot SE$$

Confidence intervals for other confidence levels

Q: How could we get a 99.7% confidence interval confidence level?

A: For normally distributed data, 99.7% of our data lie within 3 standard deviations of the mean



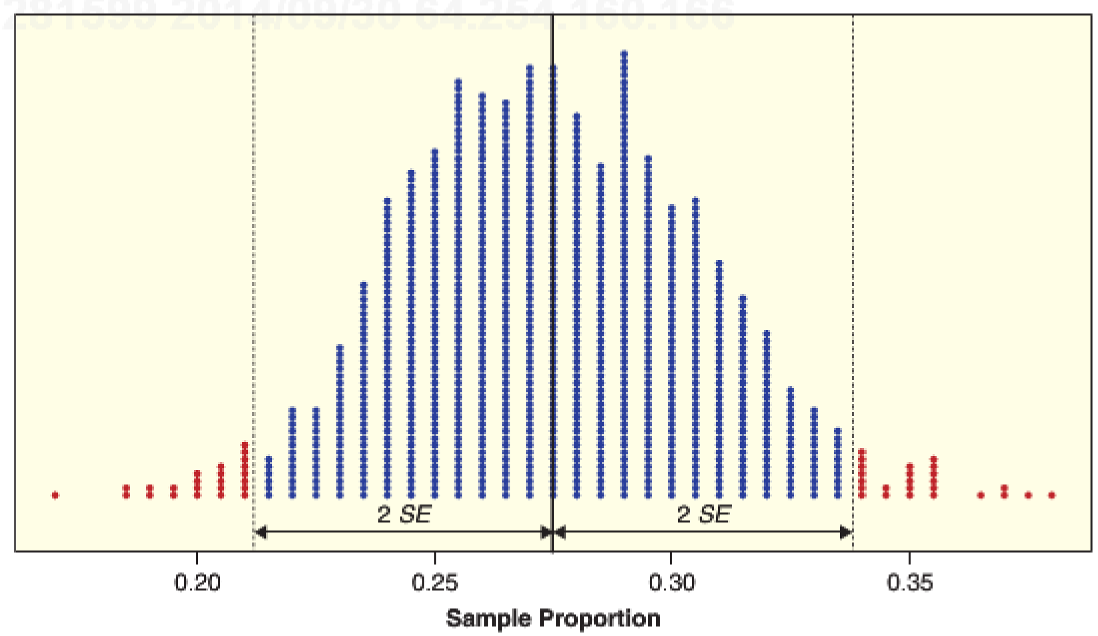
$$CI_{99.7} = \bar{x} \pm 3 \cdot SE$$

$$CI_{68} = \bar{x} \pm 1 \cdot SE$$

Confidence intervals for other confidence levels

Q: How could we get a confidence interval for the qth confidence level?

A: We need to find the critical value q* such that q% of our statistics are within $\pm q^* \cdot SE$ for a normal distribution



$$CI = \bar{x} \pm q^* \cdot SE$$

In R: > qnorm(0.975)

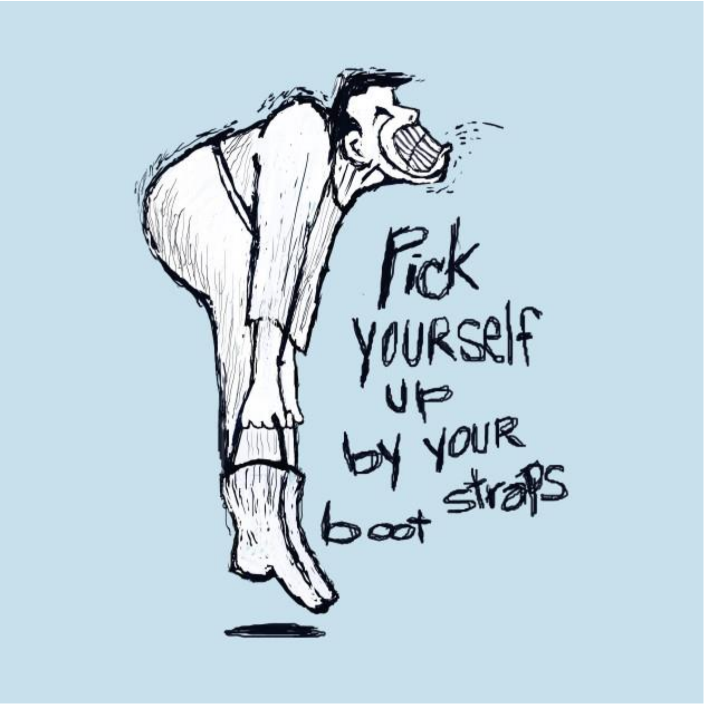
[1] 1.96

Sampling distributions

Unfortunately we can't calculate the sampling distribution ☹

- Therefore we can't get the SE from the sampling distribution ☹

We have to pick ourselves up by 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 *<u>the sample is the population</u>* (plug-in principle)

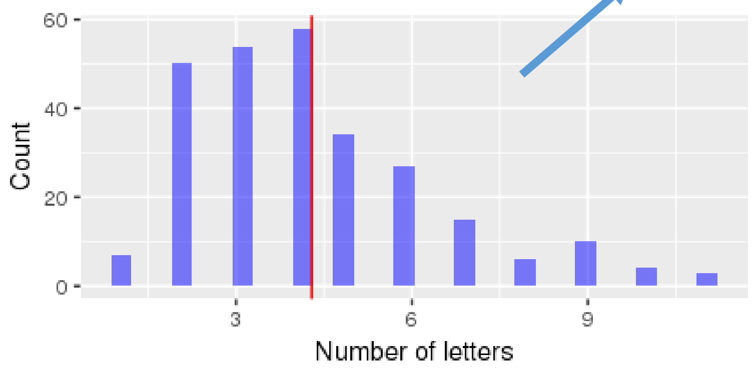
1. We then sample *n* points *<u>with replacement</u>* from our sample, and compute our statistic of interest

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 distribution illustration

μ



<u>The sample (n = 10)</u>
10, 3, 3, 3, 4, 3, 2, 6, 4, 5

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

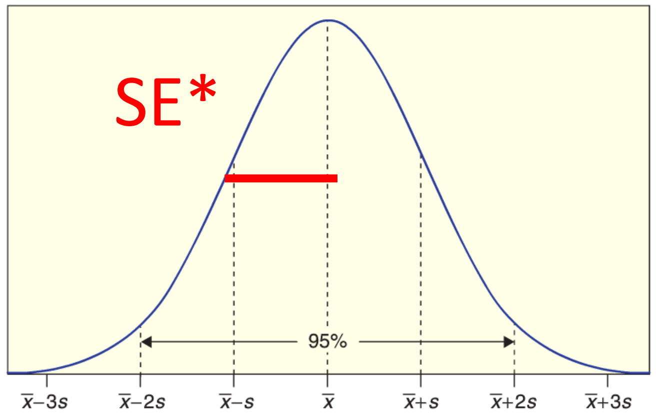
$\bar{x}^* = 4$

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

$\bar{x}^* = 4.1$

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

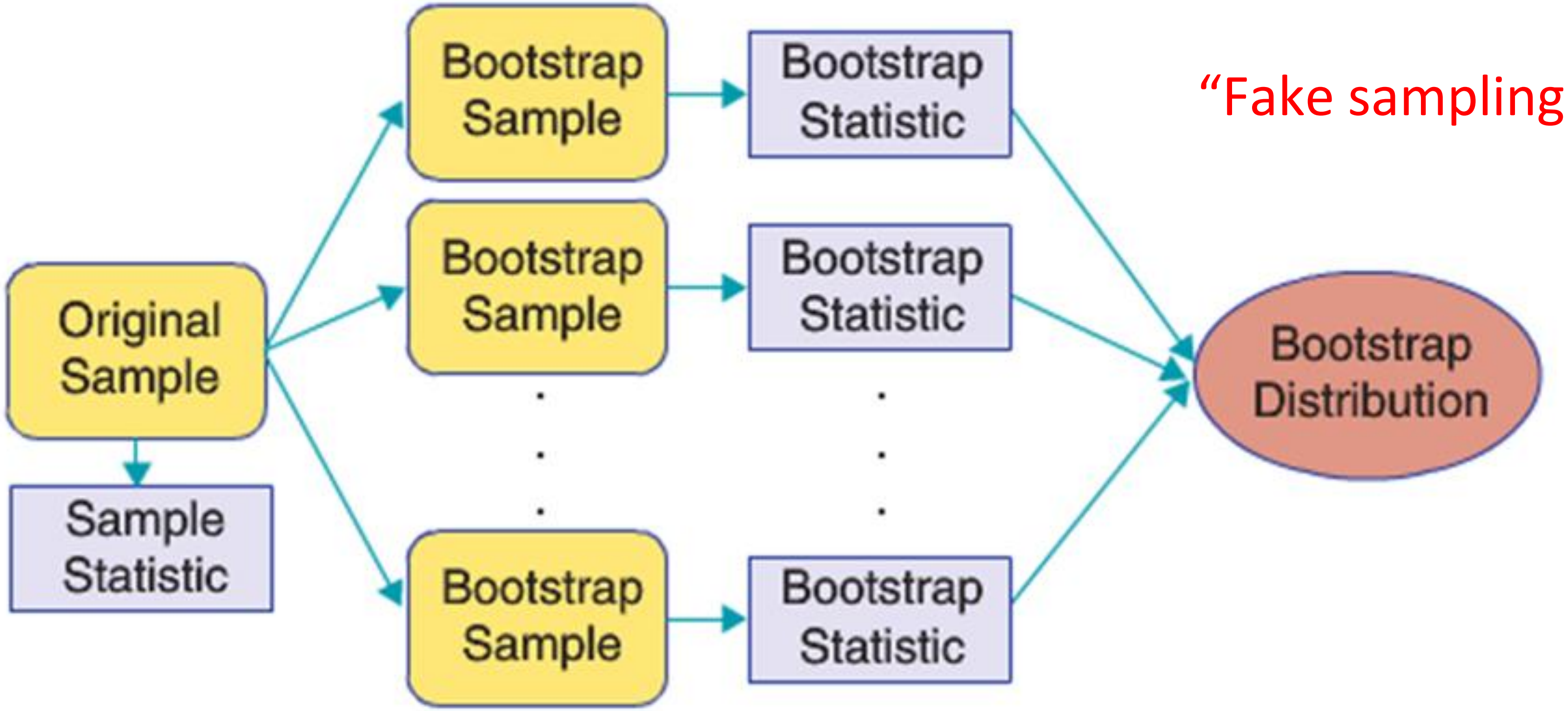
$\bar{x}^* = 3.9$



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:

$$\textit{Statistic} \pm 2 \cdot SE^*$$

Where SE* is the standard error estimated using the bootstrap

What are the steps needed to create a bootstrap SE?

1. Start with a sample

2. Repeat steps 10,000 times

    a. Resample the points in the sample to get a bootstrap sample

    b. Compute the statistic of interest on the bootstrap sample

3. Take the standard deviation of the bootstrap distribution to get SE*
SE*

Sampling with replacement from a vector

```
my_sample <- c(3, 1, 4, 1, 5, 9)
```

<u>To get a sample of size n = 6 with replacement:</u>

```
> boot_sample  <-  sample(my_sample,  6,  replace = TRUE)
```

Sampling distribution in R

```
my_sample <- c(21, 29, 25, 19, 24, 22, 25, 26, 25, 29)

bootstrap_dist <-  do_it(10000) * {

        curr_boot <- sample(my_sample , 10, replace = TRUE)
        mean(curr_boot)

}

SE_boot <- sd(bootstrap_dist)
```

Bootstrap confidence interval in R

```
obs_mean <- mean(my_sample)

CI_lower <-  obs_mean  - 2 * SE_boot

CI_upper <-  obs_mean  + 2 * SE_boot
```

Confident intervals

Q: Are we feeling confident about confidence intervals?

Worksheet to explore concepts

Please work in pairs and fill out the class worksheet on confidence intervals from sampling and bootstrap distributions

To fill out the worksheet, use the web app at: http://bit.ly/SE_app

I believe in pulling yourself up by your own bootstraps. I believe it is possible — I saw this guy do it once in Cirque du Soleil. It was magical.

— *Stephen Colbert* —

Next class: hypothesis tests!

Homework 4

- Use the link on Canvas to access homework 4 on R Studio Cloud
- Due on Gradescope at 11:30pm on Sunday February 16th