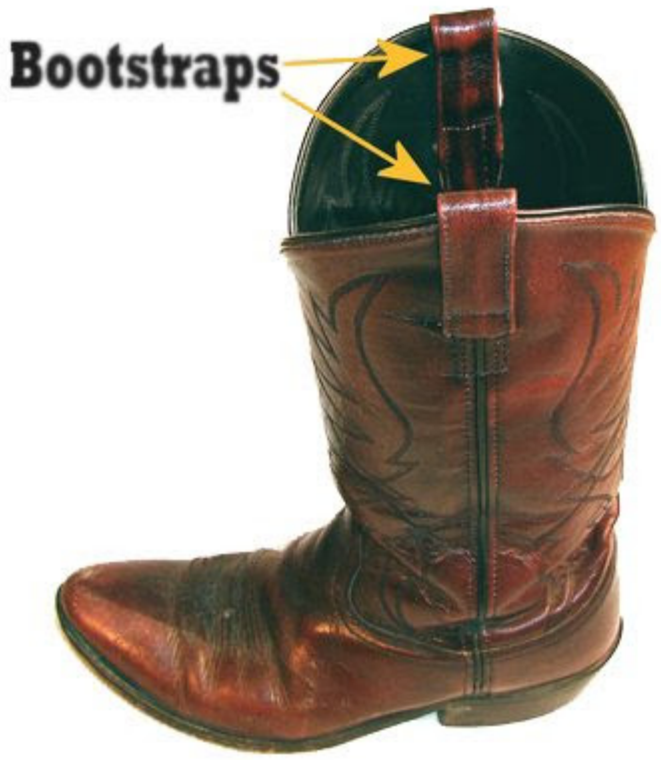
# The bootstrap



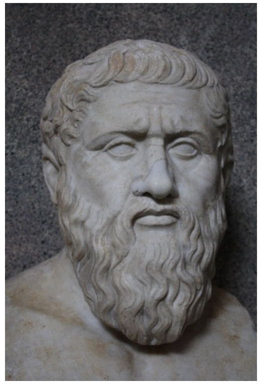
# Overview

Review: confidence intervals and the bootstrap

Calculating bootstrap confidence intervals in R

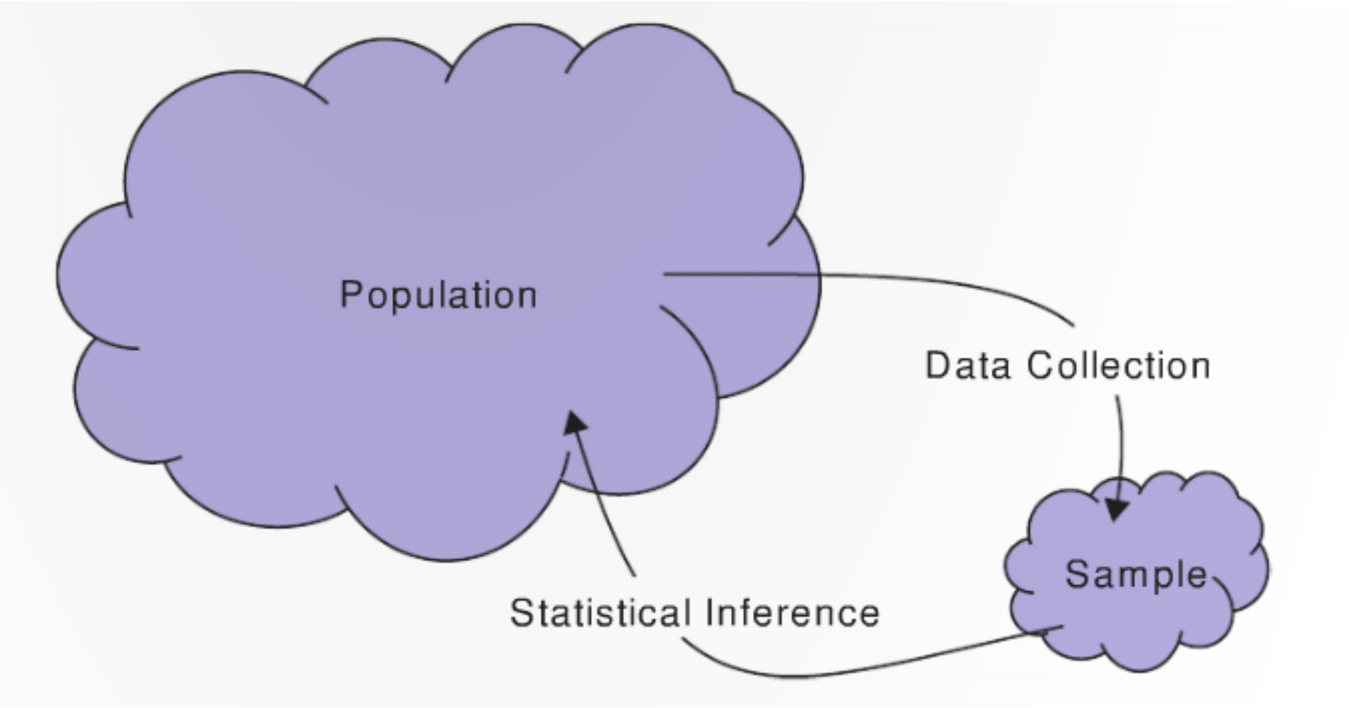
- With practice problems!

# Statistical inference: confidence intervals

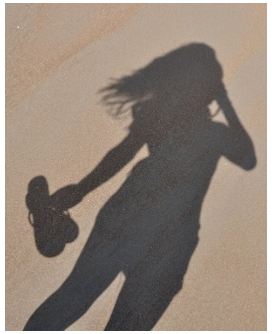


$\mu$

Confidence interval




$\bar{x}$ $\hat{SE}$

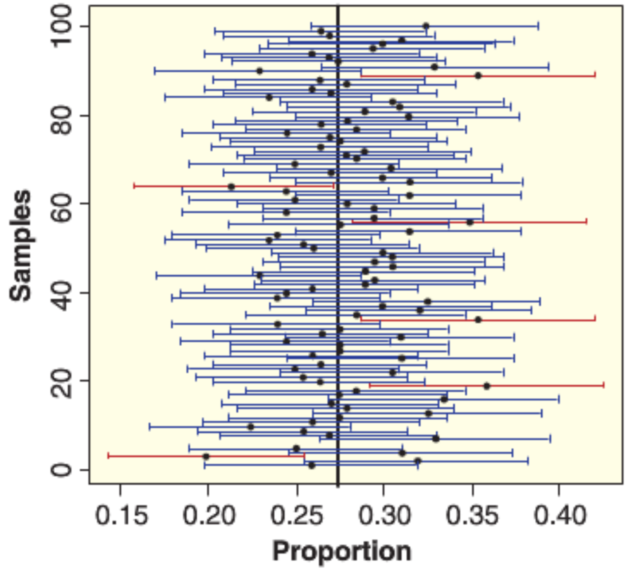


# Quick review of confidence intervals

# Review: confidence intervals

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


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




There is a tradeoff between:

- The **confidence level**
- The **confidence interval size**

# Example: Gallup poll

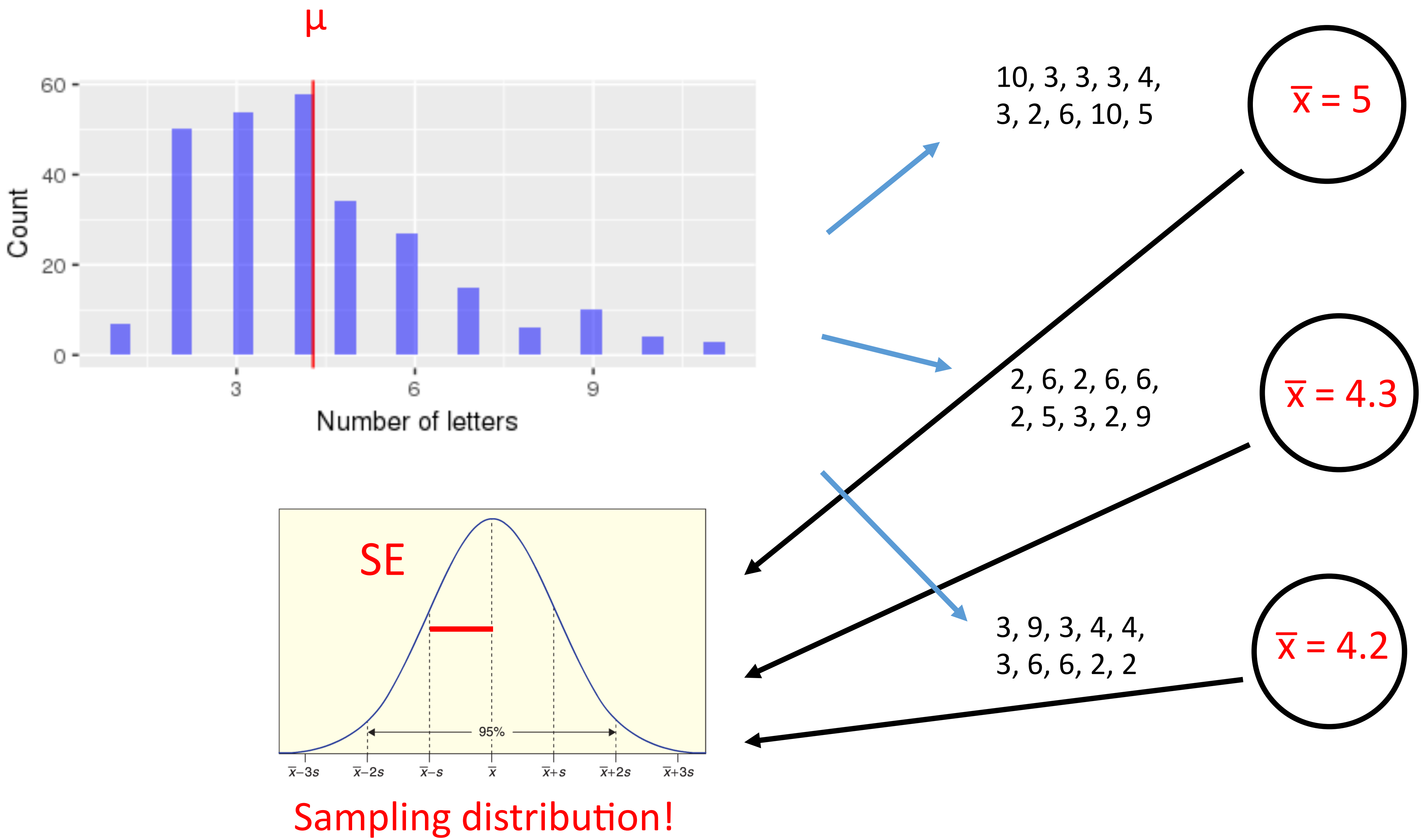
75% of Americans watched the Super Bowl plus or minus 5%

How do we interpret this?

The <u>population parameter</u> ($\pi$) lies somewhere between 70% to 80%

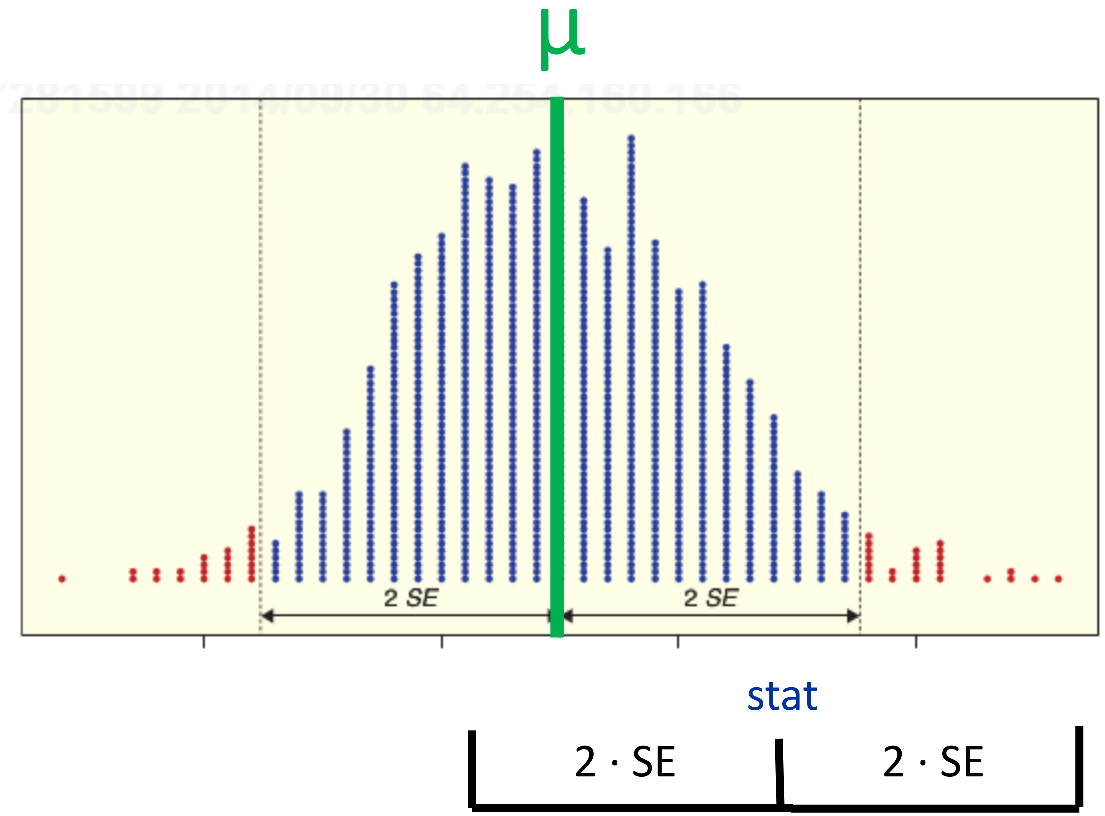
i.e., if they sampled all Americans, the true population proportion ($\pi$) of people who watched the Super Bowl is in this range

# Review: sampling distribution illustration



# Sampling distributions

For a sampling distribution that is a normal distribution, 95% of ***statistics*** lie within 2 standard deviations (SE) for the population mean?




Confidence interval

If we have:

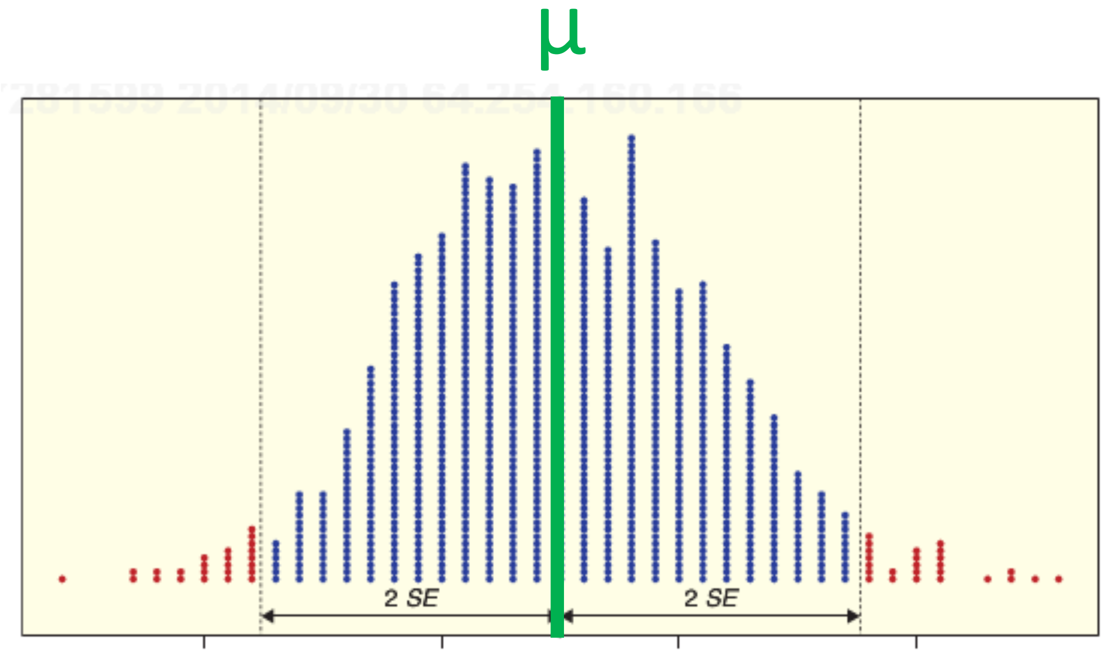
- A statistics value
- The SE

We can compute a 95% confidence interval!

$$CI_{95} = \text{stat} \pm 2 \cdot SE$$

# Sampling distributions

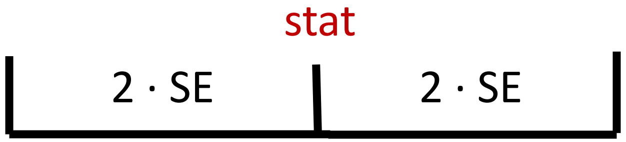
For a sampling distribution that is a normal distribution, 95% of ***statistics*** lie within 2 standard deviations (SE) for the population mean?




If we have:

- A statistics value
- The SE

We can compute a 95% confidence interval!




Confidence interval

$$CI_{95} = \text{stat} \pm 2 \cdot SE$$

# 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


$$CI_{99.7} = \text{stat} \pm 3 \cdot SE$$

stat

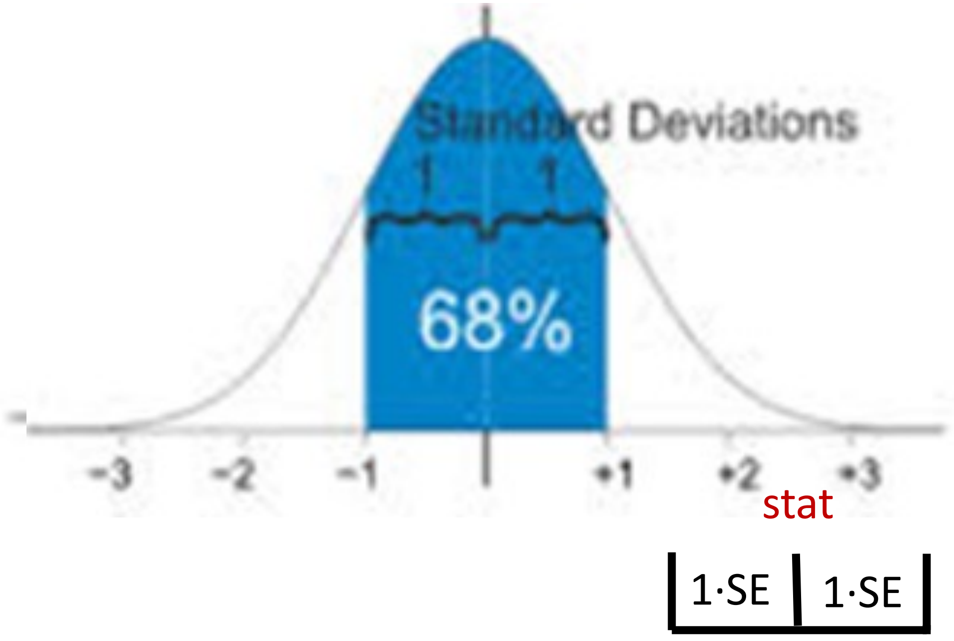
3 · SE    3 · SE

Confidence interval

# 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




Confidence interval

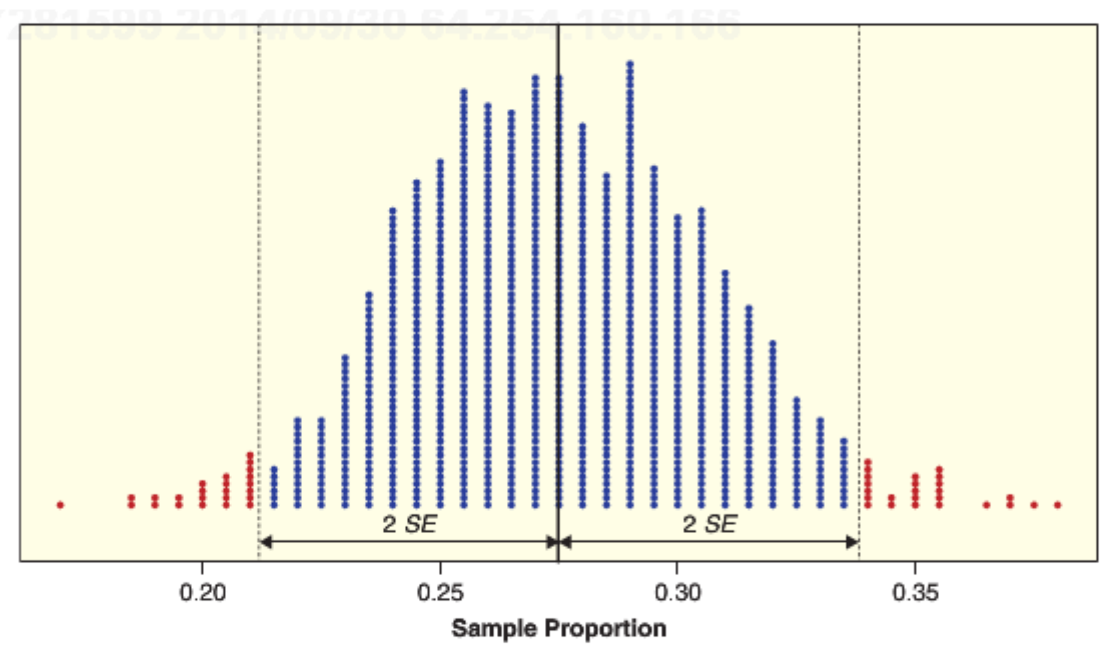
$$CI_{99.7} = \text{stat} \pm 3 \cdot SE$$

$$CI_{68} = \text{stat} \pm 1 \cdot SE$$

# 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




$$CI_{.} = \text{stat} \pm q^* \cdot SE$$

In R:  qnorm(0.975)

[1]   1.96

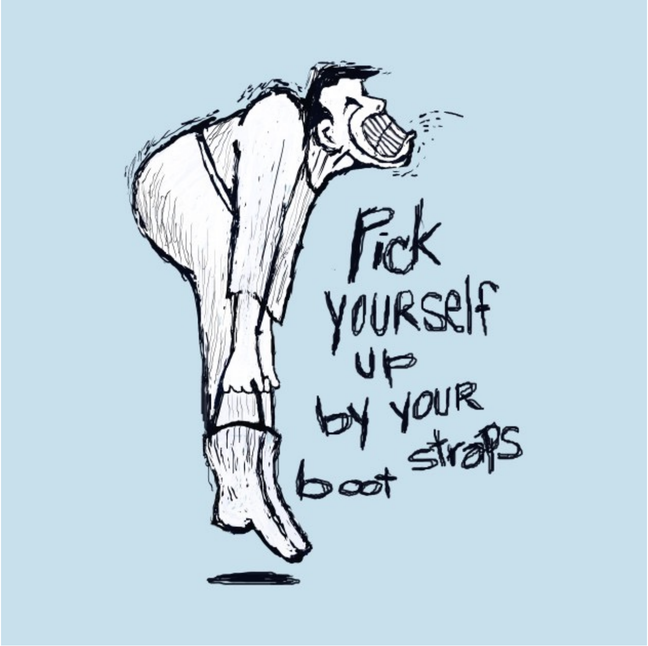
# The bootstrap continued

# Sampling distributions

As previously discussed, in practice we can't calculate the sampling distribution by repeating sampling from a population ☹

- 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 stat $\pm\ 2 \cdot \hat{SE}$ to get the 95% CI

# Plug-in principle

Suppose we get a sample from a population of size *n*

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

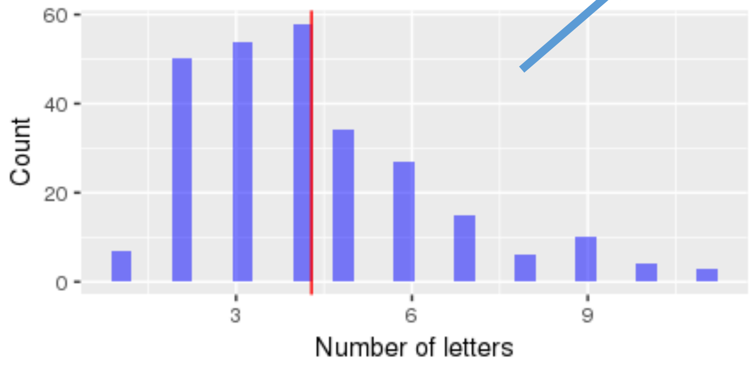
1. We then sample *n* points *with replacement* from our sample, and compute our statistic of interest

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 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

$\bar{x}^* = 4$

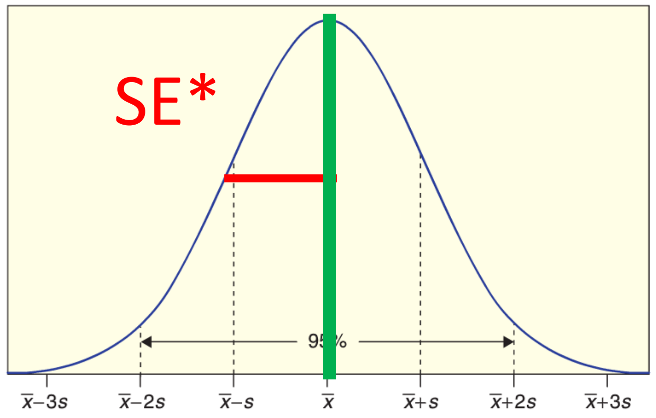
3, 3, 2, 3, 6,
4, 6, 5, 3, 6

$\bar{x}^* = 4.1$

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

$\bar{x}^* = 3.9$

$\bar{x}$




Bootstrap distribution!

Notice there is no 9's in the bootstrap samples

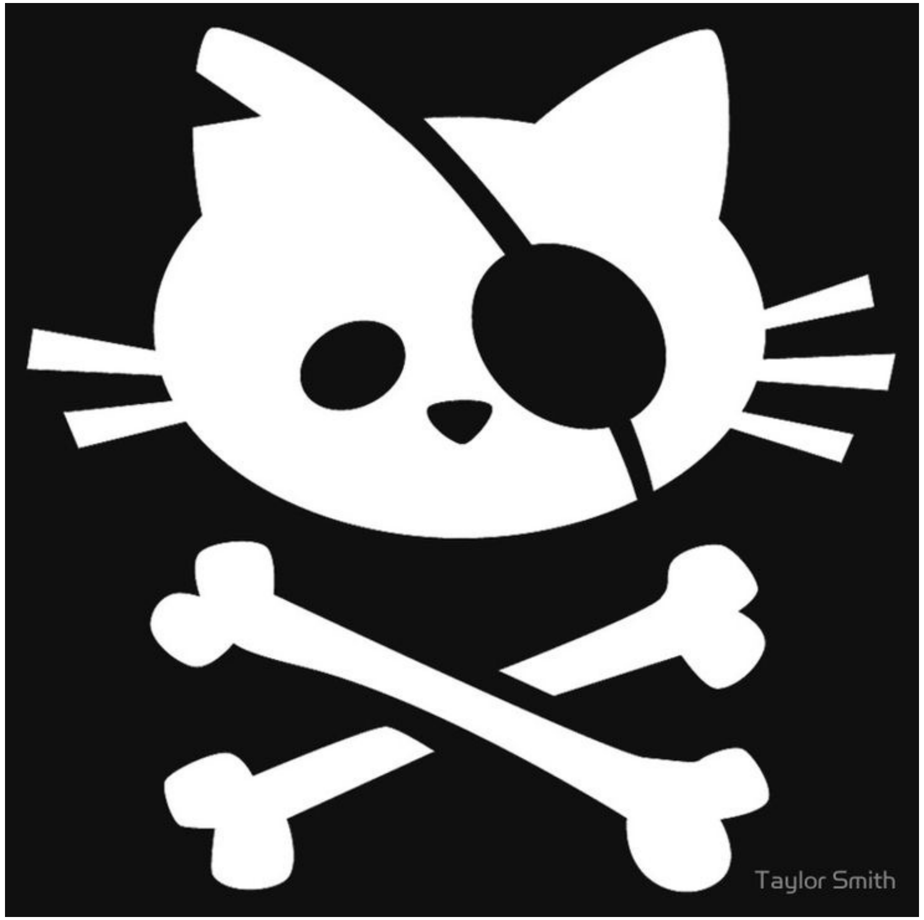
# 95% Confidence Intervals

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

# Bootstrap confidence intervals in R



# 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*

# 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
```

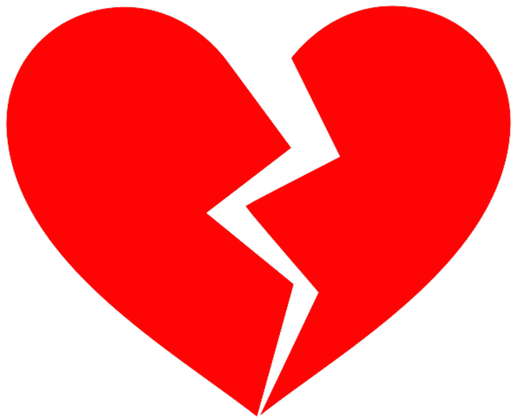
# Example: One true love?

A survey asked 2625 people whether they agreed with the statement “There is only one true love for each person”

1812 of the respondents disagreed

Compute a 90% confidence interval for the proportion who disagreed

SDS100::download_class_code(10)



# Let's try it with some real data in R!