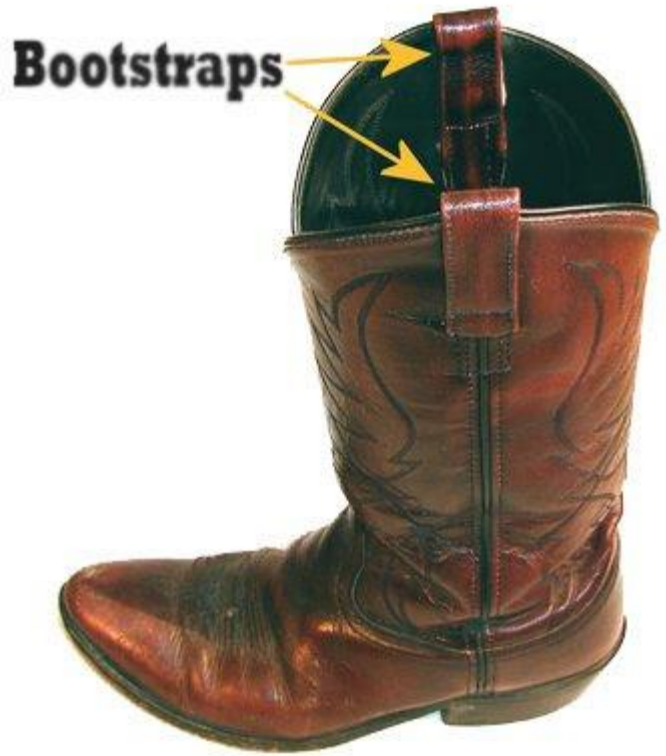
The bootstrap



Overview

Review of confidence intervals

Review and continuation of the bootstrap

Calculating bootstrap confidence intervals in R

If there is time: more practice creating confidence intervals

Announcement

Homework 5 has been posted!

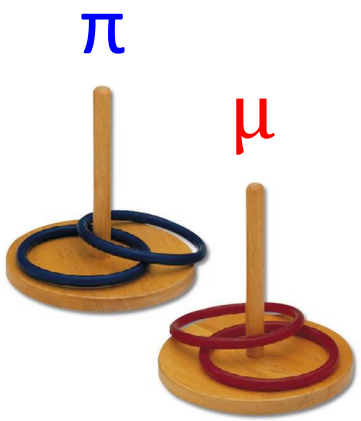
It is due on Gradescope on Sunday February 22nd at 11pm

- **Be sure to mark each question on Gradescope!**

Quick review of confidence intervals

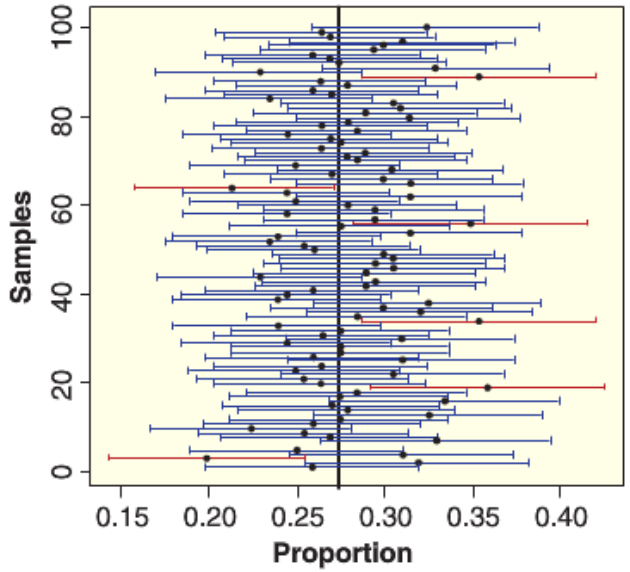
Review: confidence intervals

"Parameter catchers"



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

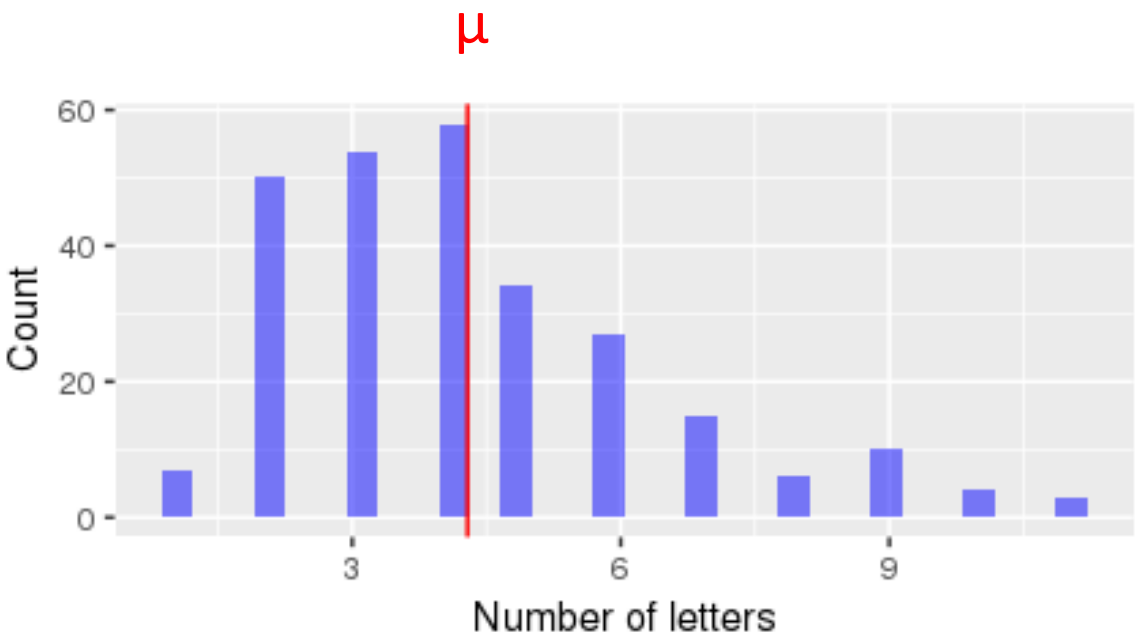
75% of Americans watched the Super Bowl plus or minus 5%

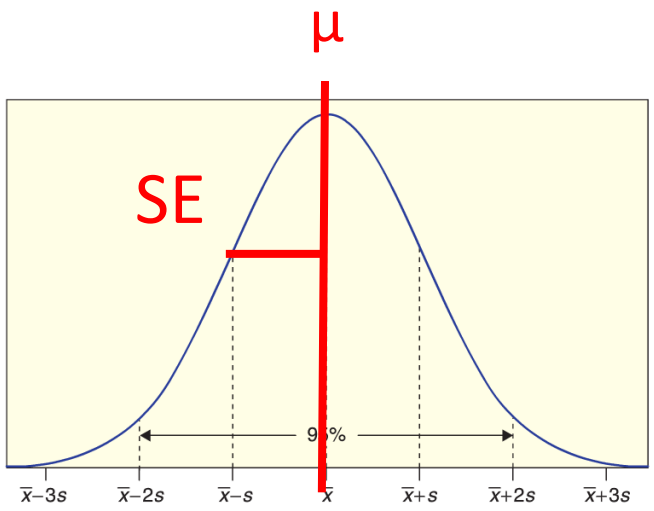
How do we interpret this?

The population parameter (π) lies somewhere between 70% to 80%

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

Review: sampling distribution illustration





Sampling distribution!

n = 10

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

$\bar{x} = 5$

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

$\bar{x} = 4.3$

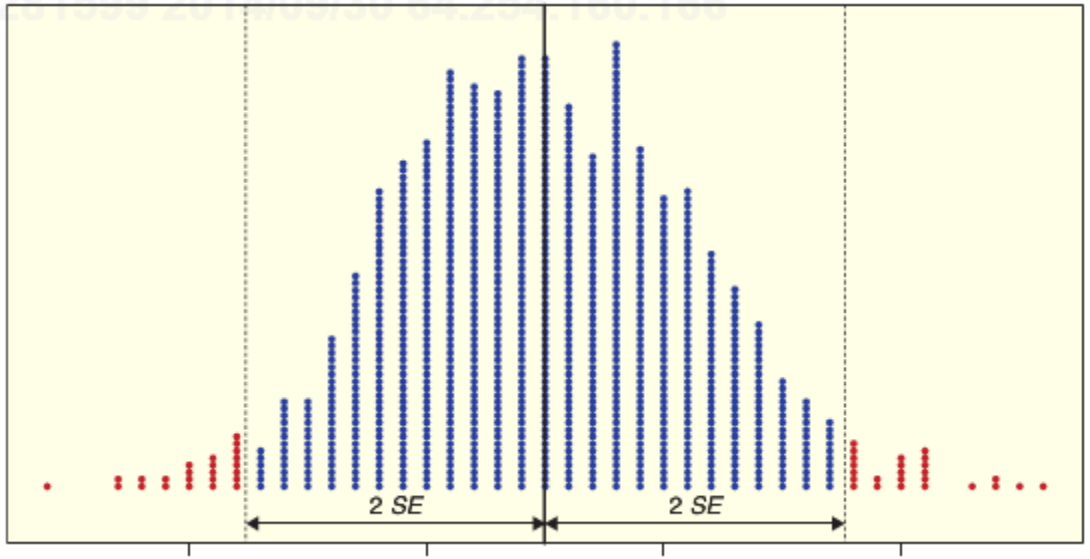
3, 9, 3, 4, 4,
3, 6, 6, 2, 2

$\bar{x} = 4.2$

Sampling 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%



stat

2 · SE | 2 · SE

Confidence interval

If we had:

- A statistic value
- The SE

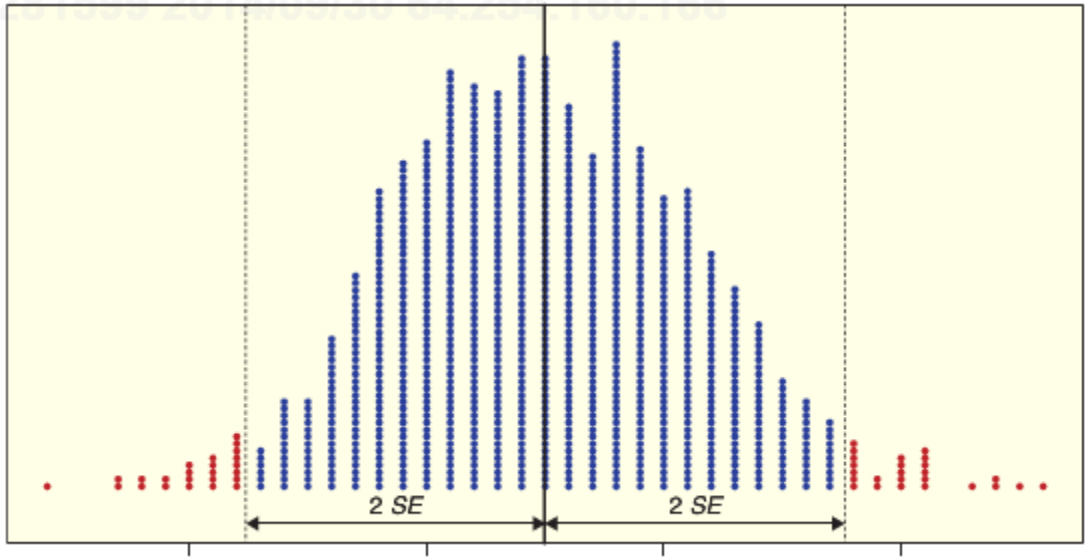
We could compute a 95% confidence interval!

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

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 statistic value
- The SE

We could compute a 95% confidence interval!

stat

2 · SE | 2 · SE

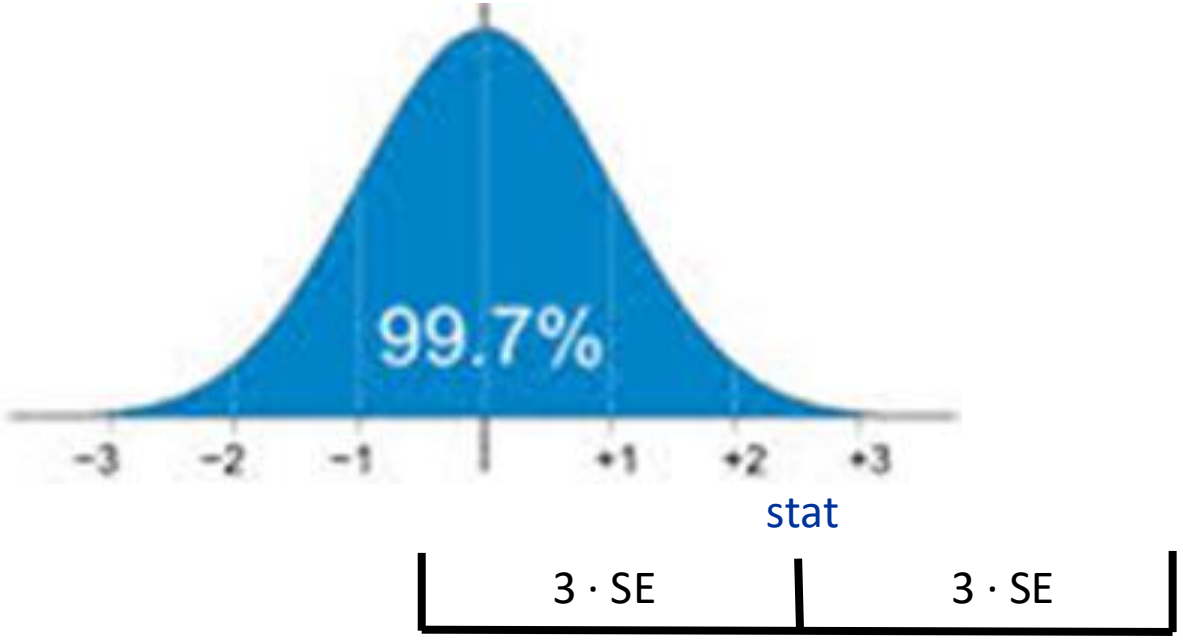
Confidence interval

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

Confidence intervals for other confidence levels

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

A: For normally distributed statistics, 99.7% of our statistics lie within 3 standard deviations (i.e., standard errors) of the mean

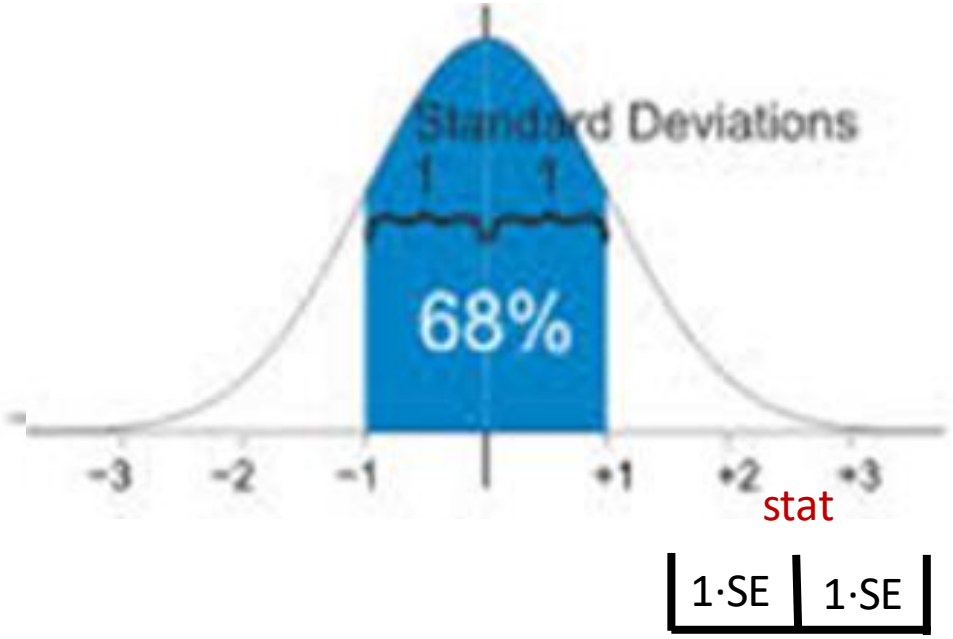


$$CI_{99.7} = \text{stat} \pm 3 \cdot SE$$

Confidence intervals for other confidence levels

Q: How could we get a 68% confidence interval confidence level?

A: For normally distributed statistics, 68% of our statistics lie within 1 standard deviation (i.e., standard error) of the mean



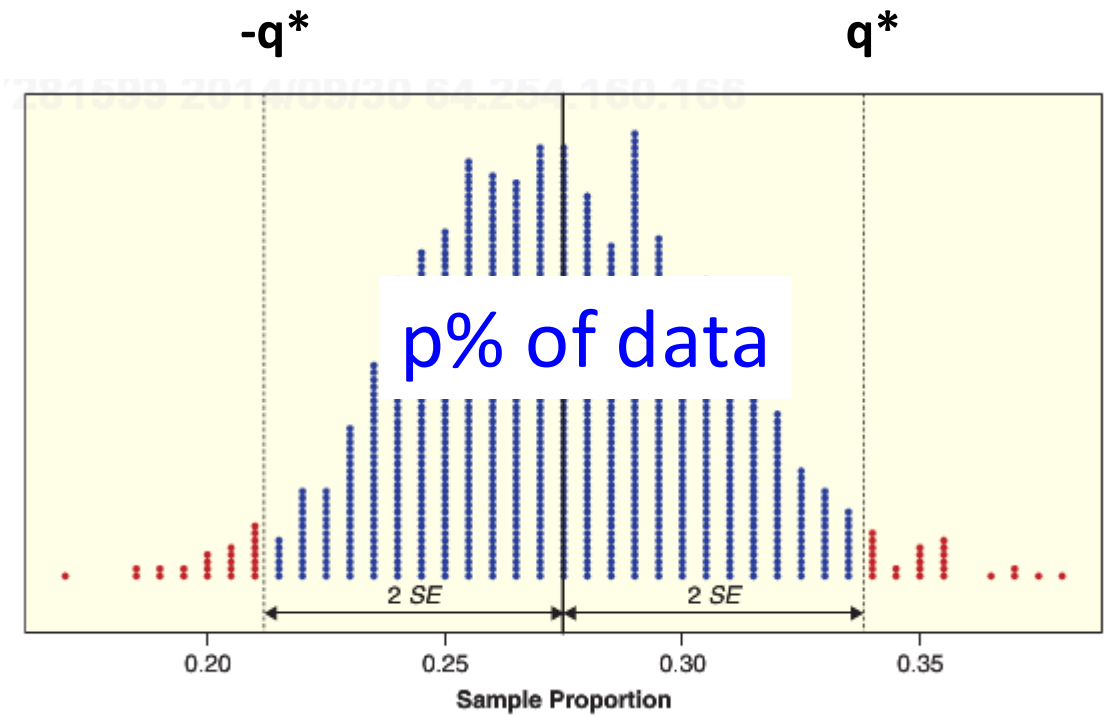
$$CI_{99.7} = \text{stat} \pm 3 \cdot SE$$

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

Confidence intervals for other confidence levels

Q: How could we get a confidence interval for the q^{th} confidence level?

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



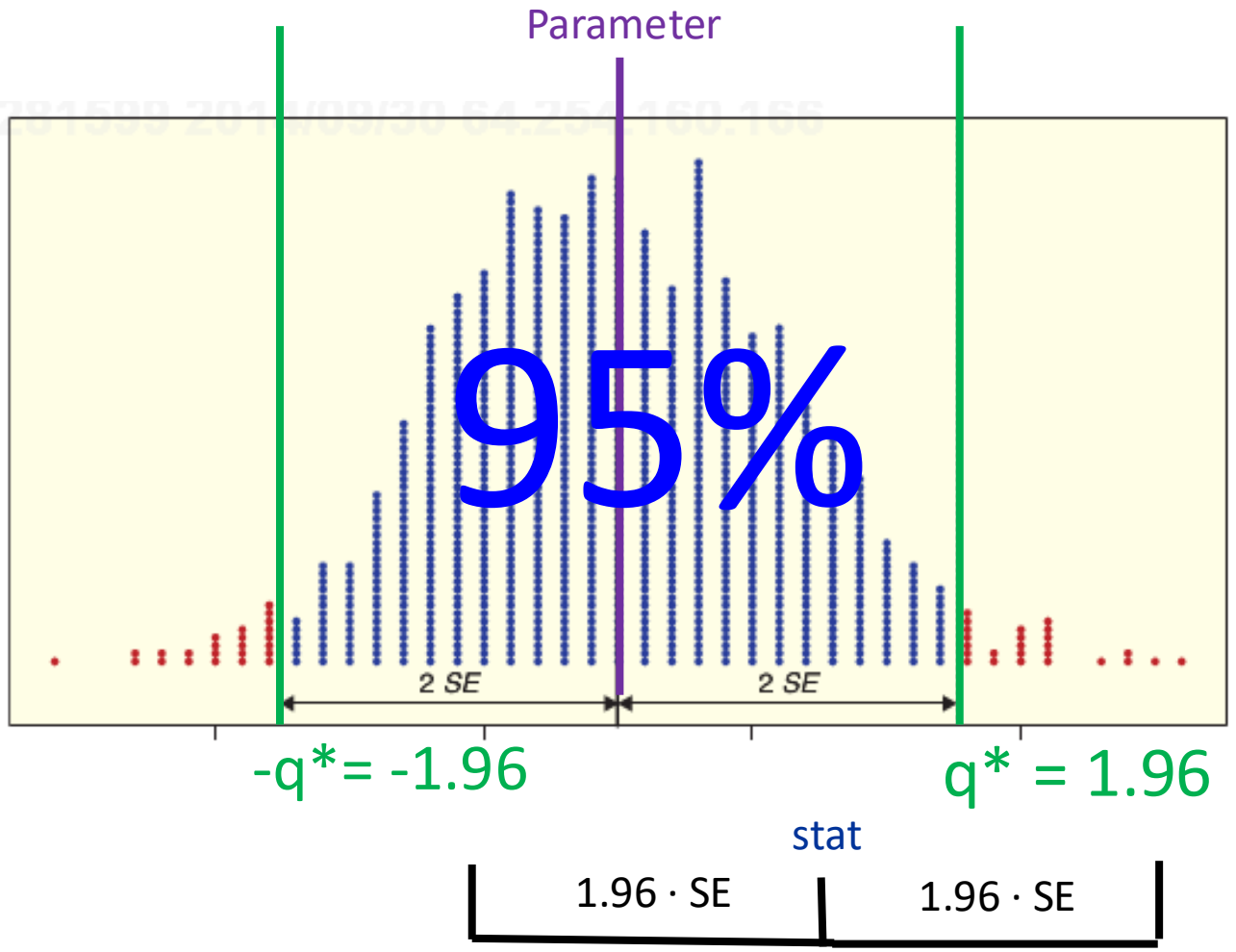
$$CI = stat \pm q^* \cdot SE$$

Critical value q*

Critical value

For a 95% CI, we choose q* = 2

- 95% of our statistics lie within
  -2 and 2 SE of the parameter
  - (1.96 to be more precise)




```
crit_val_95 <- cnorm(.95)
```

```
1.96
```

Critical value q*

confidence interval: stat ± q* · SE = stat ± 1.96 · SE

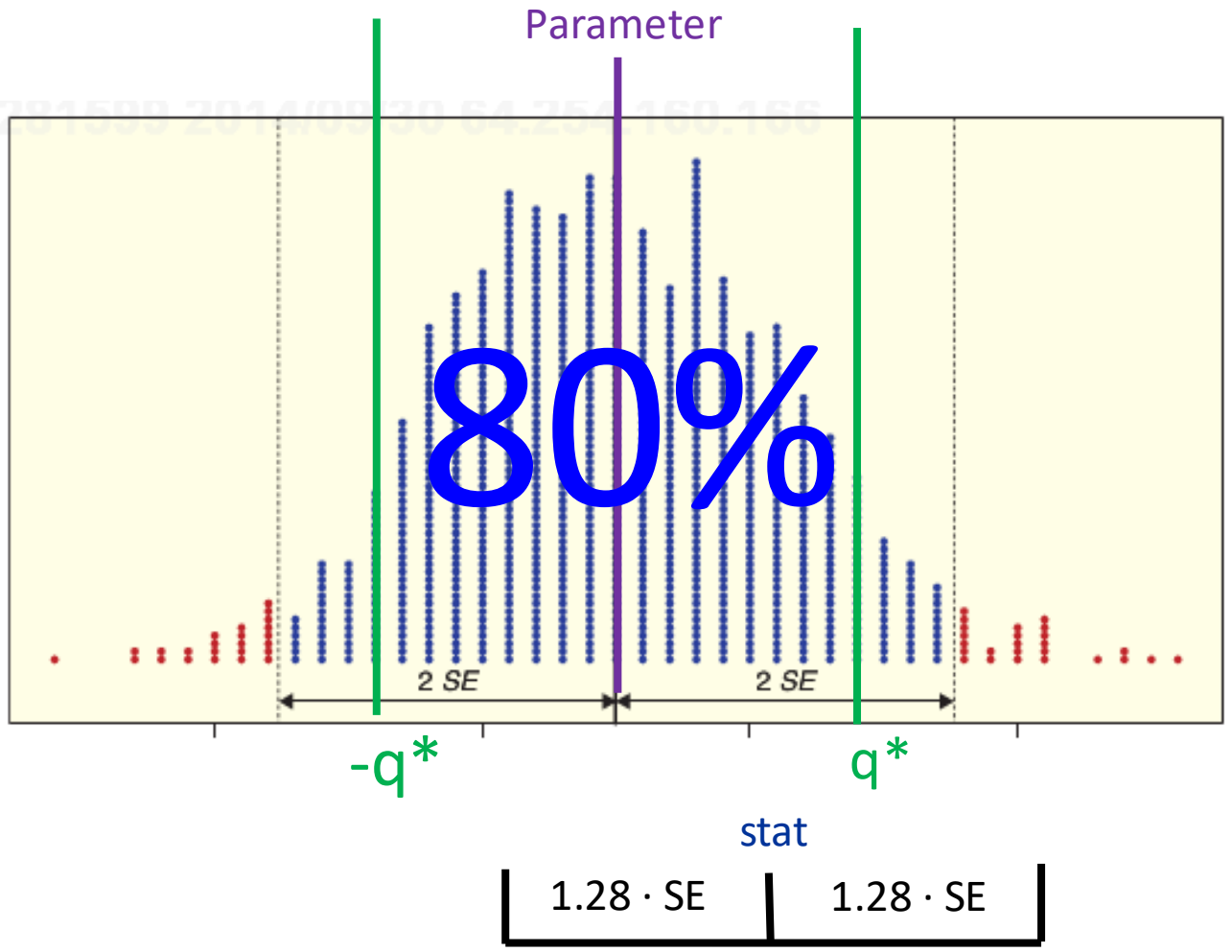
Critical value

We can choose critical values q* for other confidence levels

- E.g. for an 80% CI, we choose q* so that 80% of our statistics lie within -q* and q* SE of the parameter

```
crit_val_80 <- cnorm(.80)
```

```
1.28
```



Critical value q*

confidence interval: stat ± q* · SE = stat ± 1.28 · SE

Sampling distributions

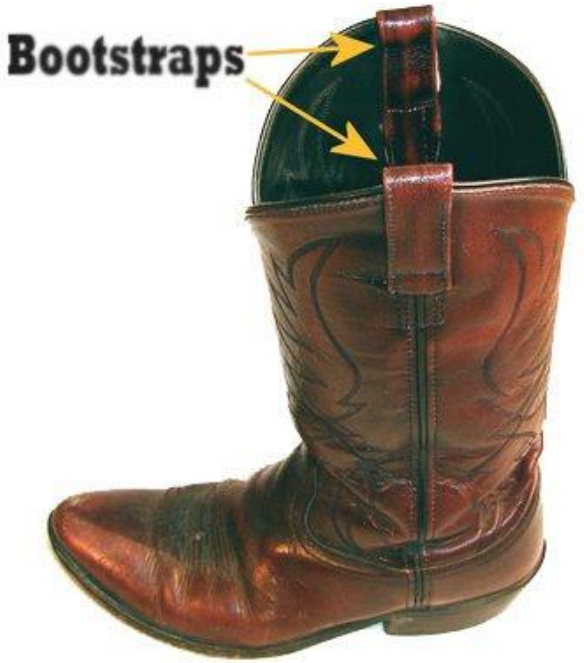
Q: Could we calculate the SE by repeatedly sampling from a population to create sampling distribution, and then take the sd of this sampling distribution?

- A: Not in the real world because it would require running our experiment over and over again…

Sampling distributions

Q: If we can't calculate the sampling distribution, what else can we do?

- A: We can pick ourselves up from the bootstraps



1. Estimate SE with $\hat{SE}$ ***from a single sample of data***
2. Then use $\bar{x} \pm 2 \cdot \hat{SE}$ to get the 95% CI

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

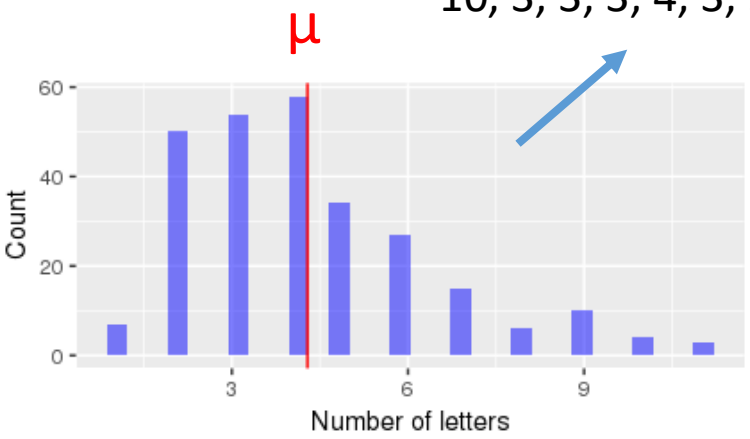
2. We repeat this process 1000's of times and get a ***bootstrap sample distribution***

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

The sample (n = 10)

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



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

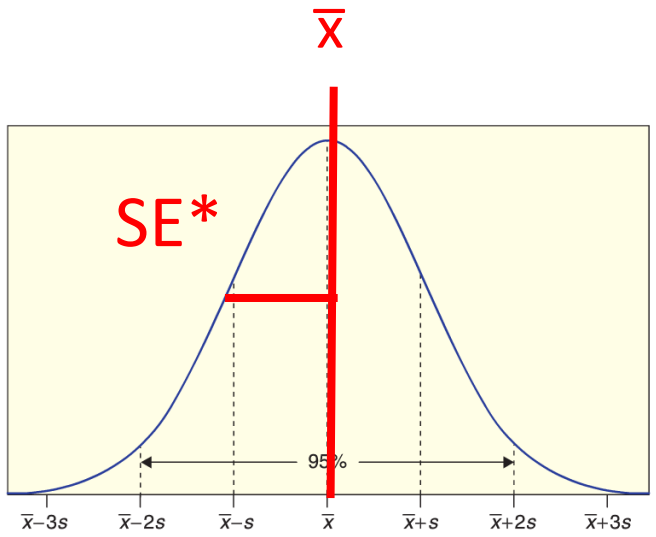
$\bar{x}^* = 4$

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

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

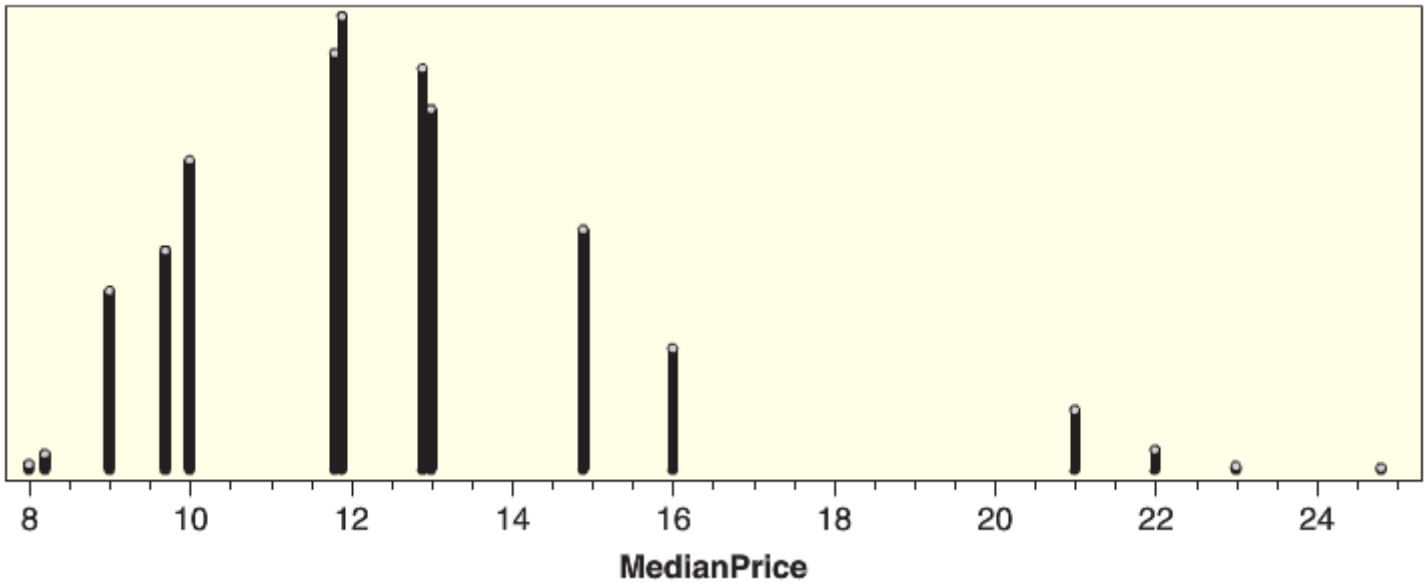
Findings CIs for many different parameters

The bootstrap method works for constructing confidence intervals for many different types of parameters!

Caution: the bootstrap does not always work

Always look at the bootstrap distribution, if it is poorly behaved (e.g., heavily skewed, has isolated clumps of values, etc.), you should not trust the intervals it produces.

Median prices of Mustangs



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)
```

To get a sample of size n = 6 with replacement:

```
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
```

Let’s try it in R!

Practice creating confidence intervals in R

Example 1: 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

Example 2: Average body temperature

Machowaik et al (1992) recorded the body temperature of 130 observations people

They calculated a 95% confidence interval for the body temperatures to be: [98.12, 98.37]

Q: How do we interpret these results?

Q: Is this what you would expect?