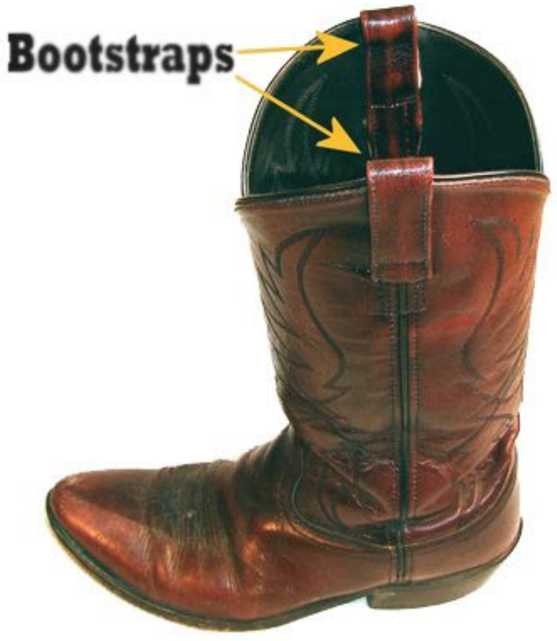
Confidence intervals and the bootstrap



Overview

Density curves and using Q-Q plots to see if data comes from a particular distribution

Confidence intervals conceptual review and wits and wagers answers

The bootstrap to create confidence intervals

Formulas for standard errors and parametric confidence intervals

Questions about anything?

Logistics?

Concepts around the sampling distributions, the bootstrap, etc.

Code?

Review density curves

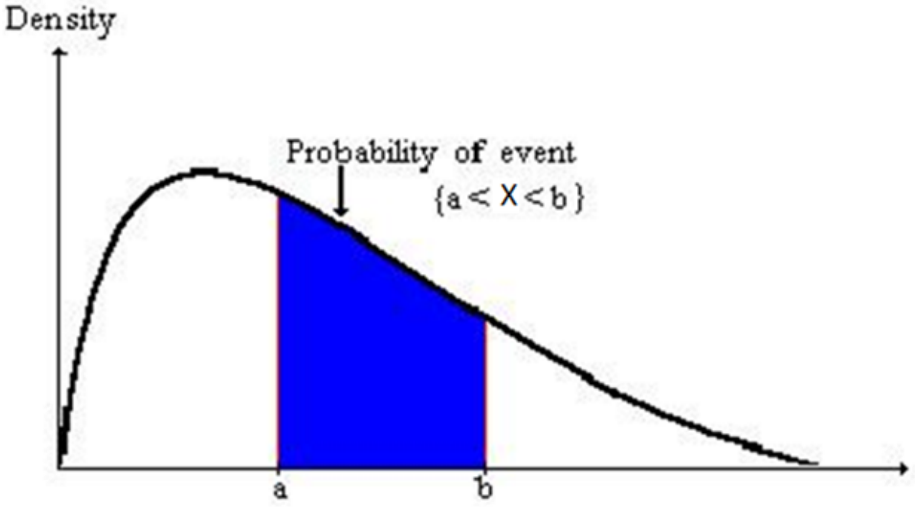
A **density curve** is a mathematical function f(x) that has two important properties:

1. The total area under the curve f(x) is equal to 1

2. The curve is always ≥ 0

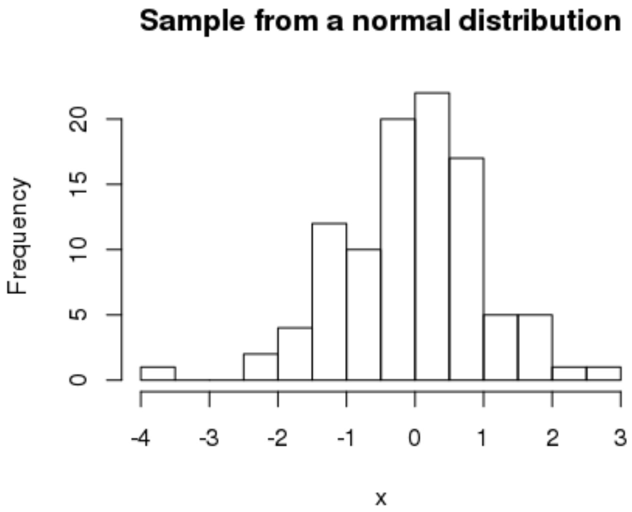
$$P(a < X < b) = \int_a^b f(x)dx$$

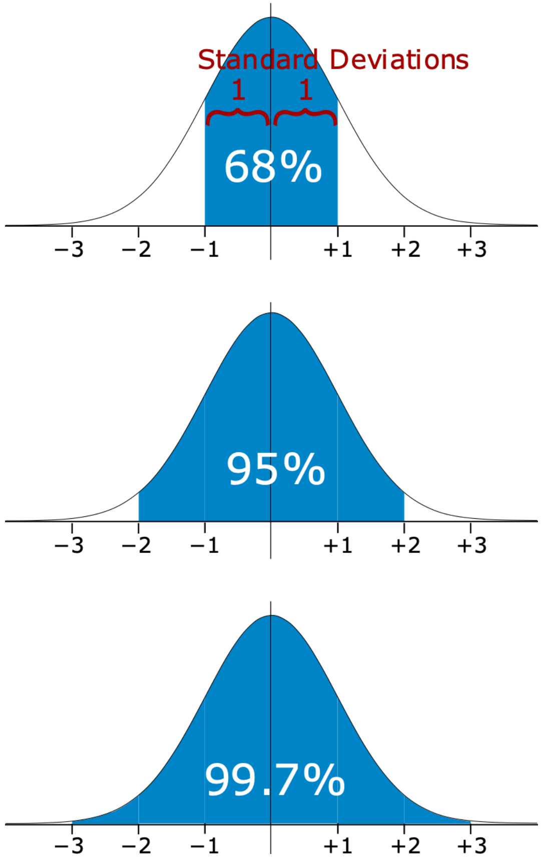
The area under the curve in an interval [a, b] models the probability that a random number X will be in the interval



Normal density function

$$f(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$


Sample from a normal distribution



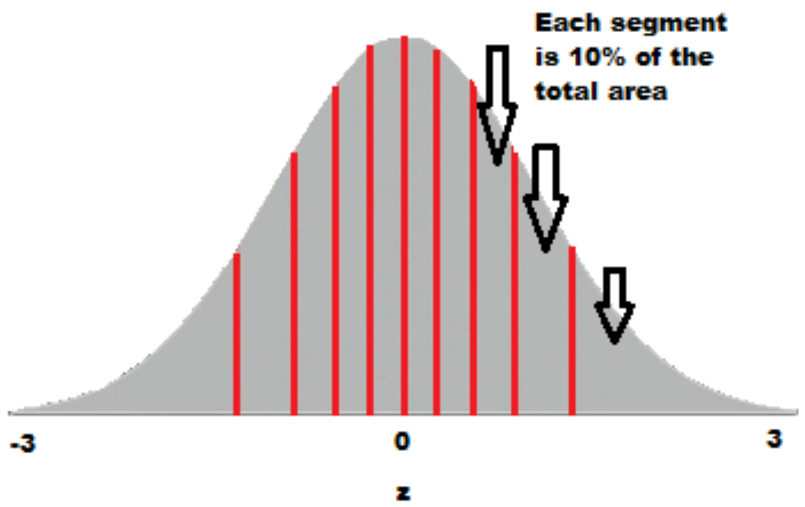
```
dnorm(x, 0, 1)
rand_data <- rnorm(100, 0, 1)
hist(rand_data)
```

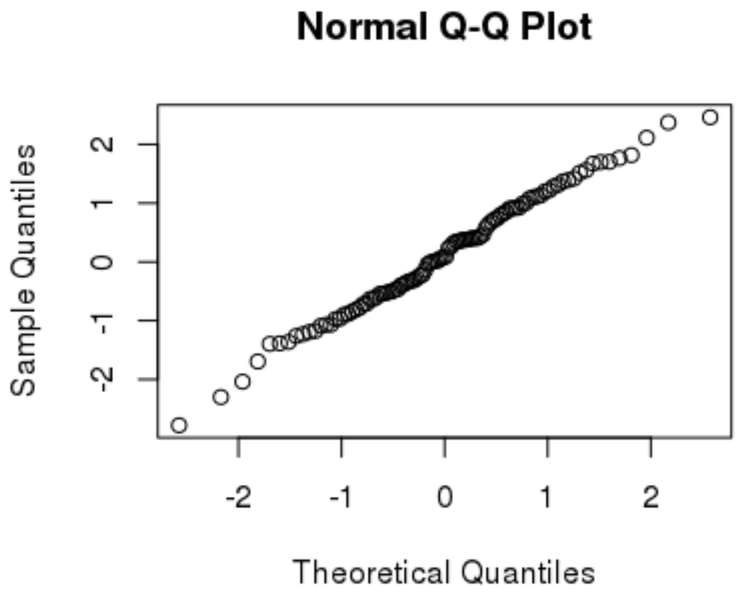
How can you assess whether data comes from a particular distribution?

Quantile-quantile plots (qq-plots)

Quantile-quantile plots (qq-plots) can be used to assess whether a data sample comes from a particular distribution

They plot the observed quantile values from a data sample against the theoretical quantile values from a known distribution





Let’s try it in R…

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

Confidence Intervals

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

- A: 90%

Question

In years, what is the longest recorded life span of a dog?

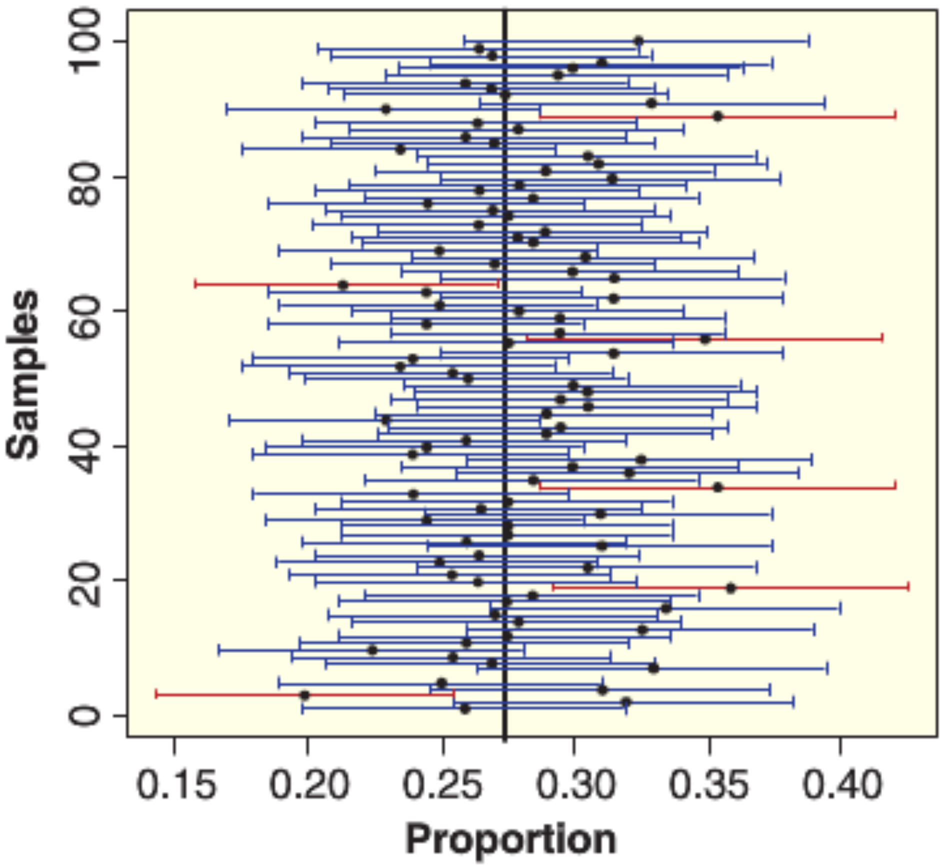
Guinness Book of Records

In miles per hour, what is the fastest recorded speed for a mammal?

University of Michigan Museum of Zoology

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

- A: No! ☹



Wits and Wagers answers

Question 1: What year was Yale University founded?

- 1701

Question 2: In what year Benjamin Franklin prove that lightning was electricity, after flying his kite in a thunderstorm?

- 1752

Question 3: In feet, how tall is The Statue of Liberty including the pedestal?

- 305 Feet

Wits and Wagers…

Question 4: In feet, how tall was the tallest giraffe ever recorded?

- 20 feet

Question 5: In pounds, what was the weight of the heaviest domesticated cat ever recorded?

- 46.81

Question 6: In years, what is the longest recorded life span of a dog?

- 29 (the average dog lives 12 years)

Question 7: How many pounds does one gallon of whole milk weigh?

- 8.6 lbs

Wits and Wagers…

Question 8: If a person weighs 100 pounds on Earth, how many pounds would they weigh on the surface of the moon?

- 16.5 lbs

Question 9: What percentage of American adults say that reading is their favorite leisure-time activity?

- 35%

Question 10: How many cups of coffee does the average American drink per year?

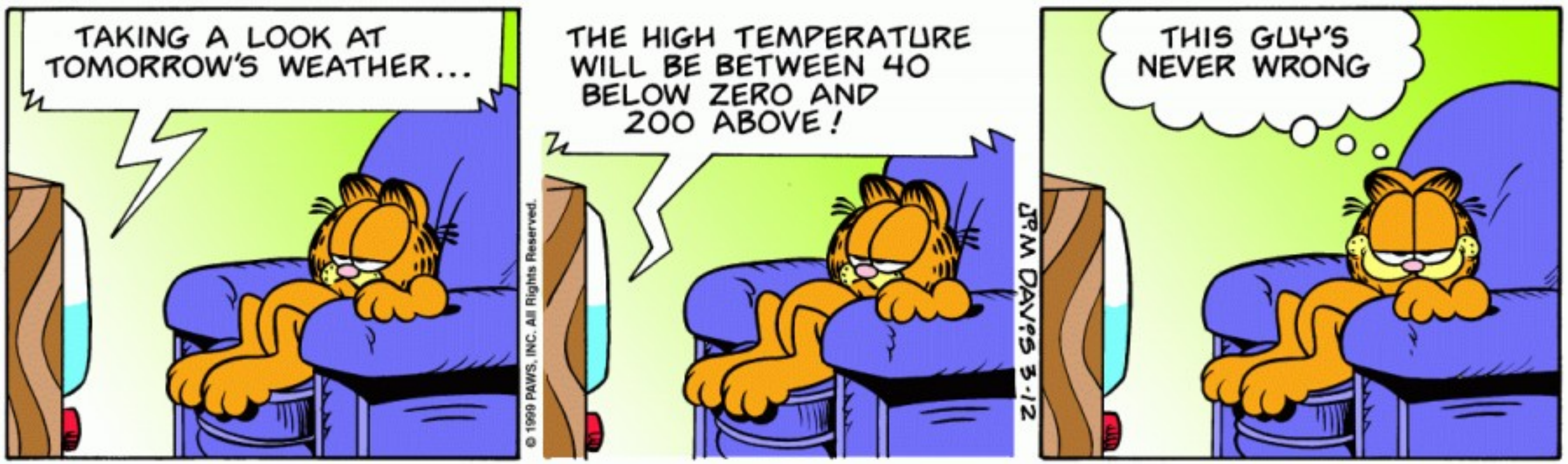
- 393.6

Results!

Please enter in the poll: how many of your intervals captured the correct answer?

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

- A: 100%


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) and the **confidence interval size**

Example

130 observations of body temperature of men were made (°F)

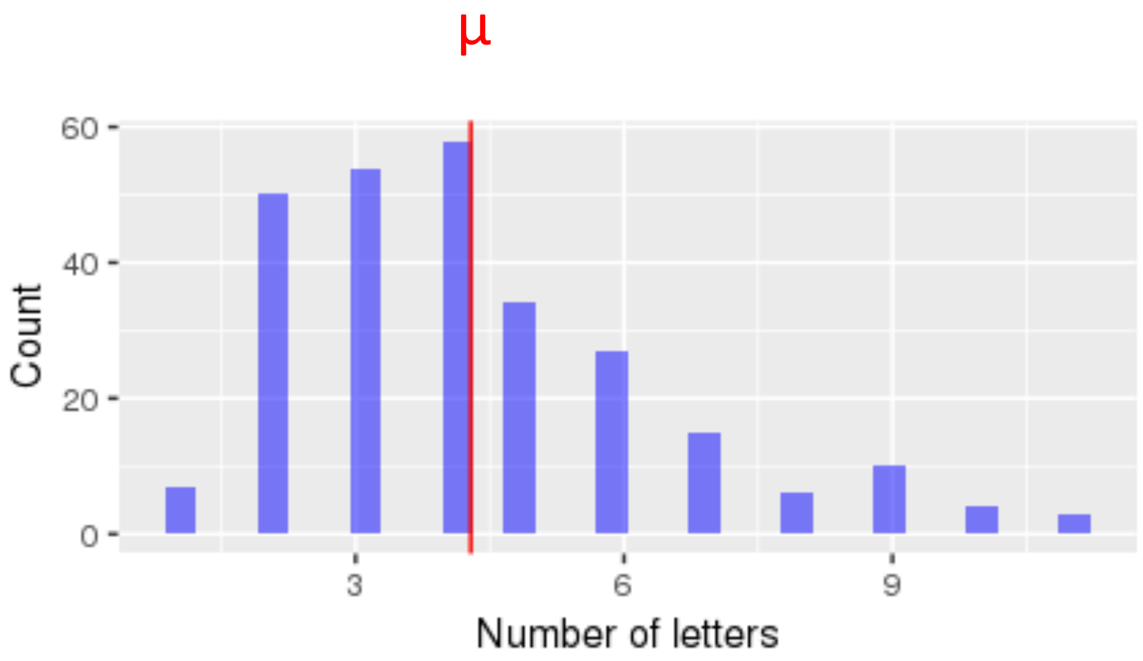
A 95% confidence interval for the mean body temperatures is:
[98.123, 98.375]

How do we interpret these results?

Is this what you would expect?

Original paper Statistic teaching resource

Review: sampling distribution illustration



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

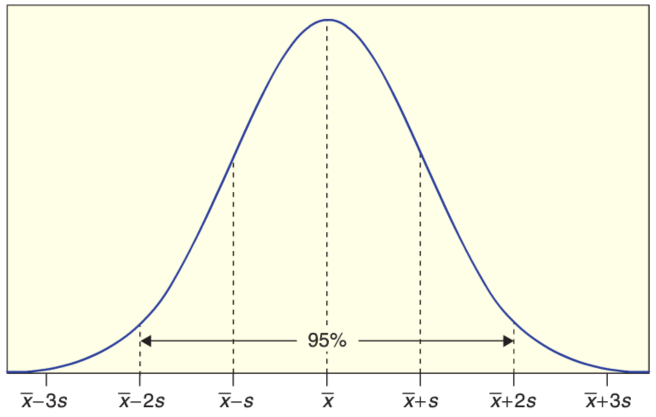
$\bar{x} = 5$

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

$\bar{x} = 4.3$

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

$\bar{x} = 4.2$

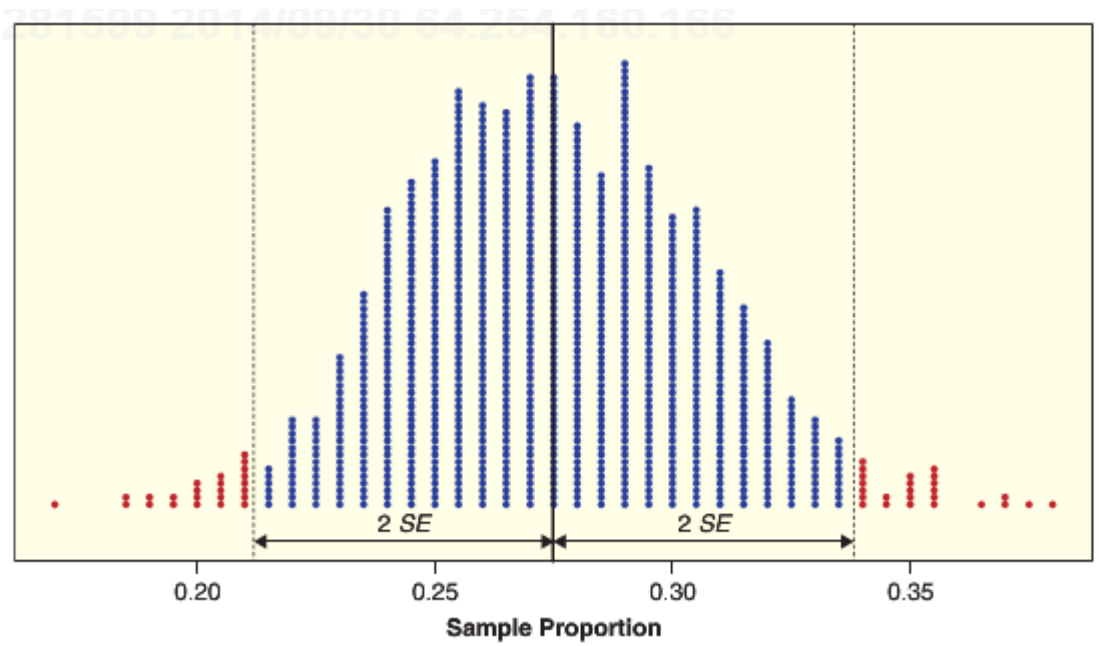


Sampling distribution!

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%



If we had:

- A statistics value
- The SE

We could compute a 95% confidence interval!

Sampling distributions

Unfortunately we can't calculate the sampling distribution ☹

We have to pick ourselves up by the bootstraps!

1. Estimate SE with $\hat{SE}$
2. Then use x̅ ± $2 \cdot \hat{SE}$ to get the 95% CI

Plug-in principle

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

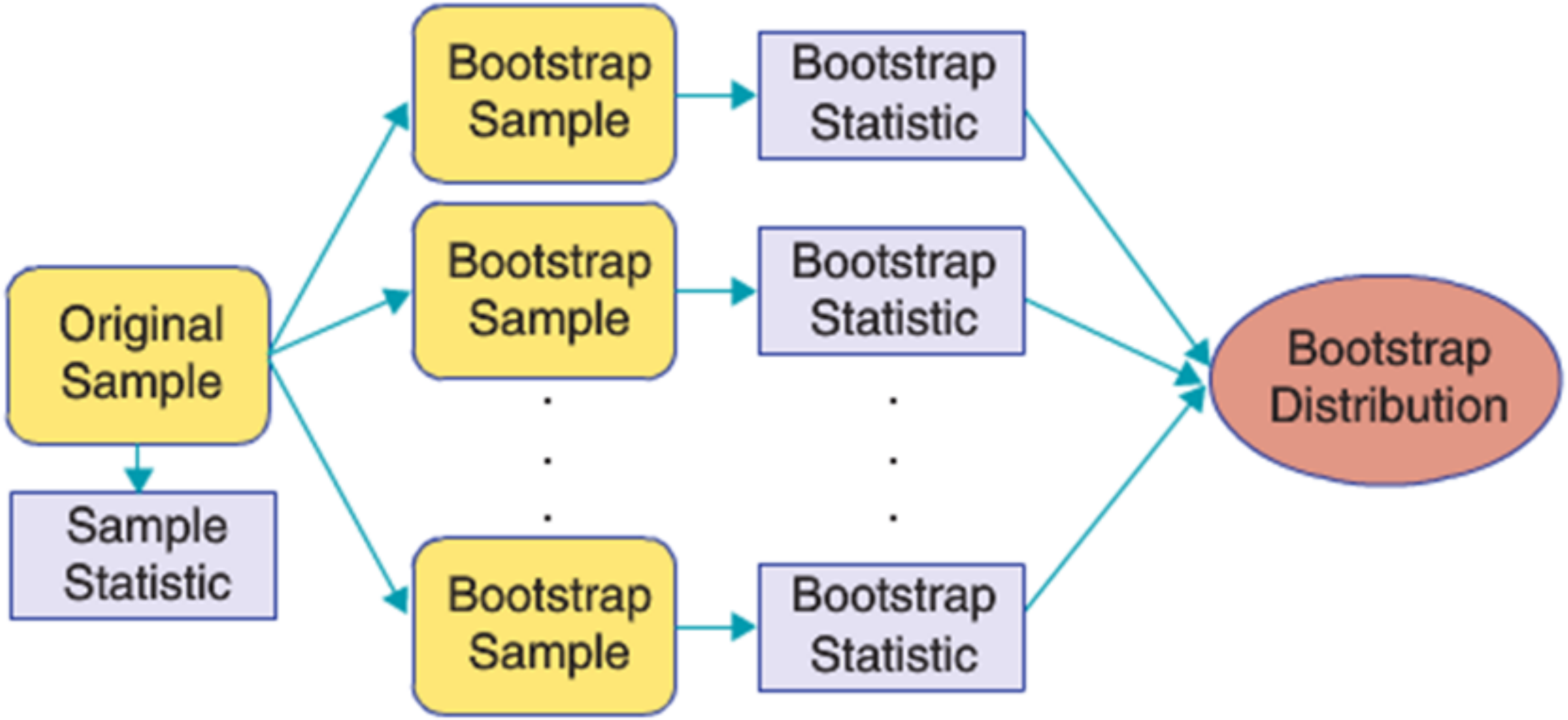
We pretend that *the sample is the population* (plug-in principle)

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

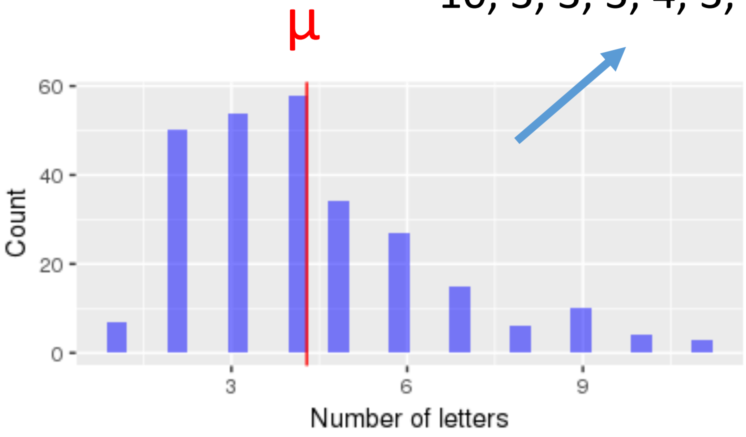
2. We repeat this process 1000's of times and get a ***bootstrap sample distribution***

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

Bootstrap process



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

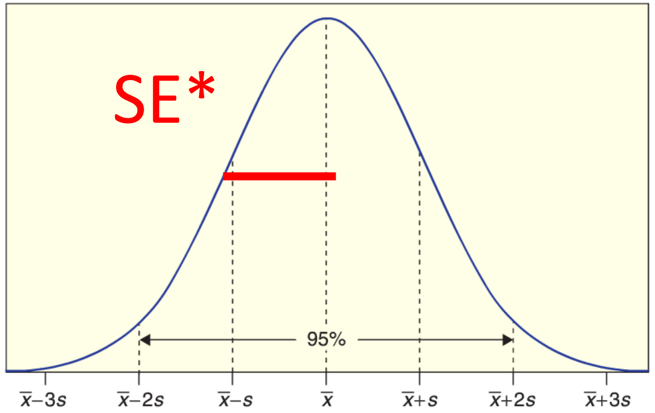
$\bar{x}^* = 4$

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

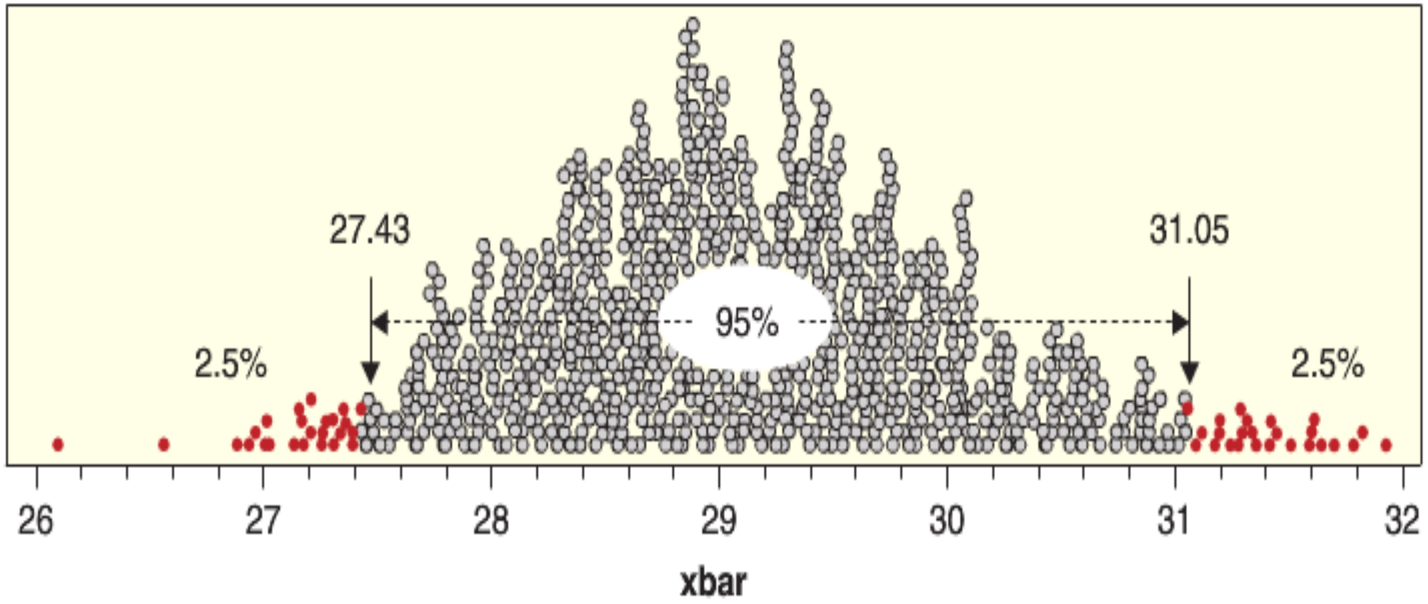
When a bootstrap distribution for a sample statistic is approximately normal, we can estimate a 95% confidence interval using:

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

Where SE* is the standard error estimated using the bootstrap

What if the bootstrap distribution is not normal?

If the bootstrap distribution is approximately symmetric, we can use percentiles in the bootstrap distribution to an interval that matches the desired confidence level.



Findings CIs for many different parameters

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

Let’s try it in R…

Formulas for the standard error of the mean

As you likely learned in intro statistics class, there is formula the **standard error of the mean (SE mean)** which is:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad\qquad s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Where:

- σ is population standard deviation parameter
- n is the sample size
- s is the sample standard deviation

Formula for the standard error of a proportion

Likewise, there is a formula for **standard error of a proportion (SE proportions)** which is:

$$\sigma_{\hat{p}} = \sqrt{\frac{\pi \ \cdot \ (1 - \pi)}{n}} \qquad s_{\hat{p}} = \sqrt{\frac{\hat{p} \ \cdot \ (1 - \hat{p})}{n}}$$

Where:

- π̂ is the population proportion parameter
- n is the sample size
- p̂ is the sample proportion statistic

Next week, hypothesis tests…