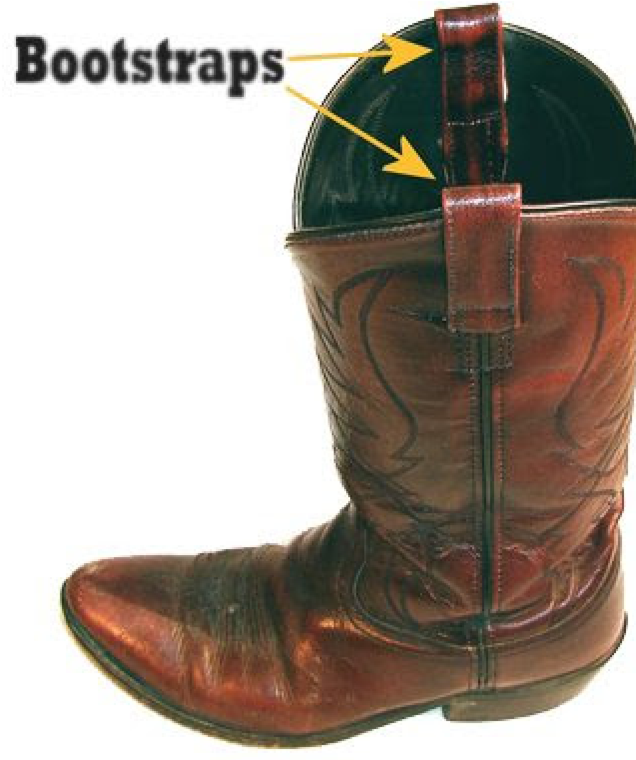
The bootstrap



Overview

Review: confidence intervals and the bootstrap

Calculate bootstrap confidence intervals in R

Quick review of confidence intervals

Review: confidence intervals

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

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

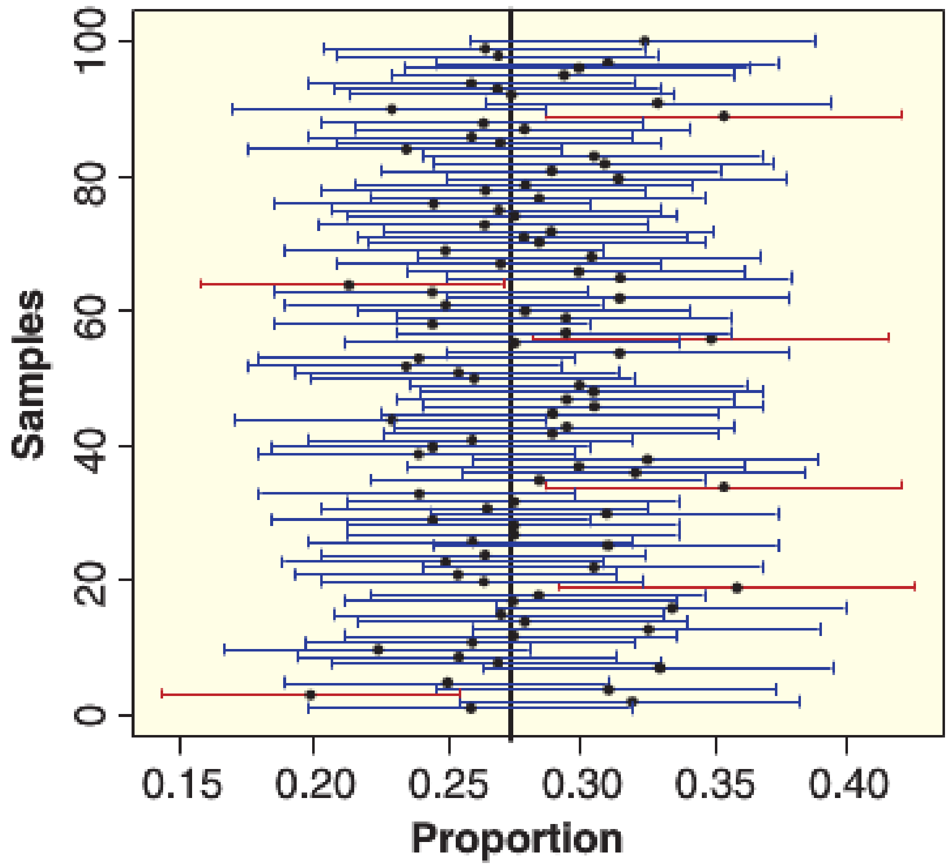
Q: What is the **confidence level**?

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

Review: confidence intervals

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

- A: 90%

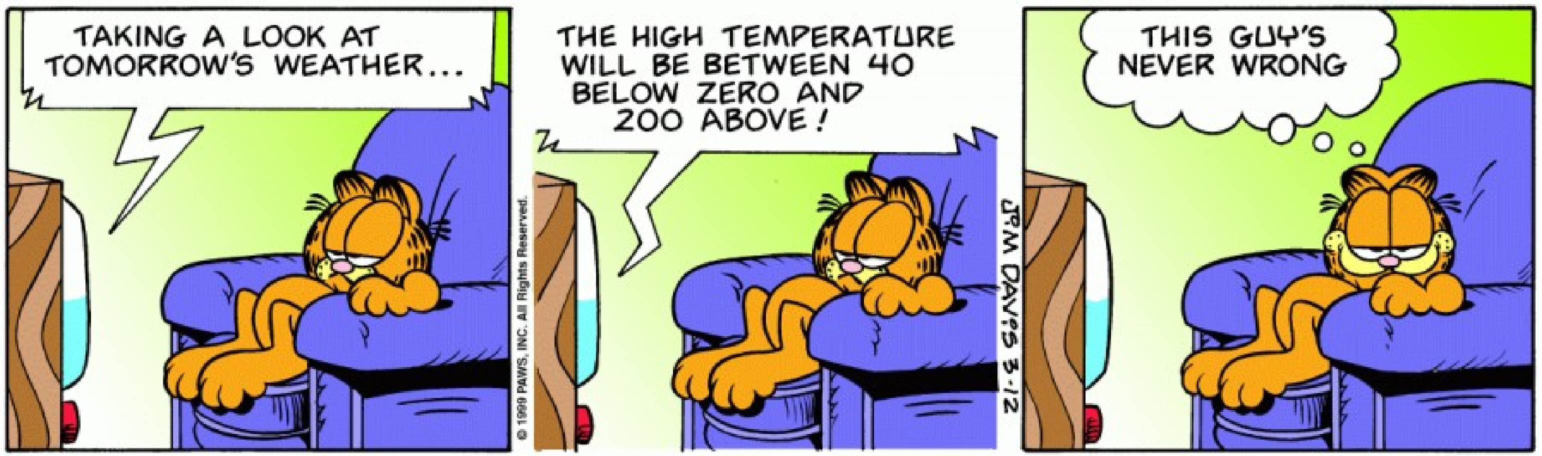


Q: For a given confidence interval, do we know if it contains the parameter?

- A: No! ☹

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

- A: 100%



There is a <u>tradeoff</u> 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:
[98.12, 98.37]

Q: How do we interpret these results?

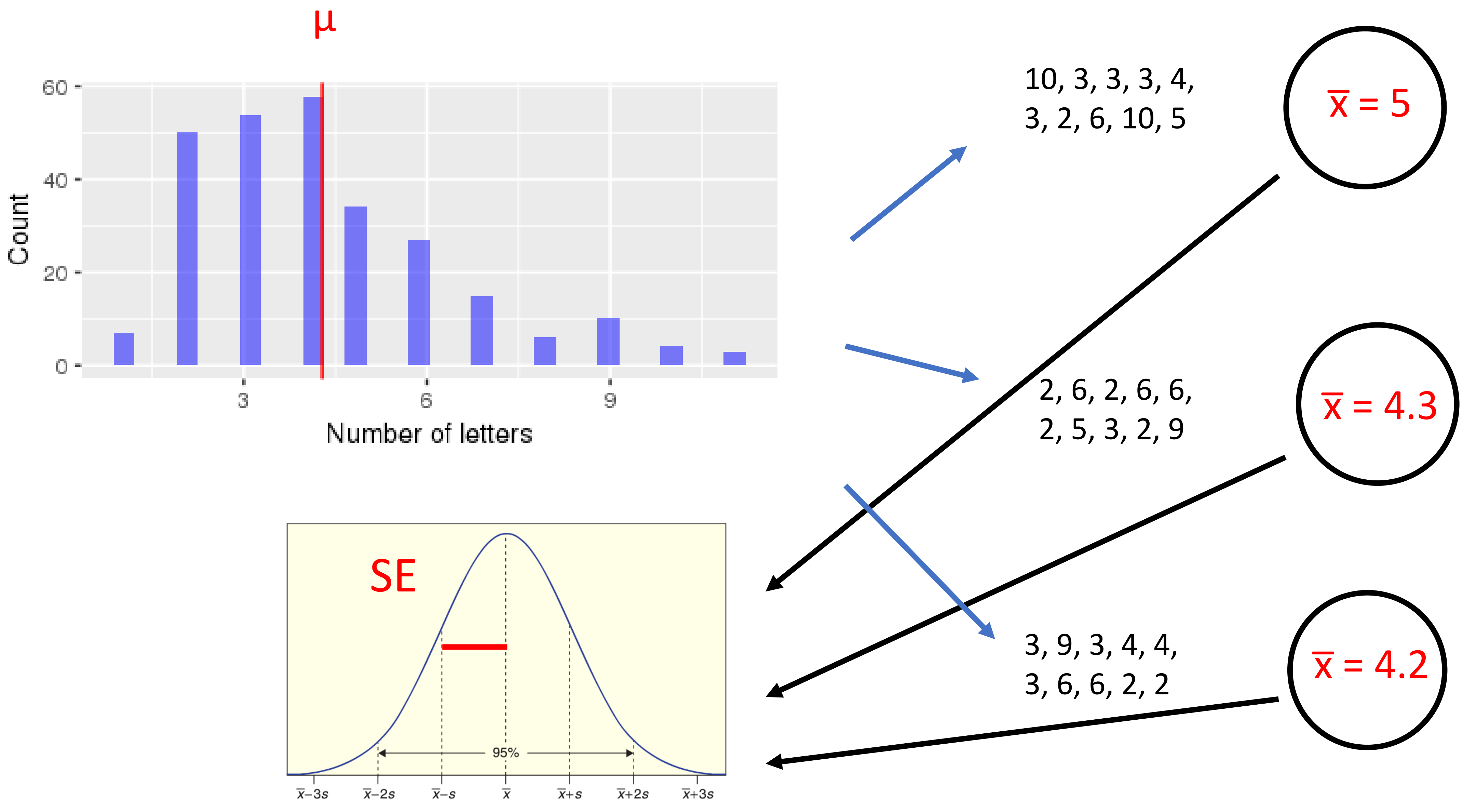
- A: The True average body temperature of humans μ is (likely) between 98.12 and 98.37 degrees

Q: Is this what you would expect?

- A: No, I was told that the average human body temperature is $\mu = 98.6$ degrees
  - It turns out that body temperatures have been decreasing

Original paper

Review: sampling distribution illustration

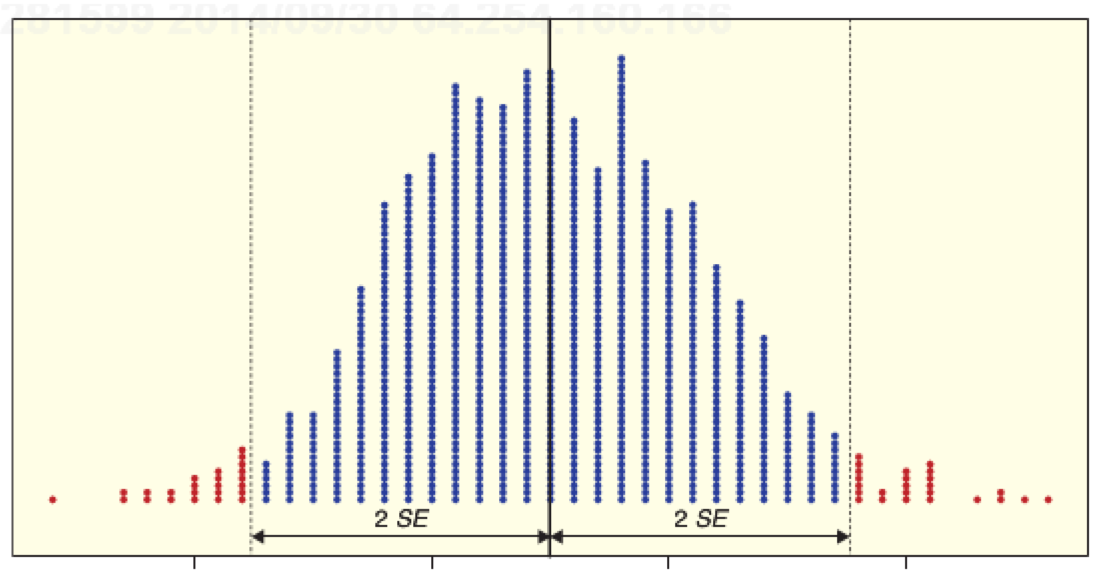


Sampling distribution!

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?

A: 95%



stat

2 · SE 2 · SE

Confidence interval

If we had:

- A statistics value
- The SE

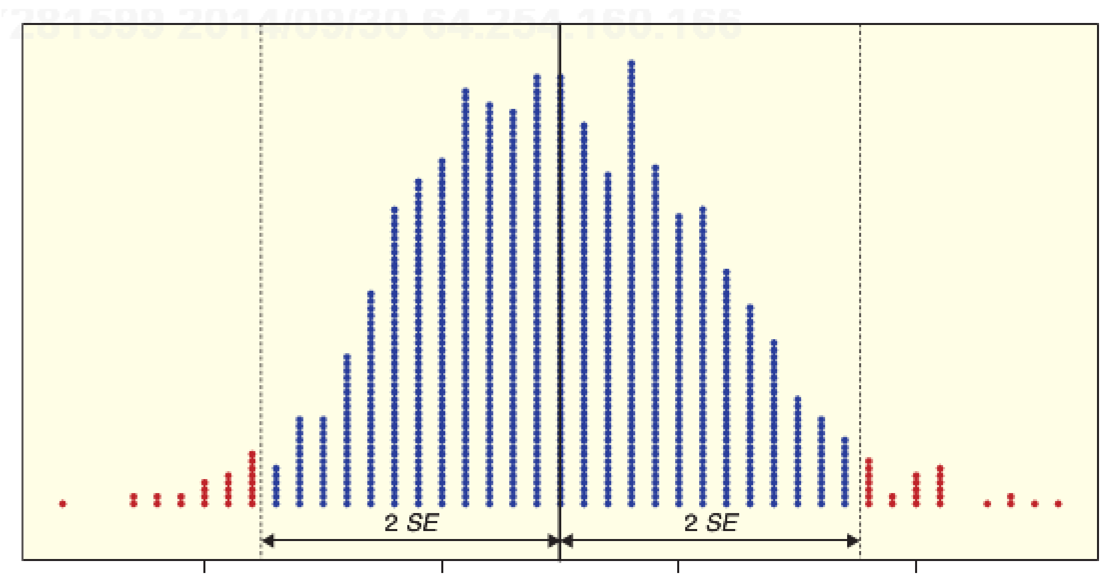
We could compute a 95% confidence interval!

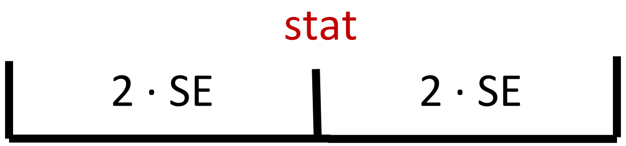
$$CI_{95} = \text{stat} \pm 2 \cdot SE$$

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%





Confidence interval

If we had:

- A statistics value
- The SE

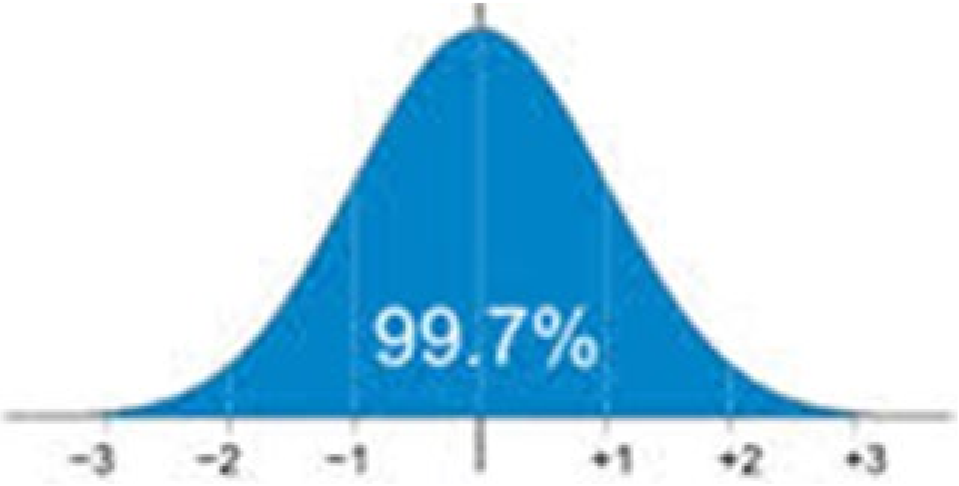
We could compute a 95% confidence interval!

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

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



stat

$3 \cdot SE$ $3 \cdot SE$

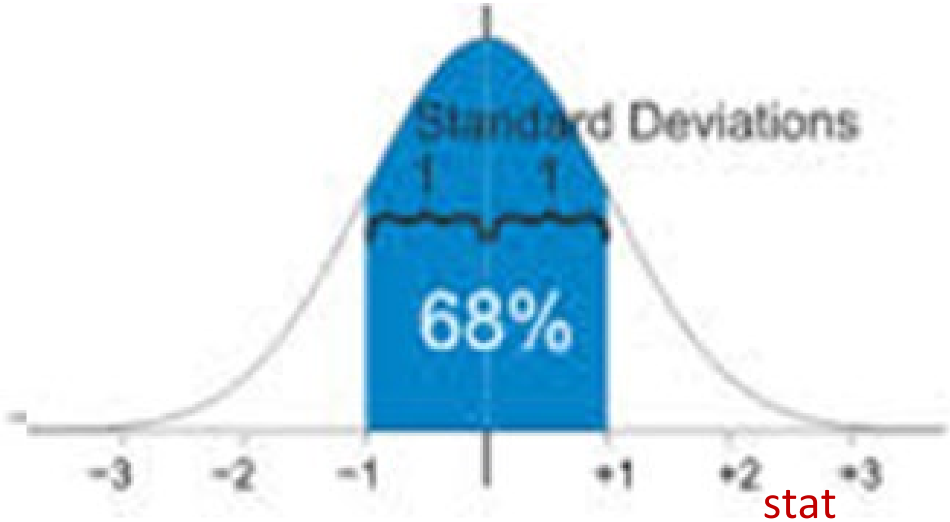
Confidence interval

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

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



stat

1·SE 1·SE

Confidence interval

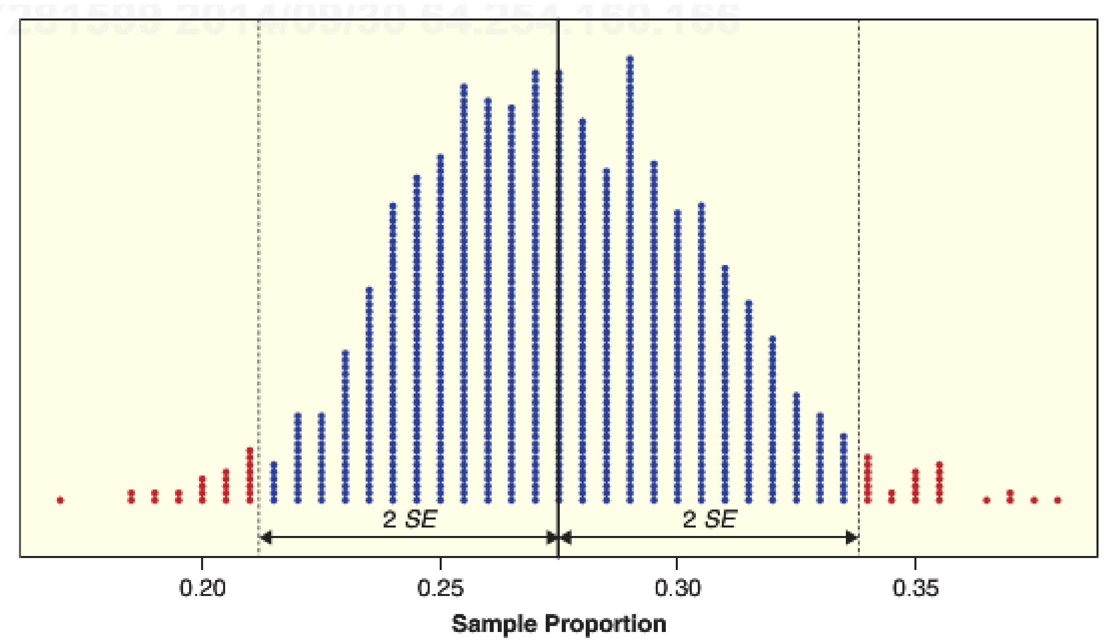
$$CI_{99.7} = \text{stat} \pm 3 \cdot SE$$

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

Confidence intervals for other confidence levels

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

A: We need to find the critical value q* such that q% of our statistics are within ± q* · 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 ± $2 \cdot \hat{SE}$ to get the 95% CI

Plug-in principle

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

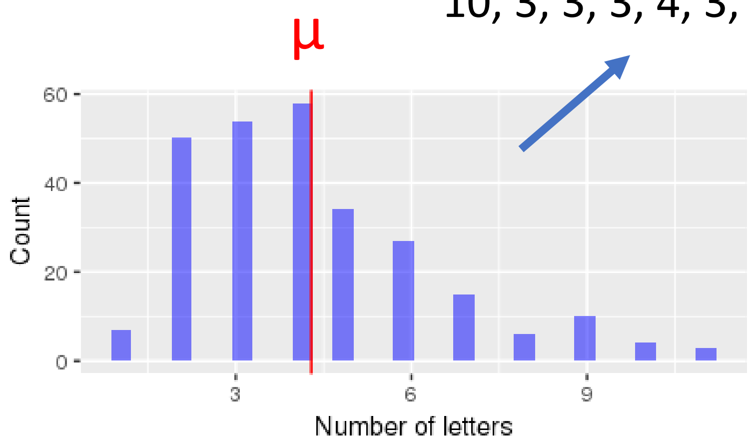
We pretend that *the sample is the population* (plug-in principle)

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



<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

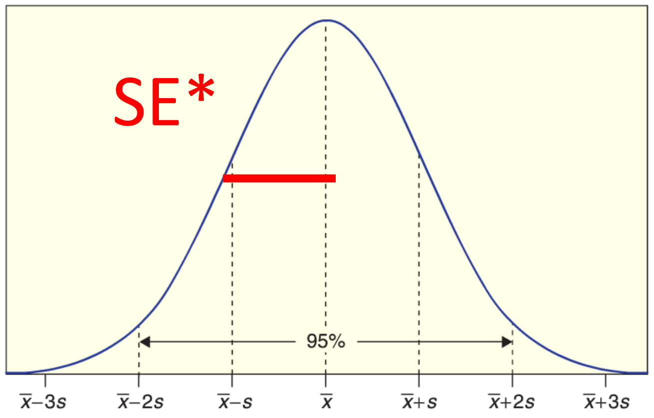
$\overline{x}^* = 4$

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

$\overline{x}^* = 4.1$

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

$\overline{x}^* = 3.9$



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

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!