Statistics for Hackers

Jake VanderPlas

PyCon 2016

Statistics is Hard.

Statistics is Hard.

Using programming skills,
it can be easy.

My thesis today:

If you can write a for-loop, you can do statistics

Warm-up:
Coin Toss

You toss a coin **30** times and see **22** heads. Is it a fair coin?

A fair coin should show 15 heads in 30 tosses. This coin is biased.

Even a fair coin could show 22 heads in 30 tosses. It might be just chance.

Classic Method:

Assume the Skeptic is correct: test the *Null Hypothesis*.

What is the probability of a fair coin showing 22 heads simply by chance?

Classic Method:

$N_H = 22,\ N_T = 8$

Start computing probabilities . . .

$$P(H) = \frac{1}{2}$$

$$P(HH) = \left(\frac{1}{2}\right)^2$$

Classic Method:

$N_H = 22,\ N_T = 8$

$$P(HHT) = \left(\frac{1}{2}\right)^3$$

$$P(2H,\ 1T) = P(HHT) + P(HTH) + P(THH)$$

$$= \frac{3}{8}$$

Classic Method:

$N_H = 22,\ N_T = 8$

$$P(N_H, N_T) = \binom{N}{N_H}\left(\frac{1}{2}\right)^{N_H}\left(1-\frac{1}{2}\right)^{N_T}$$

Number of arrangements (binomial coefficient)

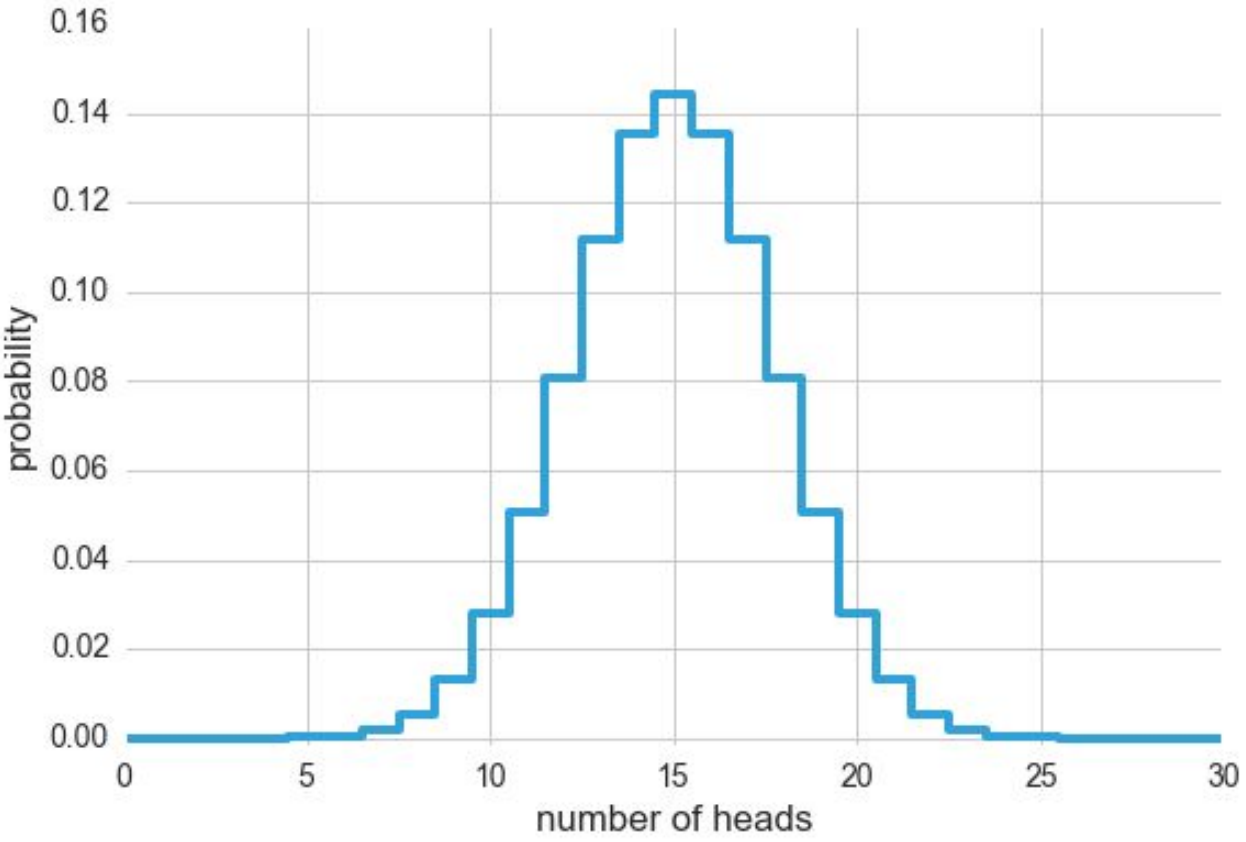
Probability of N_H heads

Probability of N_T tails

Classic Method:

$$N_H = 22,\ N_T = 8$$

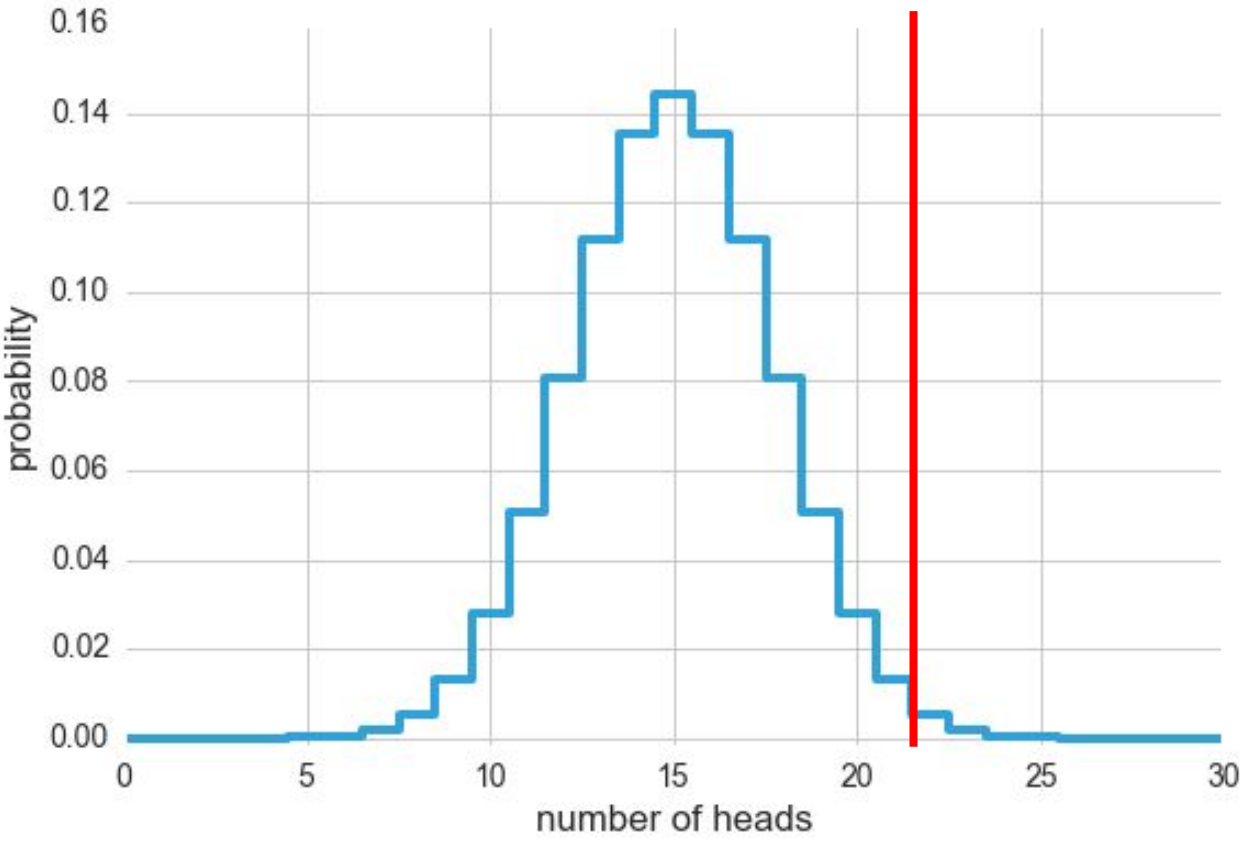
$$P(N_H, N_T) = \binom{N}{N_H}\left(\frac{1}{2}\right)^{N_H}\left(1-\frac{1}{2}\right)^{N_T}$$



Classic Method:

$$N_H = 22,\ N_T = 8$$

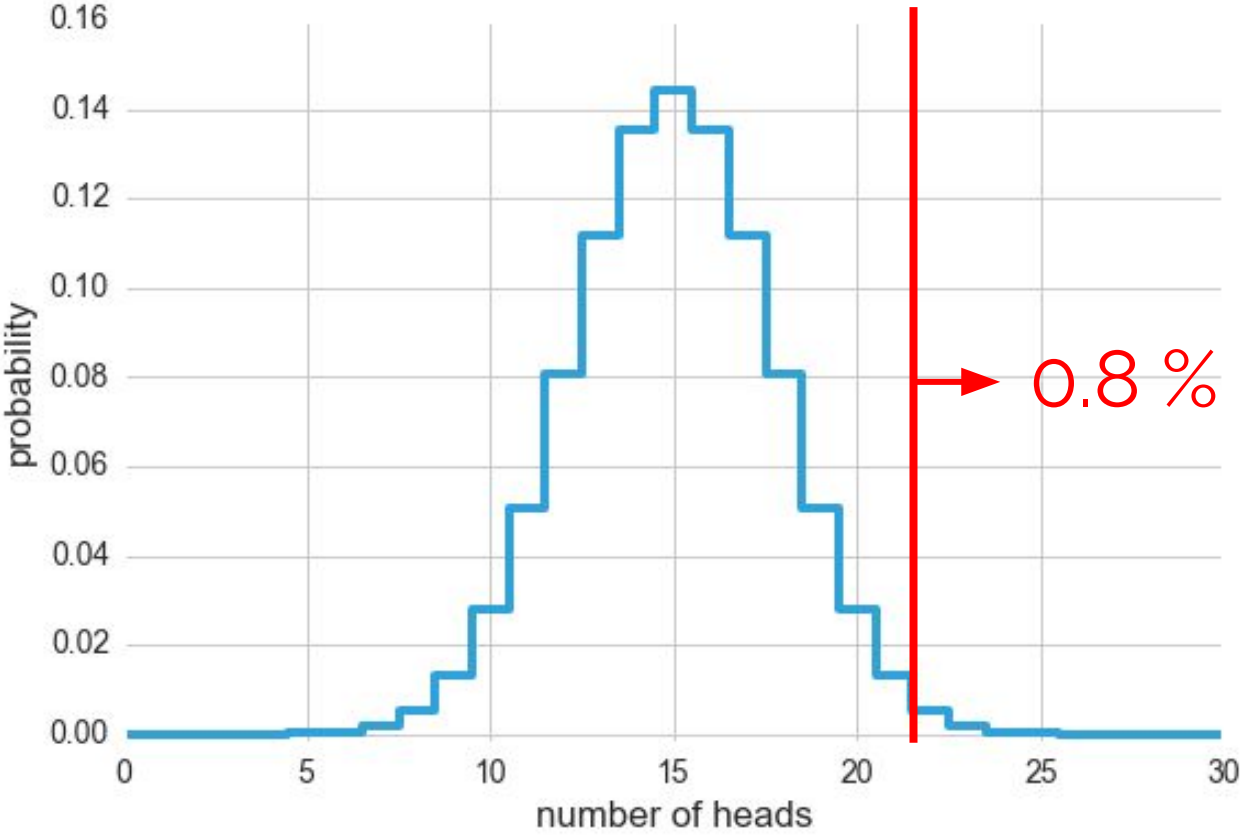
$$P(N_H, N_T) = \binom{N}{N_H}\left(\frac{1}{2}\right)^{N_H}\left(1-\frac{1}{2}\right)^{N_T}$$



Classic Method:

$$N_H = 22,\ N_T = 8$$

$$P(N_H, N_T) = \binom{N}{N_H}\left(\frac{1}{2}\right)^{N_H}\left(1-\frac{1}{2}\right)^{N_T}$$

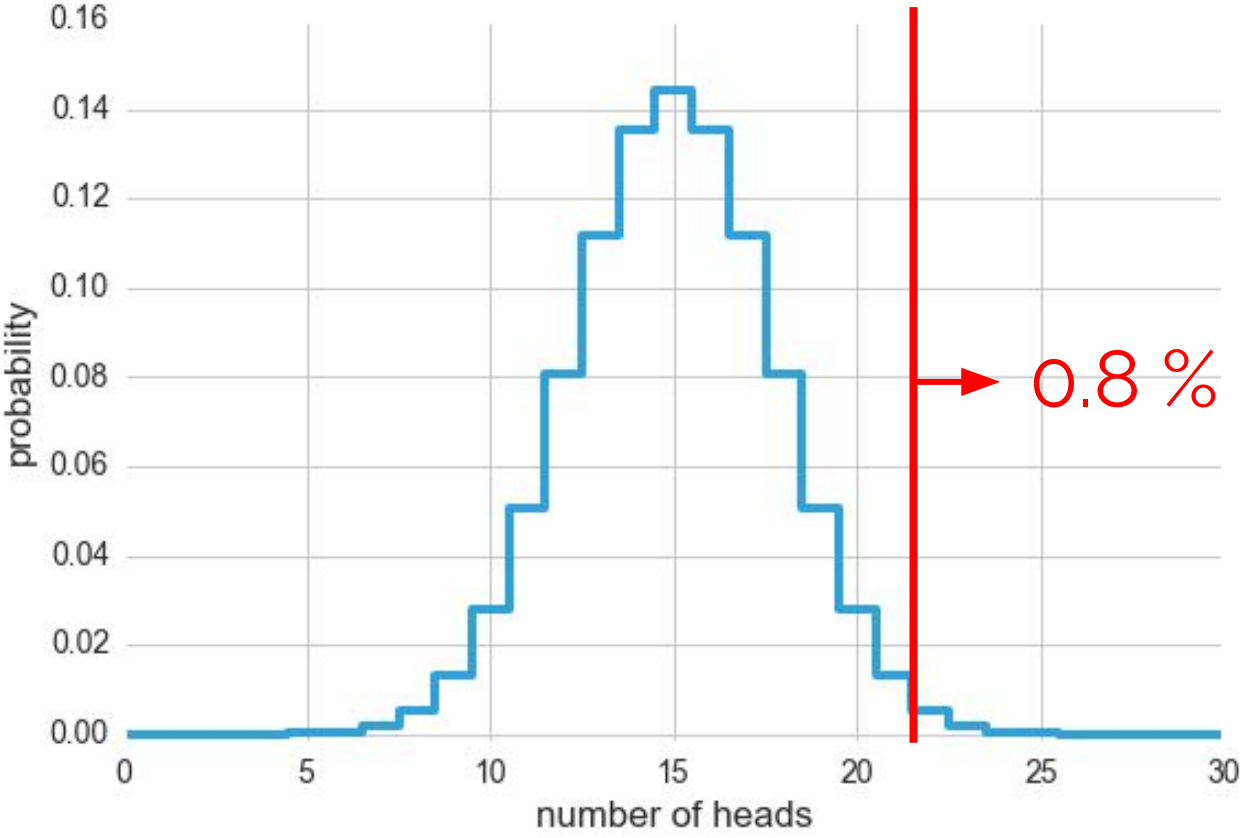


Classic Method:

$N_H = 22,\ N_T = 8$

Probability of 0.8% (i.e. p = 0.008) of observations given a fair coin.

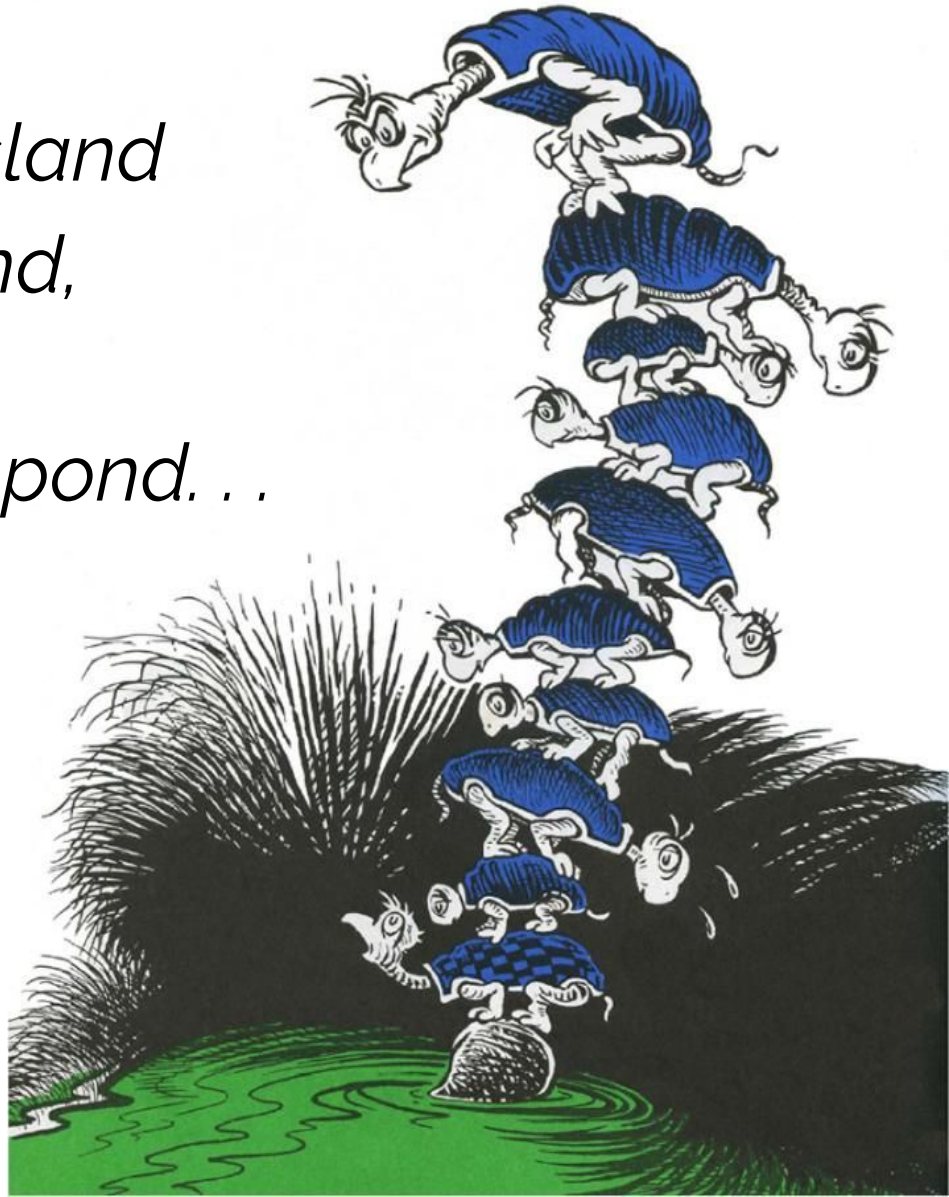
→ **reject fair coin hypothesis at p < 0.05**



Could there be an easier way?

Yertle's Turtle Tower

On the far-away island
of Sala-ma-Sond,
Yertle the Turtle
was king of the pond. . .



How High can Yertle stack his turtles?

Observe 20 of Yertle's turtle towers . . .

# of turtles										
	48	24	32	61	51	12	32	18	19	24
	21	41	29	21	25	23	42	18	23	13

- What is the mean of the number of turtles in Yertle's stack?
- What is the uncertainty on this estimate?

Classic Method:

Sample Mean:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = 28.9$$

Standard Error of the Mean:

$$\sigma_{\bar{x}} = \frac{1}{\sqrt{N}}\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2} = 3.0$$

What assumptions go into these formulae?

Can we use sampling instead?

Problem:
As before, we don't have a generating model . . .

Problem:
As before, we don't have a generating model . . .

Solution:
Bootstrap Resampling

Bootstrap Resampling:

48	24	51	12
21	41	25	23
32	61	19	24
29	21	23	13
32	18	42	18

Idea:
Simulate the distribution by *drawing samples with replacement.*

Motivation:
The data estimates its own distribution – we draw random samples from this distribution.

Bootstrap Resampling:

48	24	51	12
21	41	25	23
32	61	19	24
29	21	23	13
32	18	42	18

Idea:
Simulate the distribution by *drawing samples with replacement.*

Motivation:
The data estimates its own distribution – we draw random samples from this distribution.

Bootstrap Resampling:

48	24	51	12
21	41	25	23
32	61	19	24
29	21	23	13
32	18	42	18

Idea:
Simulate the distribution by *drawing samples with replacement.*

Motivation:
The data estimates its own distribution – we draw random samples from this distribution.

21									

Bootstrap Resampling:

48	24	51	12
21	41	25	23
32	61	19	24
29	21	23	13
32	18	42	18

Idea:
Simulate the distribution by *drawing samples with replacement.*

Motivation:
The data estimates its own distribution – we draw random samples from this distribution.

21	19								

Bootstrap Resampling:

48	24	51	12
21	41	25	23
32	61	19	24
29	21	23	13
32	18	42	18

Idea:
Simulate the distribution by *drawing samples with replacement.*

Motivation:
The data estimates its own distribution – we draw random samples from this distribution.

21	19	25							

Bootstrap Resampling:

48	24	51	12
21	41	25	23
32	61	19	24
29	21	23	13
32	18	42	18

Idea:
Simulate the distribution by *drawing samples with replacement.*

Motivation:
The data estimates its own distribution – we draw random samples from this distribution.

21	19	25	24						

Bootstrap Resampling:

48	24	51	12
21	41	25	23
32	61	19	24
29	21	23	13
32	18	42	18

Idea:
Simulate the distribution by *drawing samples with replacement.*

Motivation:
The data estimates its own distribution – we draw random samples from this distribution.

21	19	25	24	23					

Bootstrap Resampling:

48	24	51	12
21	41	25	23
32	61	19	24
29	21	23	13
32	18	42	18

Idea:
Simulate the distribution by *drawing samples with replacement.*

Motivation:
The data estimates its own distribution – we draw random samples from this distribution.

21	19	25	24	23	19				

Bootstrap Resampling:

48	24	51	12
21	41	25	23
32	61	19	24
29	21	23	13
32	18	42	18

Idea:
Simulate the distribution by *drawing samples with replacement.*

Motivation:
The data estimates its own distribution – we draw random samples from this distribution.

21	19	25	24	23	19	41			

Bootstrap Resampling:

48	24	51	12
21	41	25	23
32	61	19	24
29	21	23	13
32	18	42	18

Idea:
Simulate the distribution by *drawing samples with replacement.*

Motivation:
The data estimates its own distribution – we draw random samples from this distribution.

21	19	25	24	23	19	41	23		

Bootstrap Resampling:

48	24	51	12
21	41	25	23
32	61	19	24
29	21	23	13
32	18	42	18

Idea:
Simulate the distribution by *drawing samples with replacement.*

Motivation:
The data estimates its own distribution – we draw random samples from this distribution.

21	19	25	24	23	19	41	23	41	

Bootstrap Resampling:

48	24	51	12
21	41	25	23
32	61	19	24
29	21	23	13
32	18	42	18

Idea:
Simulate the distribution by *drawing samples with replacement.*

Motivation:
The data estimates its own distribution – we draw random samples from this distribution.

21	19	25	24	23	19	41	23	41	18

Bootstrap Resampling:

48	24	51	12
21	41	25	23
32	61	19	24
29	21	23	13
32	18	42	18

Idea:
Simulate the distribution by *drawing samples with replacement.*

Motivation:
The data estimates its own distribution – we draw random samples from this distribution.

21	19	25	24	23	19	41	23	41	18
61									

Bootstrap Resampling:

48	24	51	12
21	41	25	23
32	61	19	24
29	21	23	13
32	18	42	18

Idea:
Simulate the distribution by *drawing samples with replacement.*

Motivation:
The data estimates its own distribution – we draw random samples from this distribution.

21	19	25	24	23	19	41	23	41	18
61	12								

Bootstrap Resampling:

48	24	51	12
21	41	25	23
32	61	19	24
29	21	23	13
32	18	42	18

Idea:
Simulate the distribution by *drawing samples with replacement.*

Motivation:
The data estimates its own distribution – we draw random samples from this distribution.

21	19	25	24	23	19	41	23	41	18
61	12	42							

Bootstrap Resampling:

48	24	51	12
21	41	25	23
32	61	19	24
29	21	23	13
32	18	42	18

Idea:
Simulate the distribution by *drawing samples with replacement.*

Motivation:
The data estimates its own distribution – we draw random samples from this distribution.

21	19	25	24	23	19	41	23	41	18
61	12	42	42						

Bootstrap Resampling:

48	24	51	12
21	41	25	23
32	61	19	24
29	21	23	13
32	18	42	18

Idea:
Simulate the distribution by *drawing samples with replacement.*

Motivation:
The data estimates its own distribution – we draw random samples from this distribution.

21	19	25	24	23	19	41	23	41	18
61	12	42	42	42					

Bootstrap Resampling:

48	24	51	12
21	41	25	23
32	61	19	24
29	21	23	13
32	18	42	18

Idea:
Simulate the distribution by *drawing samples with replacement.*

Motivation:
The data estimates its own distribution – we draw random samples from this distribution.

21	19	25	24	23	19	41	23	41	18
61	12	42	42	42	19				

Bootstrap Resampling:

48	24	51	12
21	41	25	23
32	61	19	24
29	21	23	13
32	18	42	18

Idea:
Simulate the distribution by *drawing samples with replacement.*

Motivation:
The data estimates its own distribution – we draw random samples from this distribution.

21	19	25	24	23	19	41	23	41	18
61	12	42	42	42	19	18			

Bootstrap Resampling:

48	24	51	12
21	41	25	23
32	61	19	24
29	21	23	13
32	18	42	18

Idea:
Simulate the distribution by *drawing samples with replacement.*

Motivation:
The data estimates its own distribution – we draw random samples from this distribution.

21	19	25	24	23	19	41	23	41	18
61	12	42	42	42	19	18	61		

Bootstrap Resampling:

48	24	51	12
21	41	25	23
32	61	19	24
29	21	23	13
32	18	42	18

Idea:
Simulate the distribution by *drawing samples with replacement.*

Motivation:
The data estimates its own distribution – we draw random samples from this distribution.

21	19	25	24	23	19	41	23	41	18
61	12	42	42	42	19	18	61	29	

Bootstrap Resampling:

48	24	51	12
21	41	25	23
32	61	19	24
29	21	23	13
32	18	42	18

Idea:
Simulate the distribution by *drawing samples with replacement.*

Motivation:
The data estimates its own distribution – we draw random samples from this distribution.

21	19	25	24	23	19	41	23	41	18
61	12	42	42	42	19	18	61	29	41

Bootstrap Resampling:

48	24	51	12
21	41	25	23
32	61	19	24
29	21	23	13
32	18	42	18

Idea:
Simulate the distribution by *drawing samples with replacement.*

Motivation:
The data estimates its own distribution – we draw random samples from this distribution.

21	19	25	24	23	19	41	23	41	18
61	12	42	42	42	19	18	61	29	41

→ 31.05

Repeat this
several thousand times . . .