Review of confidence intervals and introduction to the bootstrap

Overview

Extensive review and continuation of sampling distributions and confidence intervals

The bootstrap

If there is time:

- Using the bootstrap to create confidence intervals in R

Announcement

Homework 4 has been posted!

It is due on Gradescope on Sunday September 28th at 11pm

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

The material this week is going to be a bit more conceptually challenging.

Please attend the practice sessions and office hours to reinforce your understanding!

Review of sampling distributions, standard errors and confidence intervals

Sampling distributions

Q_2: What is a sampling distribution?

Q_3: What does a sampling distribution show us?

Art time

Please draw:

1. Population
2. Plato
3. Population parameter with appropriate symbol
4. One sample that has 10 points
5. Sample statistic for the mean with appropriate symbol
6. Nine more samples that have 10 points
7. Nine more sample statistics with appropriate symbol
8. A sampling distribution

The standard error

Q_4: What is the **standard error**?

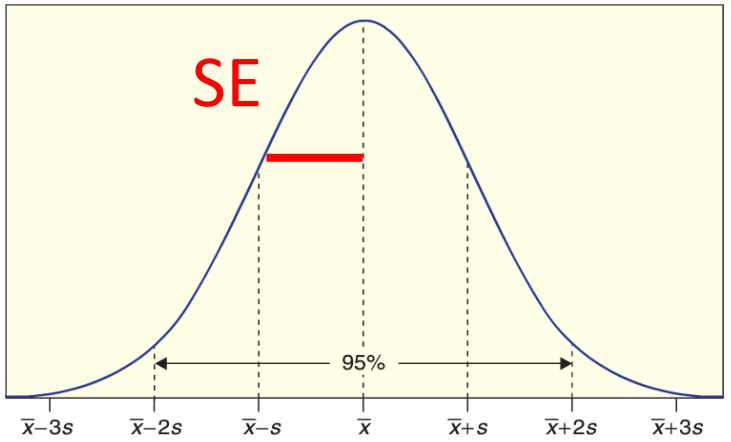
Q_5: What symbol do we use to denote the standard error?

Sampling distribution in R

Q_6: Suppose we had a function called get_sample() that could generate samples from a population

How could we estimate the SE of the mean using R?

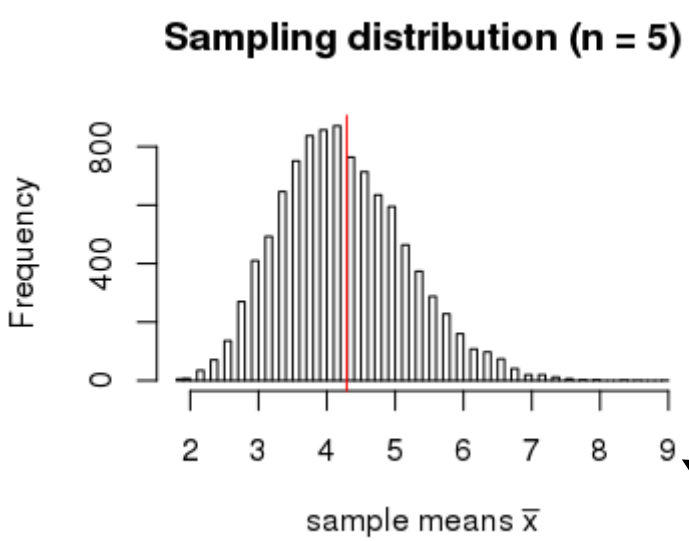
The standard error

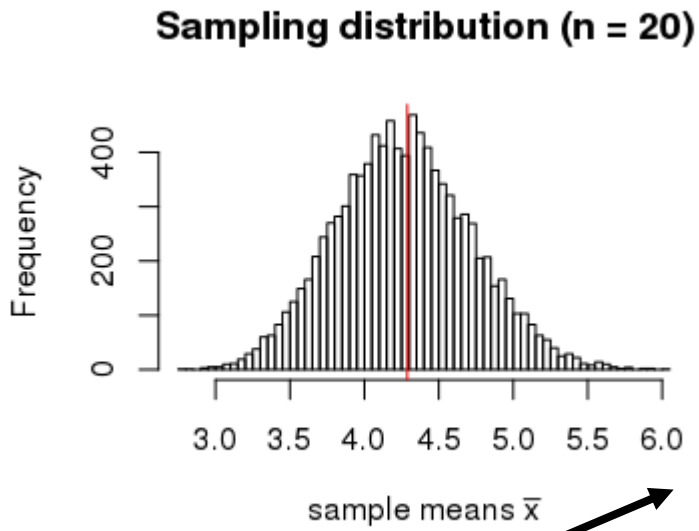


Q_7: What does the size of the standard error tell us?

Q_8: What would it mean if there is a large SE?

Q_9: What are two ways that sampling distribution for the mean $\bar{x}$ changes with larger sample size n?





x-axis range 9 vs. 6

Shapes of sampling distributions

Q_{10}: What is a commonly seen shape for sampling distributions?

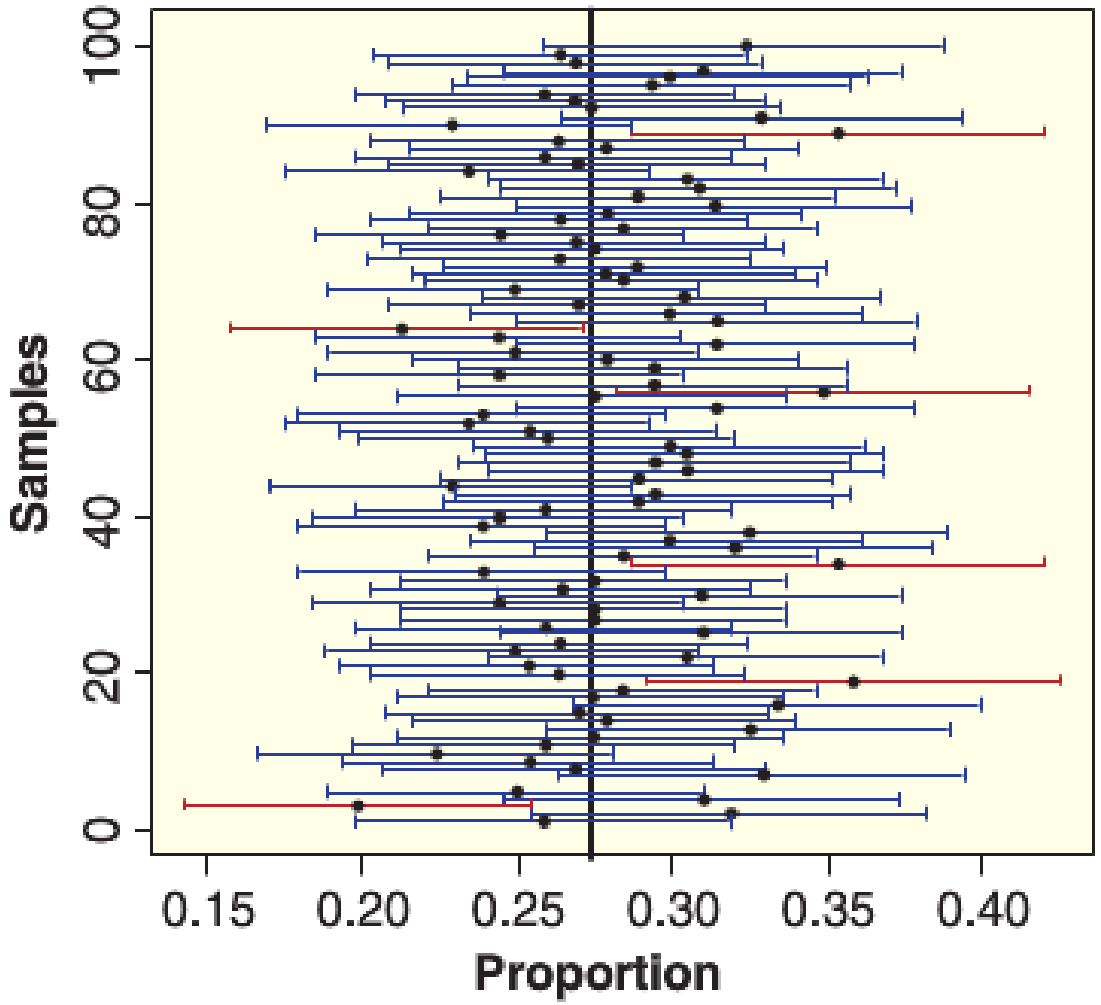
Confidence Intervals

Q_{11}: What is a **confidence interval**?

Q_{12}: What is the **confidence level**?

Confidence Intervals

Q_{13}: For a **confidence level** of 90%, what percent of the intervals will contain the population parameter?

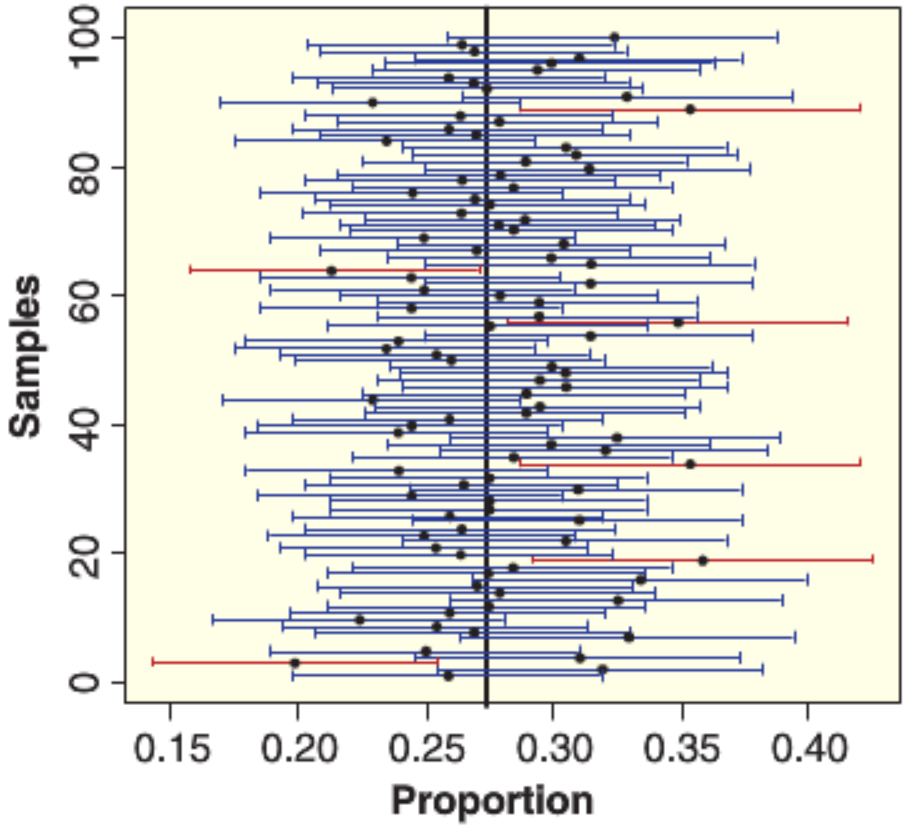


Confidence Intervals

Q_{14}: Is there a tradeoff between the **confidence level** (percent of times we capture the parameter) and the **confidence interval size?**

Confidence Intervals

Q_{15}: For any given confidence interval we compute, do we know whether it has really captured the parameter?



Normal distributions

Q_{16}: For a normal distribution, what percentage of points lie within 2 standard deviations for the population mean?

Sampling distributions

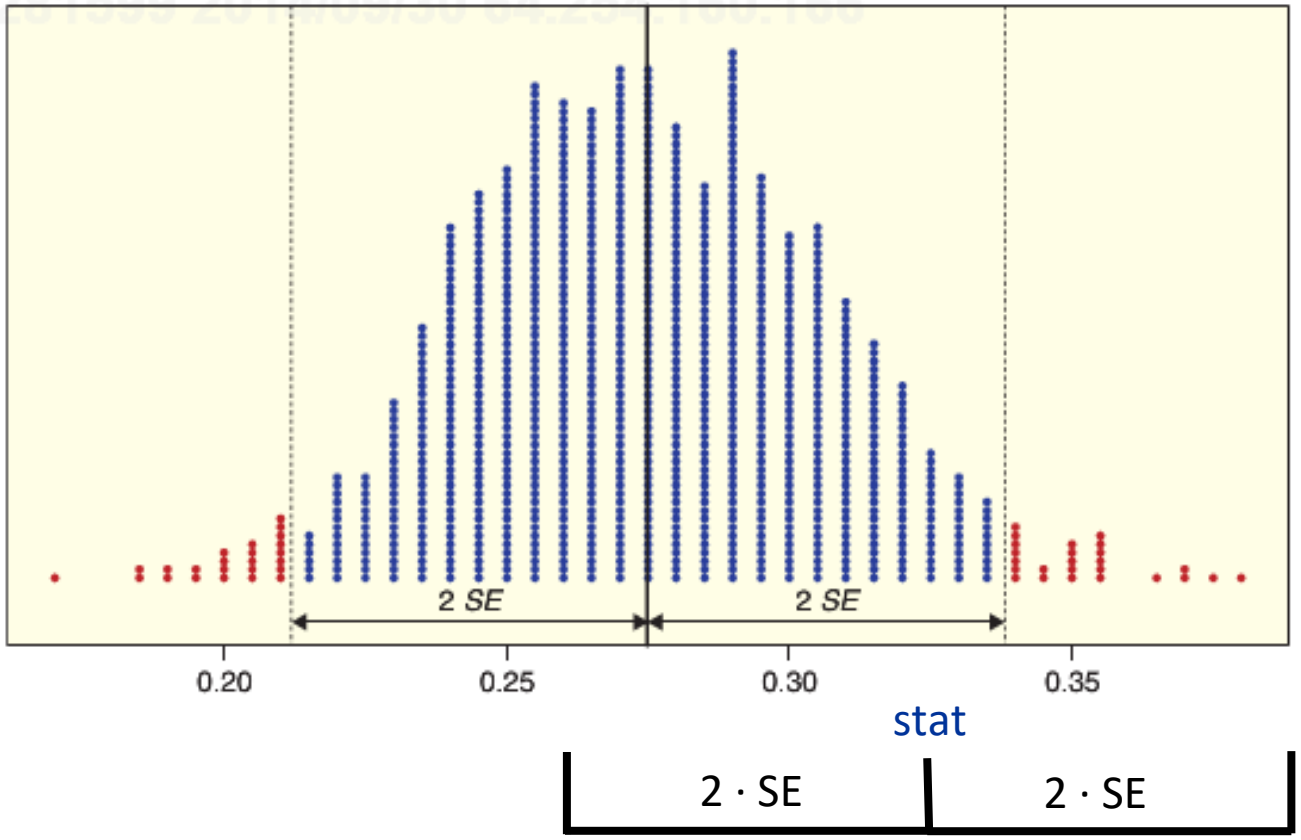
Q_{17}: For a sampling distribution that is a normal distribution, what percentage of ***statistics*** lie within 2 standard deviations (SE) for the population mean?

Q_{18}: If we had a *statistic value* and the value of the *SE*, could we compute a 95% confidence interval?

Confidence intervals

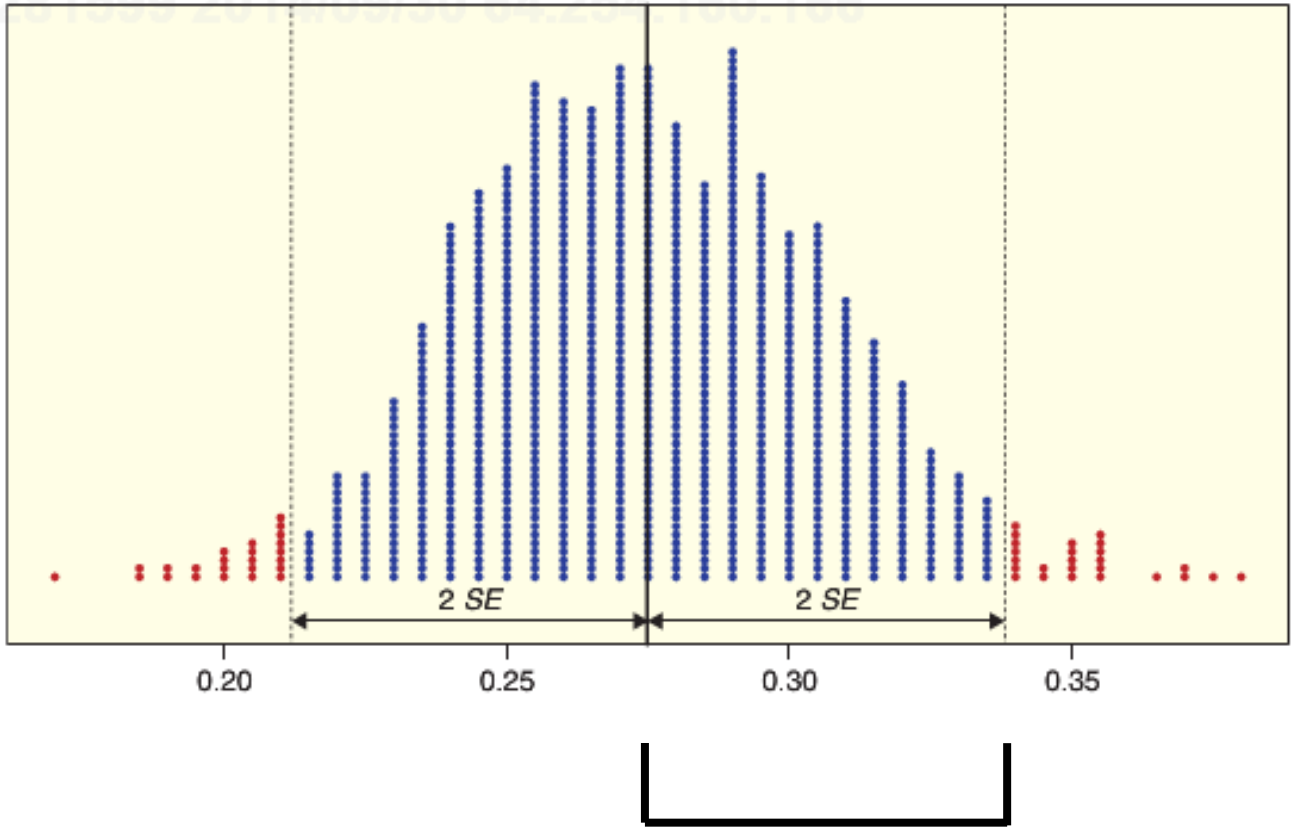
Q_{19}: Suppose we create an interval interval centered at a randomly chosen statistic value

How large would the interval need to be to overlap with the parameter 95% of the time?



Confidence interval

Confidence intervals



Q_{20}: What is a formula we can use to calculate 95% confidence intervals?

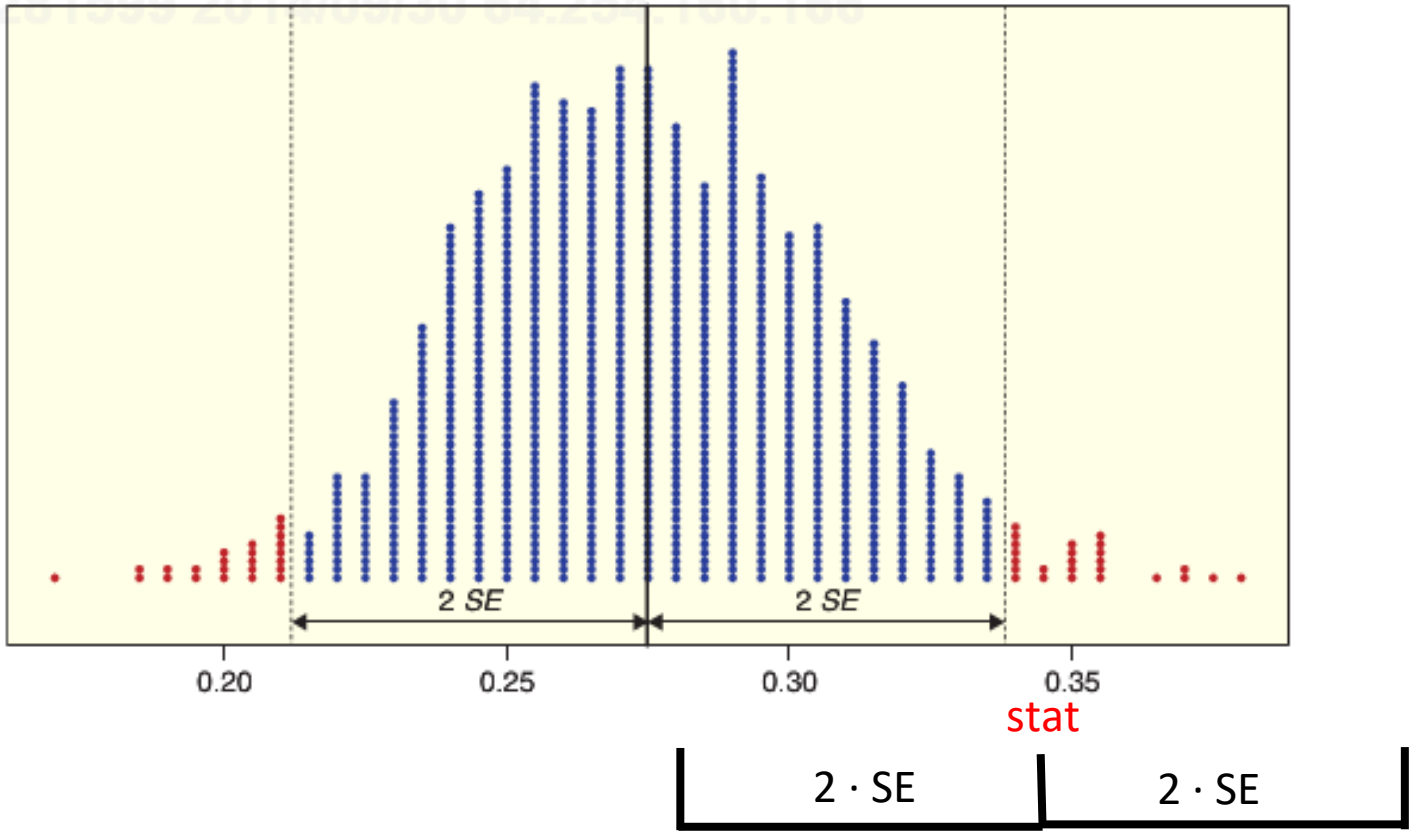
95% confidence interval: stat $\pm$ 2 $\cdot$ SE

Q_{21}:What is this quantity called?

Confidence intervals

Q_{22}: How frequently do 95% confidence intervals fail to capture the parameter of interest?

- 5% of the time

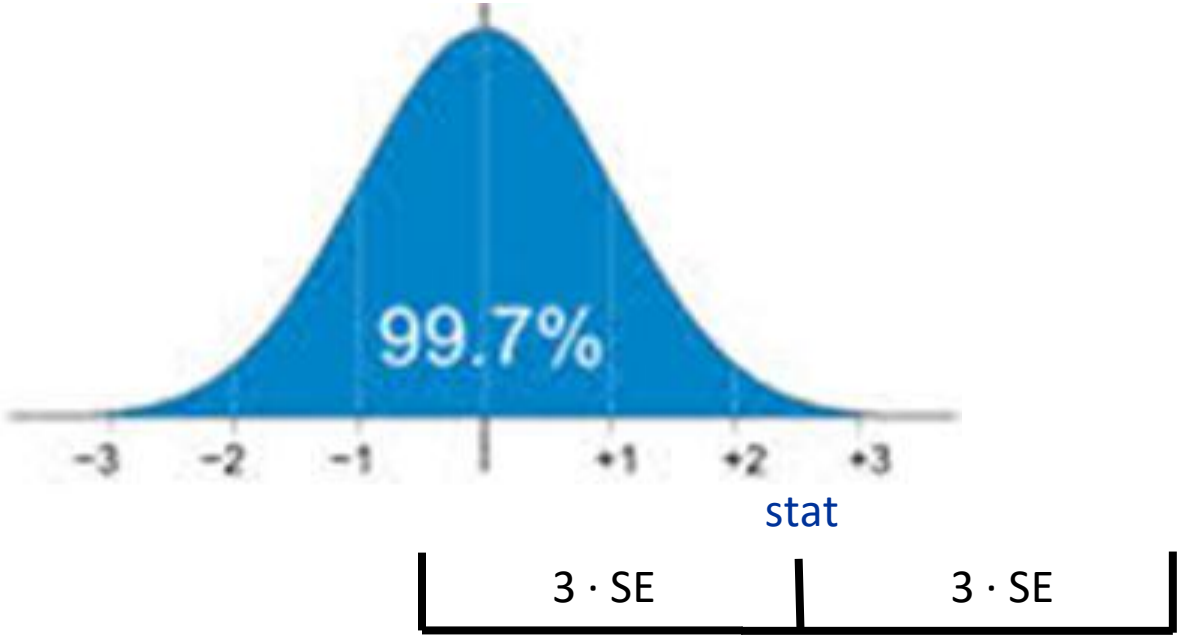


Confidence interval

95% confidence interval: stat ± 2 · SE

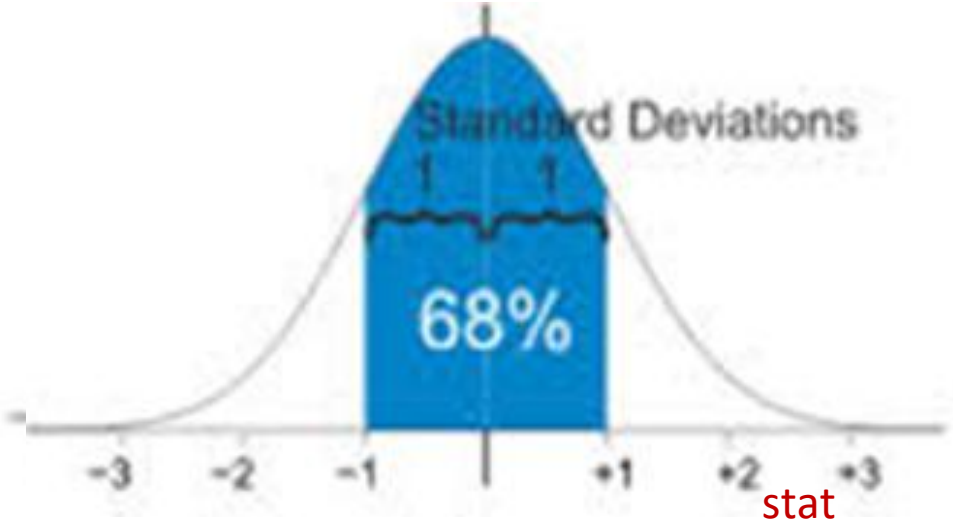
Confidence intervals for other confidence levels

Q_{23}: How could we get a 99.7% confidence interval confidence level?



Confidence intervals for other confidence levels

Q_{24}: How could we get a 68% confidence interval confidence level?



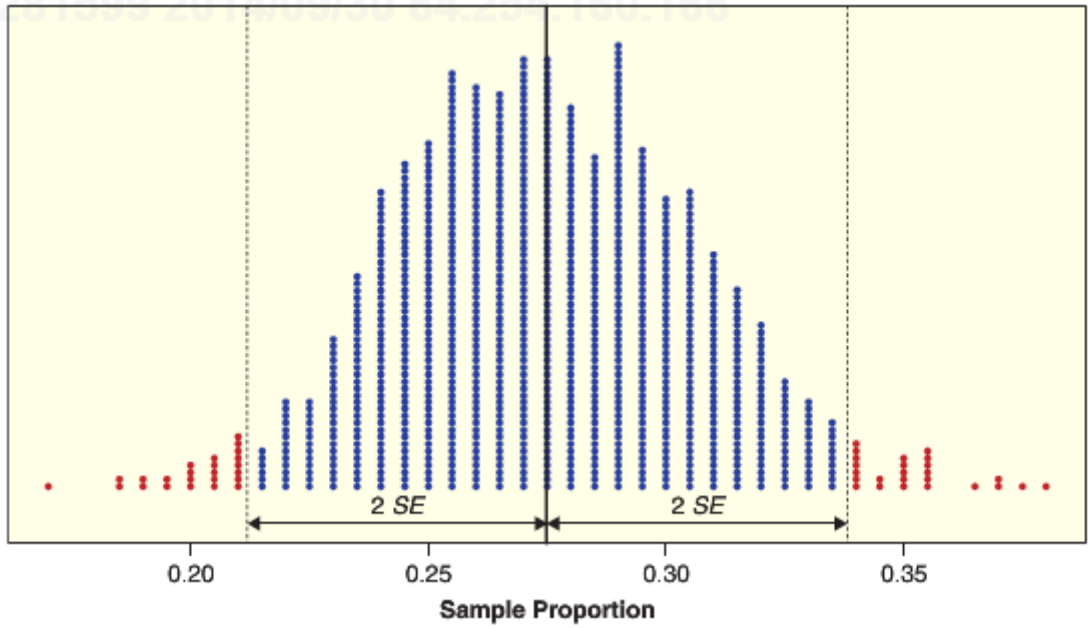
stat

1·SE | 1·SE

Confidence interval

Confidence intervals for other confidence levels

Q_{25}: How could we get a confidence interval for the q^{th} confidence level?



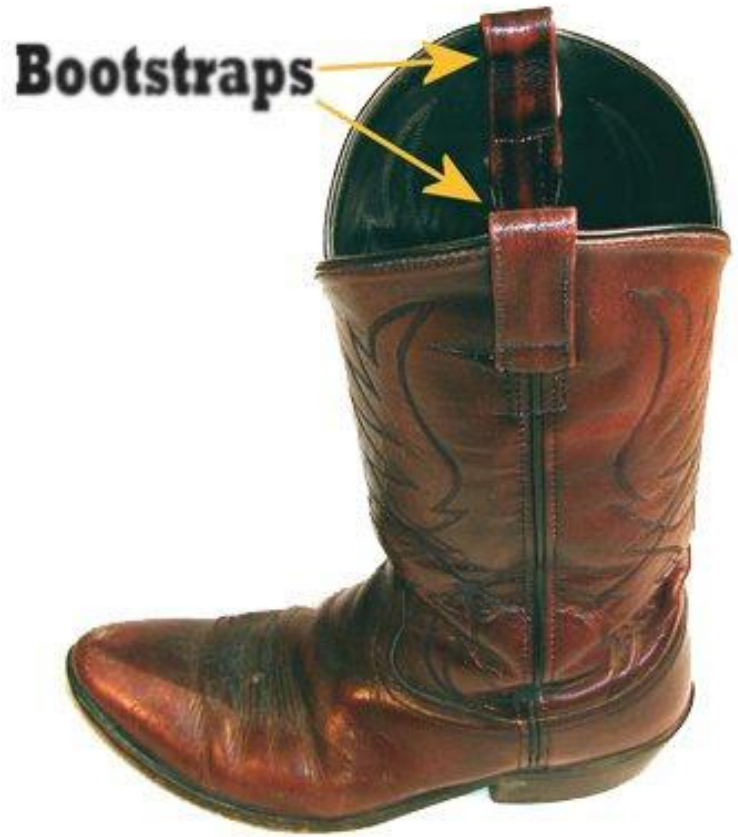
Sampling distributions

Q_{26}: Could we calculate the SE by repeatedly sampling from a population to create sampling distribution, and then take the sd of this sampling distribution?

Sampling distributions

Q_{27}: If we can't calculate the sampling distribution, what else could we do?

The bootstrap



The bootstrap

The bootstrap is a method to estimate the standard error

- $\hat{SE}$ is an estimate for SE
- We will use the symbol SE* as the ***bootstrap estimate*** for SE (rather than $\hat{SE}$)

1. Estimate SE with SE*
2. Then use $\bar{x} \pm 2 \cdot SE^*$ to get the 95% CI

Plug-in principle

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

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

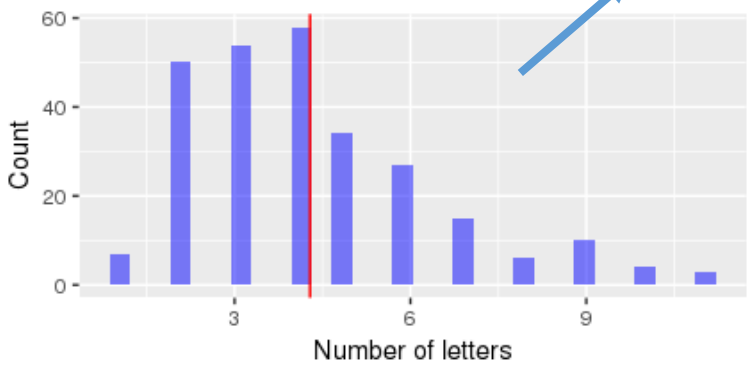
3. The standard deviation of this bootstrap distribution (SE* bootstrap) is a good approximate for standard error SE from the real sampling distribution

Gettysburg address word length bootstrap distribution

The sample (n = 10)

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

μ



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

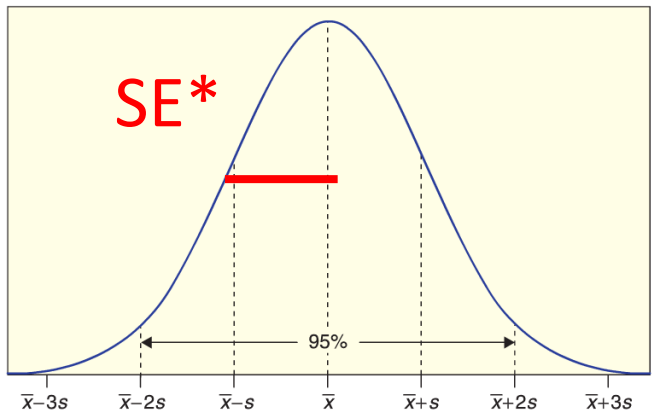
$\bar{x}^* = 4$

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

$\bar{x}^* = 4.1$

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

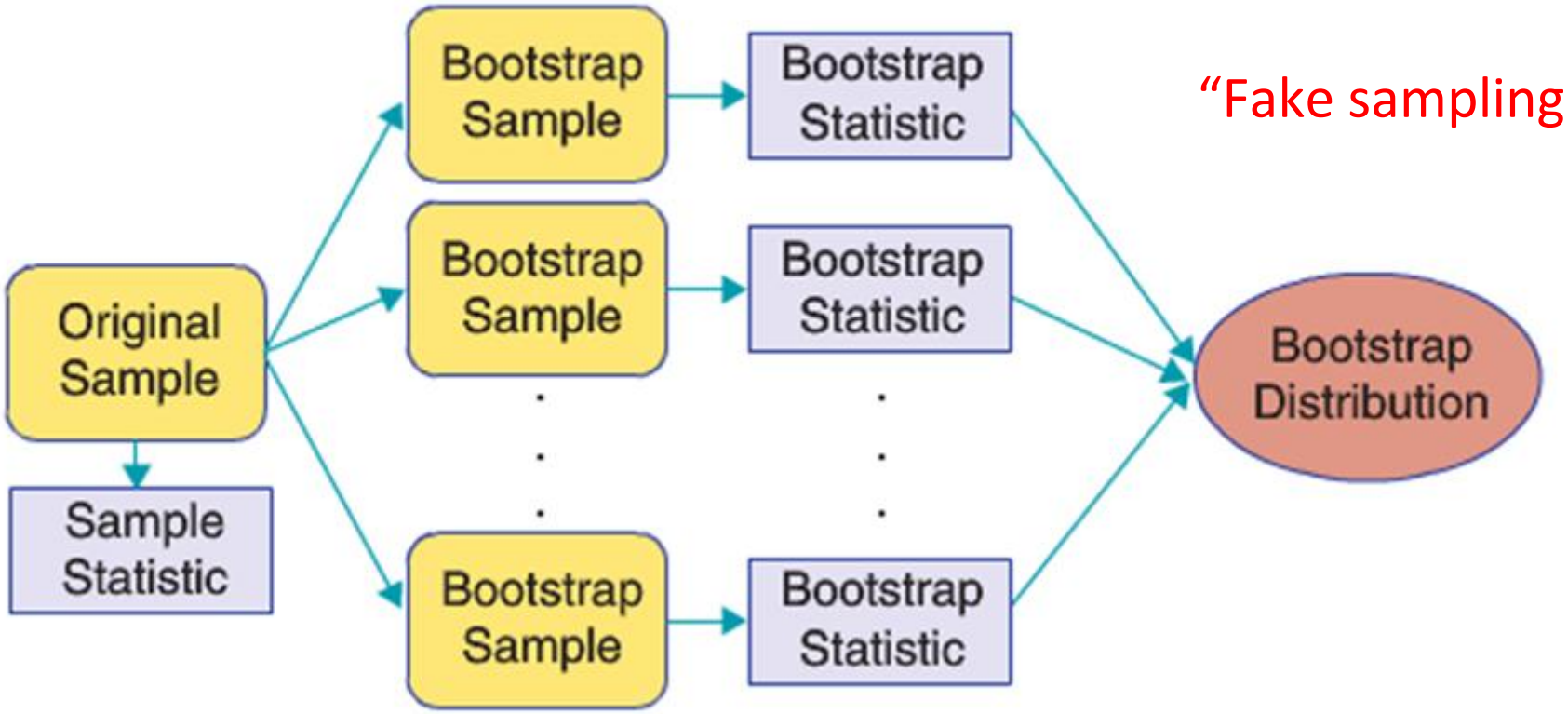
$\bar{x}^* = 3.9$



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

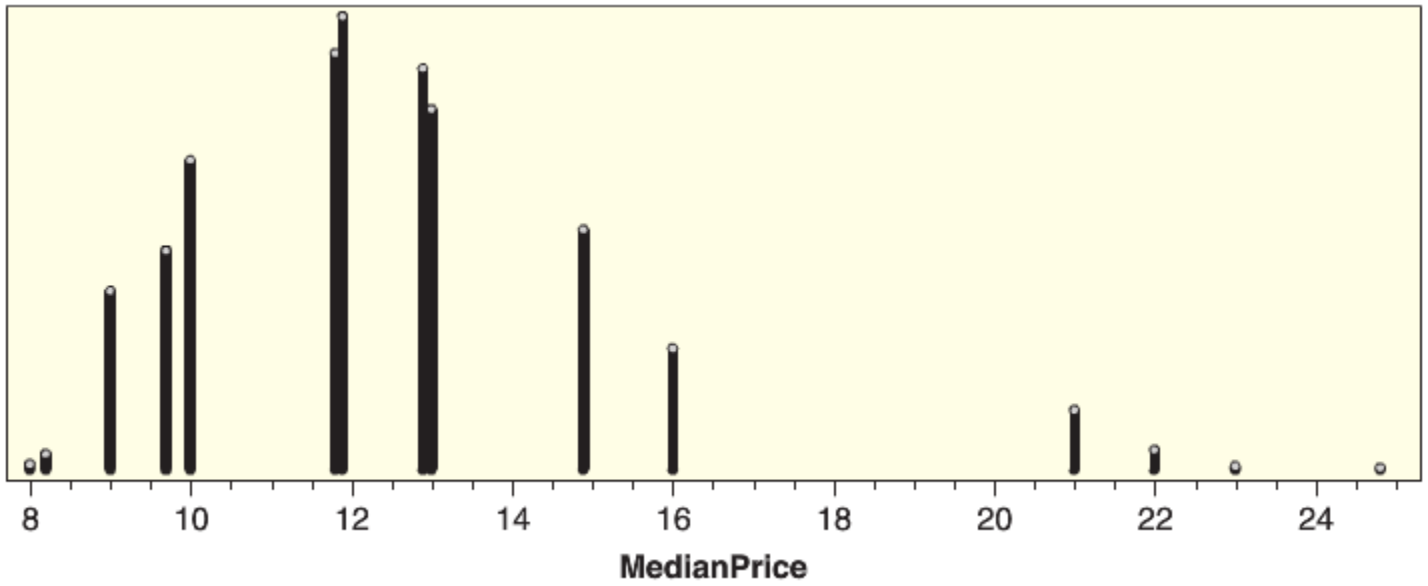
Findings CIs for many different parameters

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



Calculating 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
```