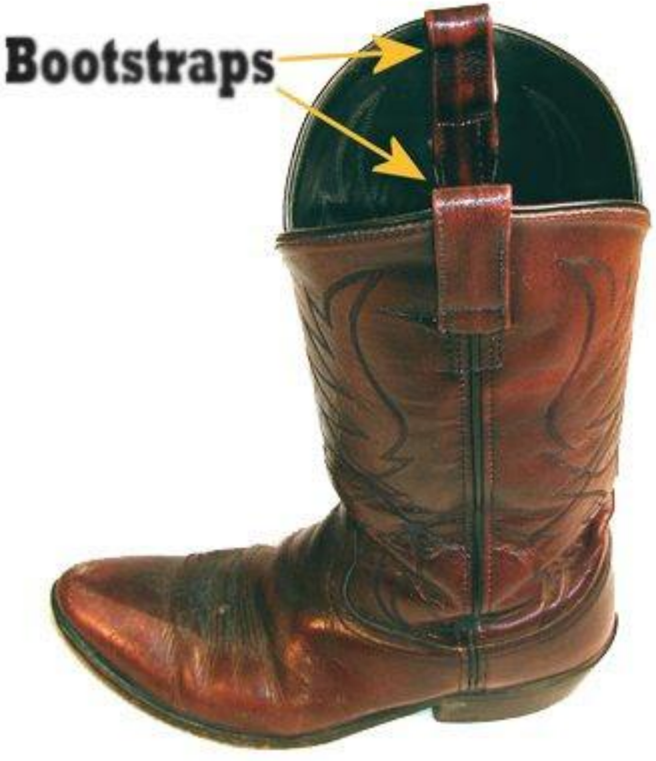
The bootstrap



Overview

Quick review of confidence intervals

The bootstrap

Calculating bootstrap confidence intervals in R

If there is time: Introduction to hypothesis tests

Quick review of confidence intervals

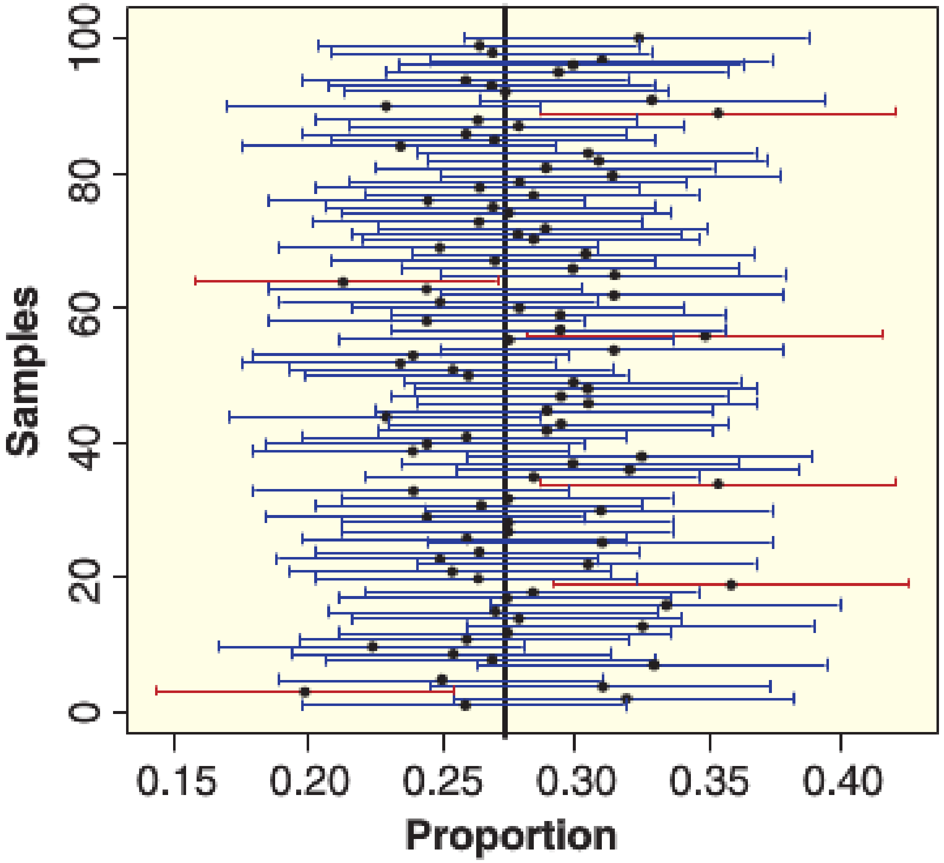
Review: confidence intervals

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

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

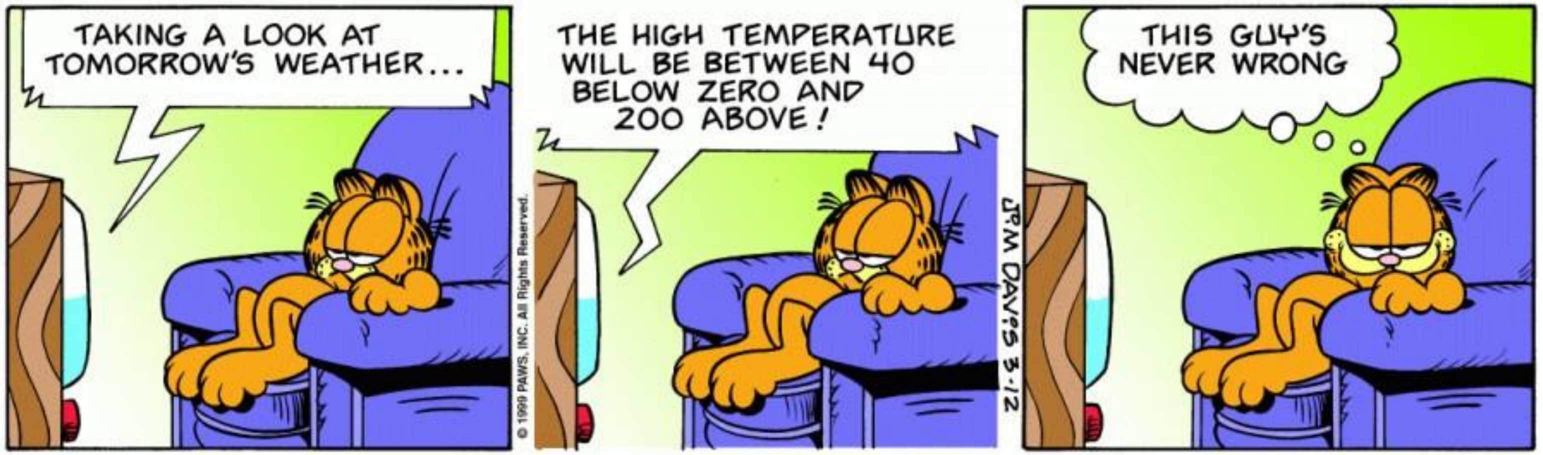
Review: confidence intervals

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



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

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


TAKING A LOOK AT TOMORROW'S WEATHER...
THE HIGH TEMPERATURE WILL BE BETWEEN 40 BELOW ZERO AND 200 ABOVE!
THIS GUY'S NEVER WRONG
© 1999 PAWS, INC. All Rights Reserved.
JPM DAVIS 3-12

There is a tradeoff between:

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

Example

Machowaik et al (1992) made 130 observations of body temperatures were made

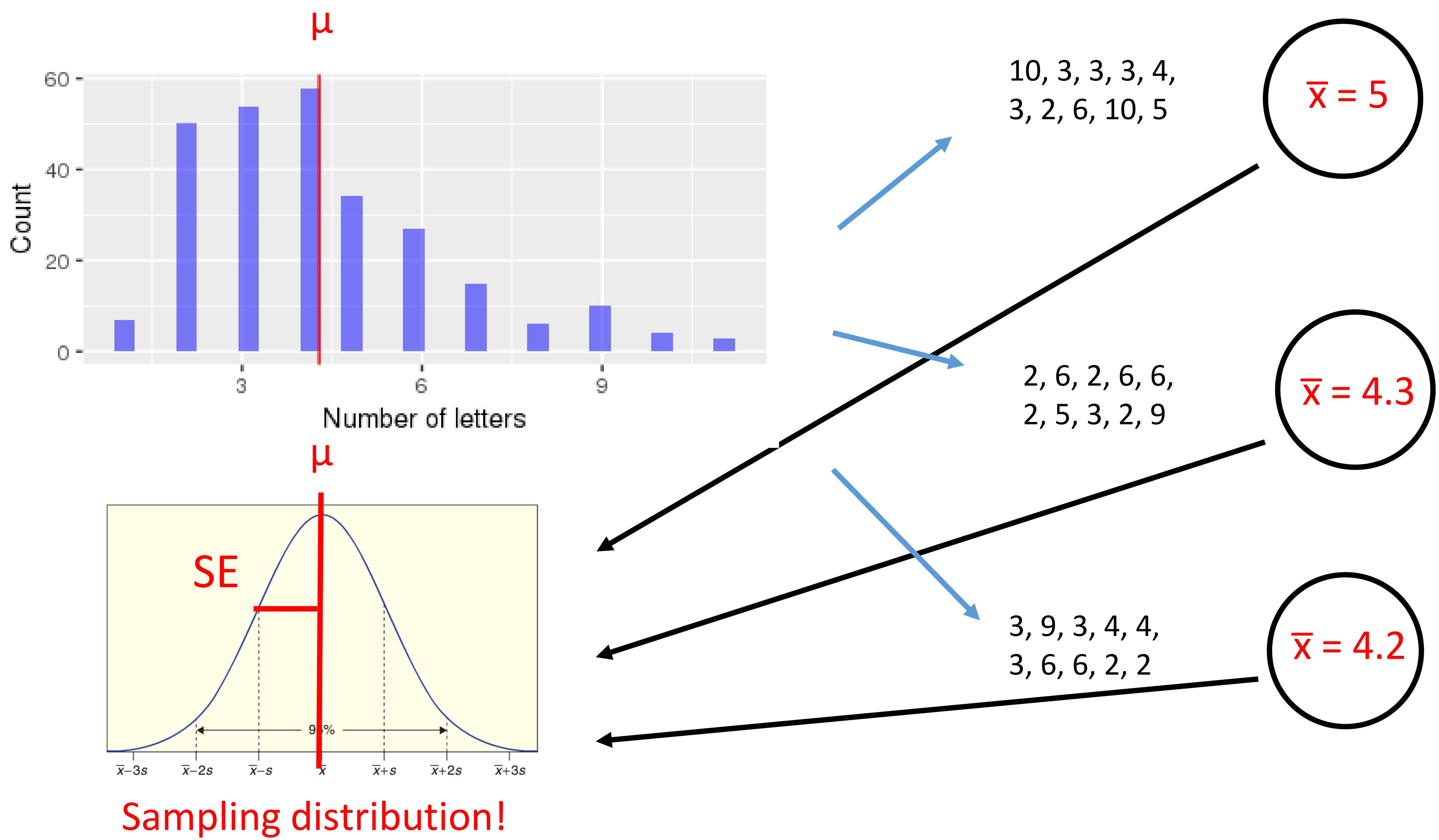
They calculated a 95% confidence interval for the body temperatures to be: [98.12, 98.37]

Q: How do we interpret these results?

Q: Is this what you would expect?

Original paper

Review: sampling distribution illustration



μ

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

$\bar{x} = 4.2$

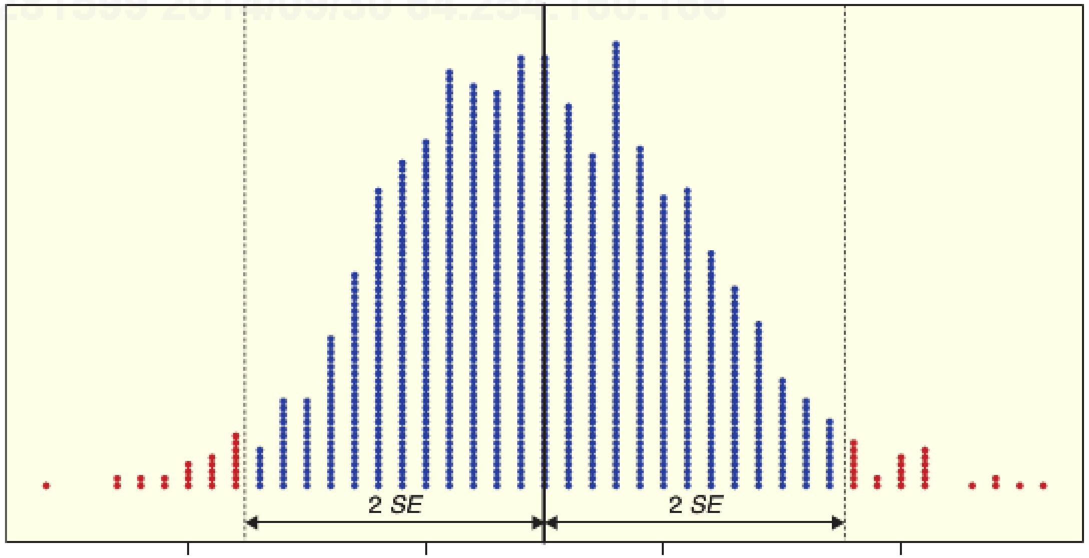
μ

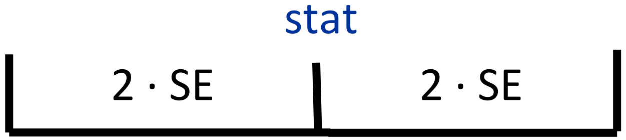
SE

Sampling distribution!

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?




stat

2 · SE 2 · SE

Confidence interval

If we had:

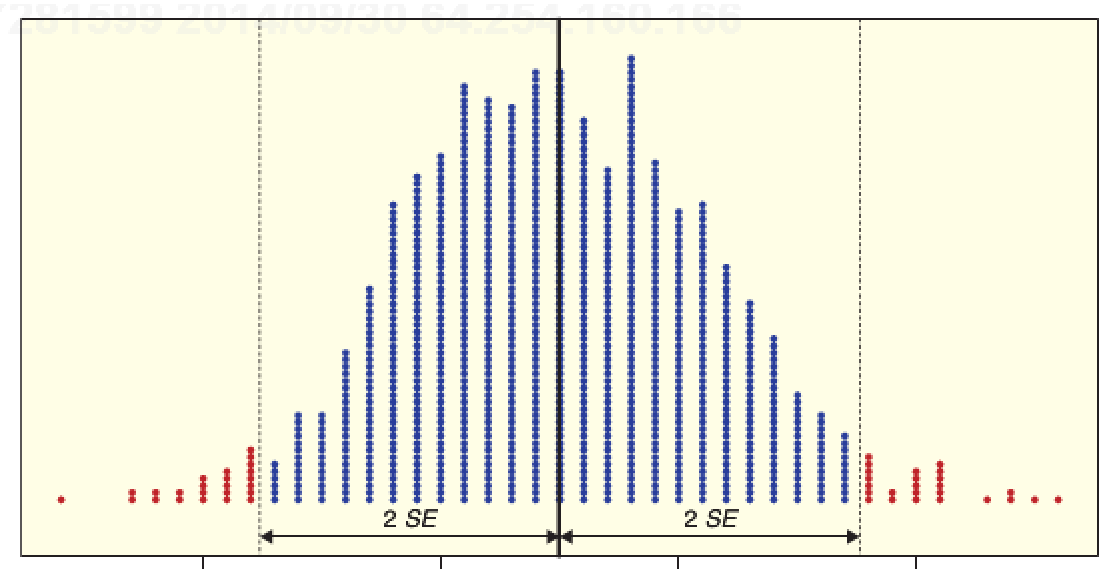
- A statistic value
- The SE

We could compute a 95% confidence interval!

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

Sampling distributions

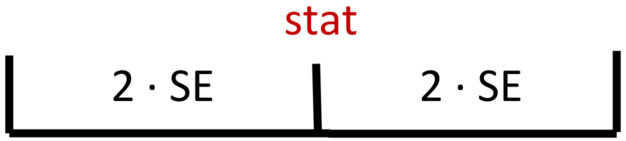
Q: For a sampling distribution that is a normal distribution, what percentage of ***statistics*** lie within 2 standard deviations (SE) for the population mean?



If we had:

- A statistic value
- The SE

We could compute a 95% confidence interval!

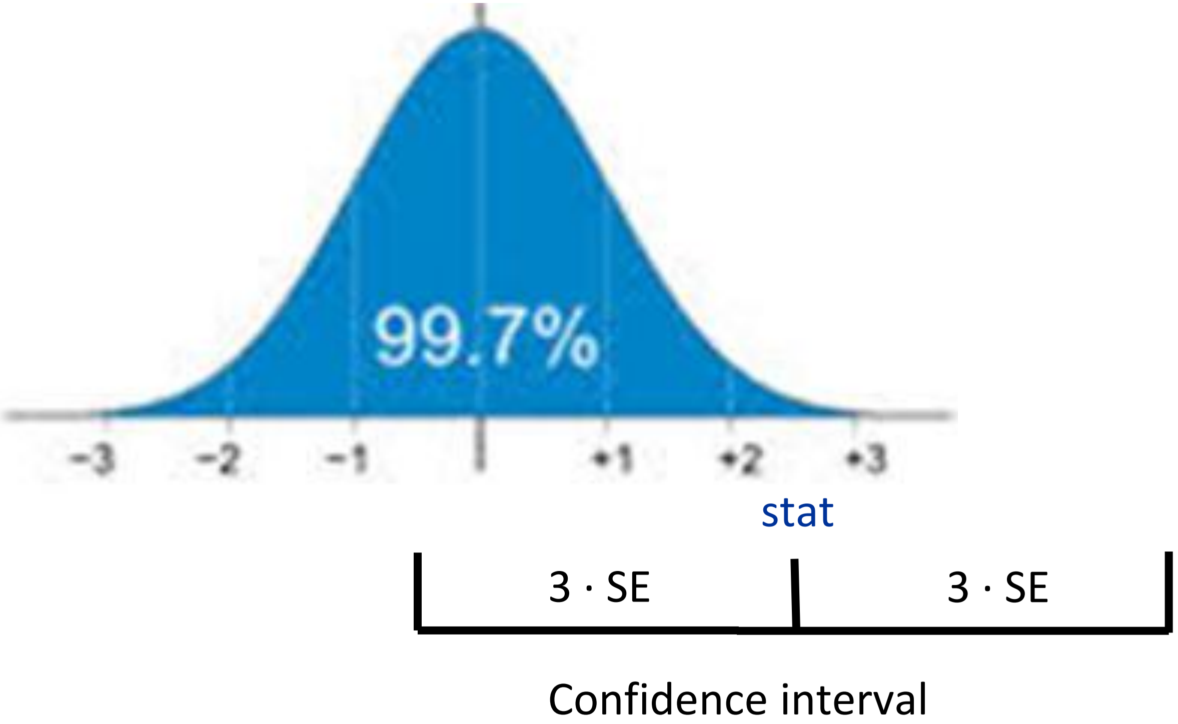


Confidence interval

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

Confidence intervals for other confidence levels

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

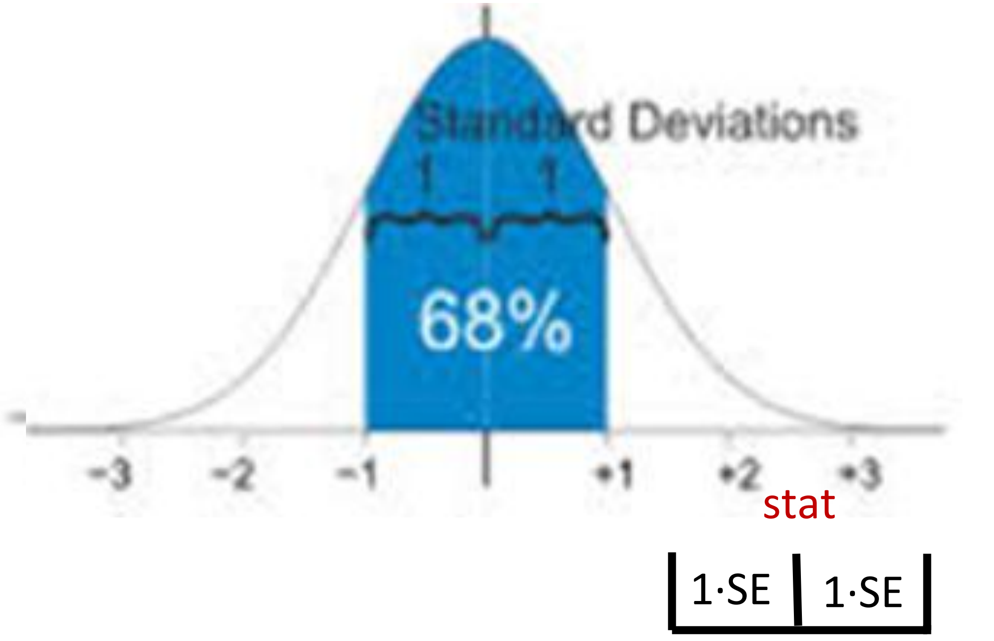


Confidence interval

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

Confidence intervals for other confidence levels

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



stat

1·SE 1·SE

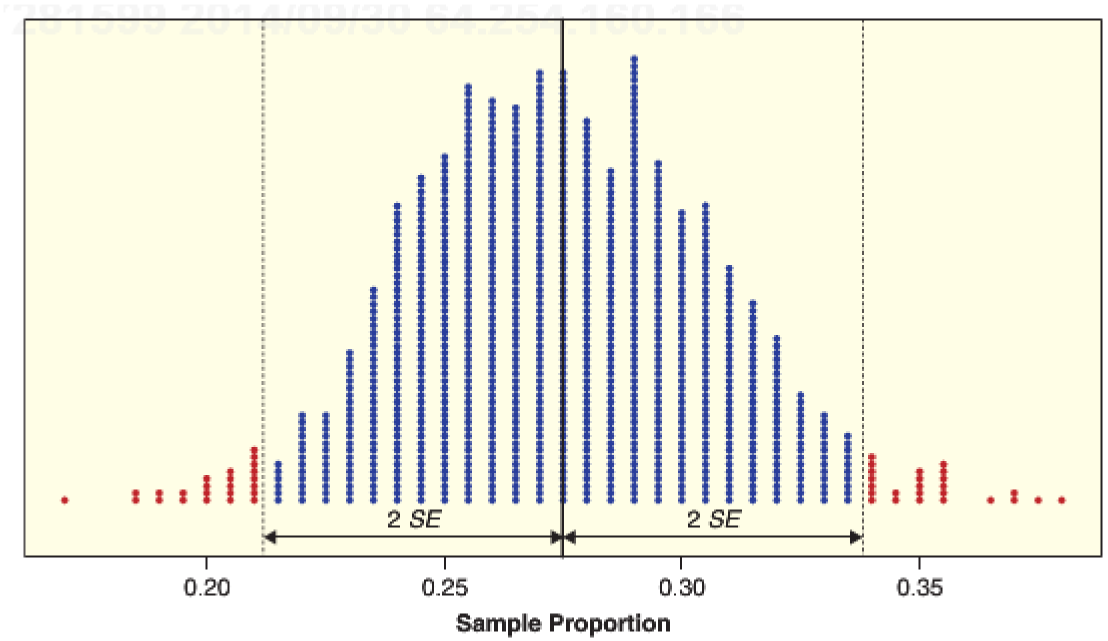
Confidence interval

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

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

Confidence intervals for other confidence levels

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



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

In R:

```
qnorm(0.975)
[1]   1.96
```

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?

Sampling distributions

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

The bootstrap

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 ± $2 \cdot \hat{SE}$ 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)

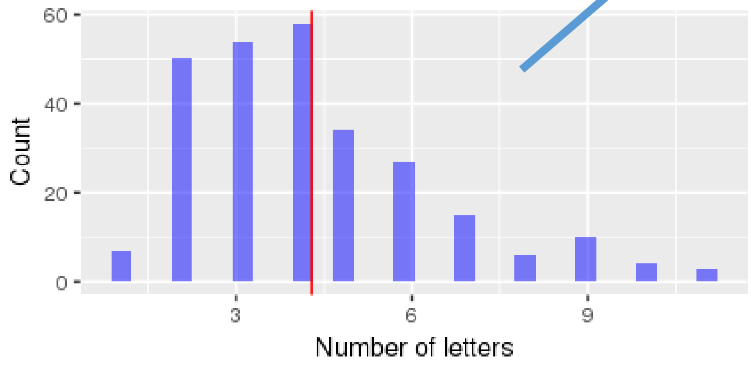
1. We then sample *n* points *with replacement* from our sample, and compute our statistic of interest

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



The sample (n = 10)
10, 3, 3, 3, 4, 3, 2, 6, 4, 5

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

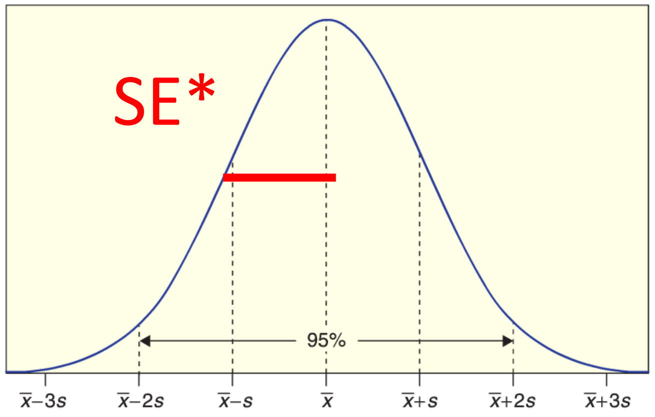
$\bar{x}^* = 4$

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:

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

Where SE* is the standard error estimated using the bootstrap

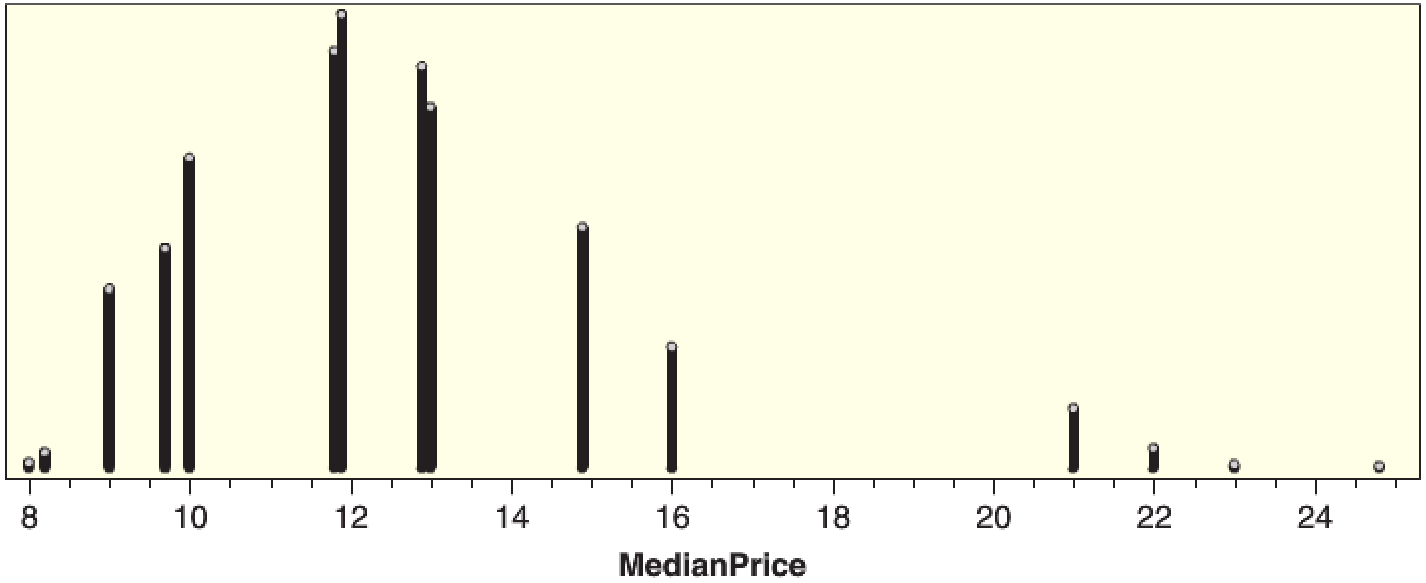
Findings CIs for many different parameters

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 with some real data in R!