BOOTSTRAPPING AND CONFIDENCE INTERVALS

MPA 630: Data Science for Public Management

November 8, 2018

Fill out your reading report on Learning Suite

PLAN FOR TODAY

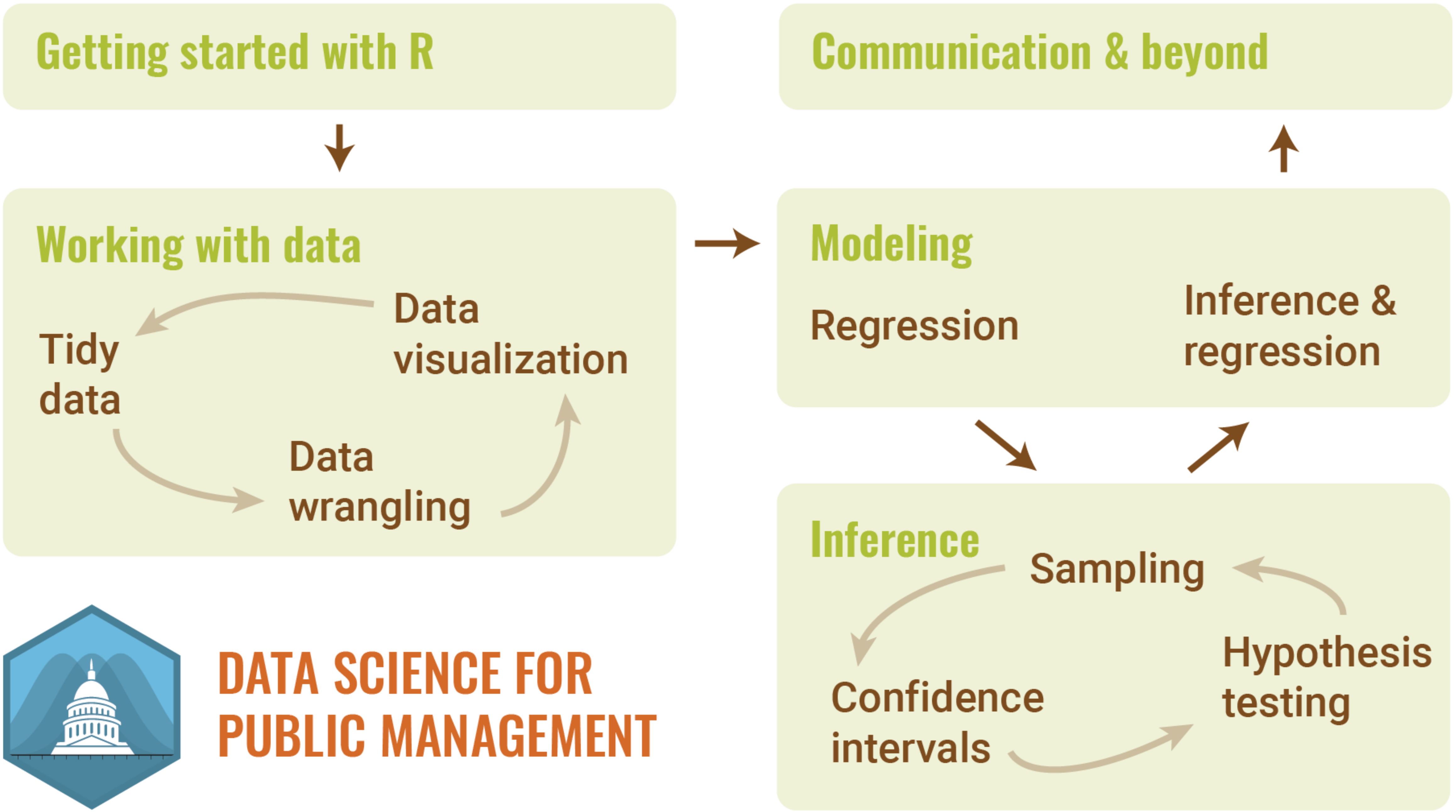
Why are we even doing this?

Confidence intervals

Bootstrapping

Precision vs. accuracy

WHY ARE WE EVEN DOING THIS?


Getting started with R
Communication & beyond
Working with data
Tidy data
Data visualization
Data wrangling
Modeling
Regression
Inference & regression
Inference
Sampling
Confidence intervals
Hypothesis testing
DATA SCIENCE FOR PUBLIC MANAGEMENT

POPULATION PARAMETERS

Population

A collection of things in the world

Population parameter

Something we want to know about the population

TYPES OF PARAMETERS

Proportion

Difference between proportions

Mean / median

Difference between means / medians

Intercept

Slope

% of Mia Love supporters in Utah County

Difference in student loan default rates for private vs. public universities

Median commute time for workers in Idaho

Relationship between property taxes and # of households with kids in 5 Western states

Difference in average test scores in large and small classes in the US

TYPES OF PARAMETERS

Proportion	p
Mean	μ
Difference between proportions	$p_1 - p_2$
Difference between means	$\mu_1 - \mu_2$
Intercept	β_0
Slope	β_1
Standard deviation	σ

POPULATION PARAMETERS

Key assumption in the flavor of statistics we're doing:

There are true, fixed population parameters out in the world

KNOWING THE POPULATION

In general, we *cannot* measure population parameters directly

How do we find out what they are?

Inference!

INFERENCE

Use sample data to make conclusions about the underlying population that the sample came from

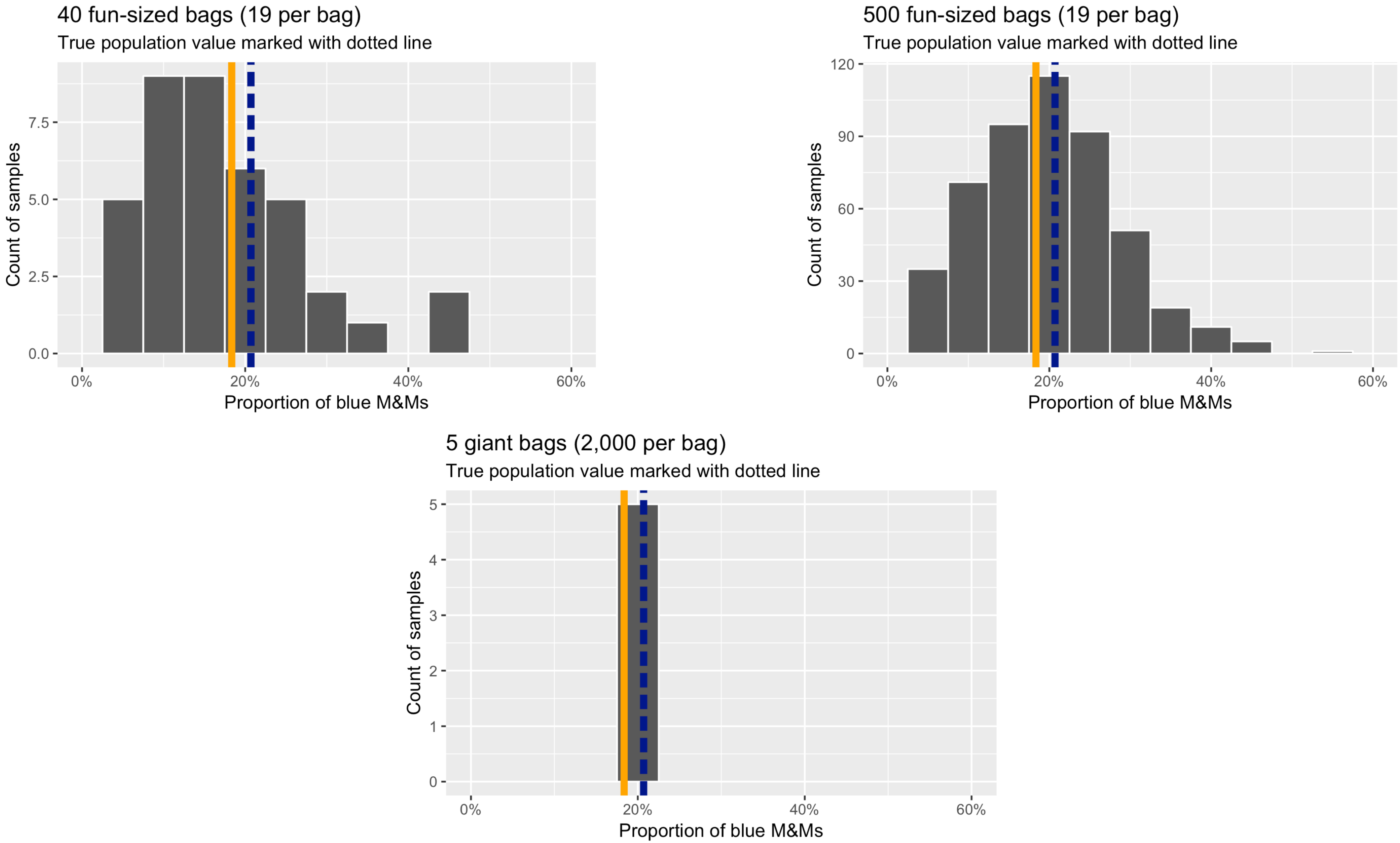
POPULATION VS. SAMPLE

Proportion	p	$\hat{p}$
Mean	μ	$\bar{x}$
Difference between proportions	$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$
Difference between means	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$
Intercept	β_0	$\hat{\beta}_0$
Slope	β_1	$\hat{\beta}_1$
Standard deviation	σ	s

SAMPLES AND SIZES

What happens to your sample statistic/point estimate as you increase the size of the sample?

What's better:
a small shovel you use a bunch of times
or a big shovel you use a few times (or even once)?



40 fun-sized bags (19 per bag)
True population value marked with dotted line
Count of samples
0.0
2.5
5.0
7.5
0%
20%
40%
60%
Proportion of blue M&Ms
500 fun-sized bags (19 per bag)
True population value marked with dotted line
Count of samples
0
30
60
90
120
0%
20%
40%
60%
Proportion of blue M&Ms
5 giant bags (2,000 per bag)
True population value marked with dotted line
Count of samples
0
1
2
3
4
5
0%
20%
40%
60%
Proportion of blue M&Ms

CONFIDENCE INTERVALS

GOAL OF INFERENCE

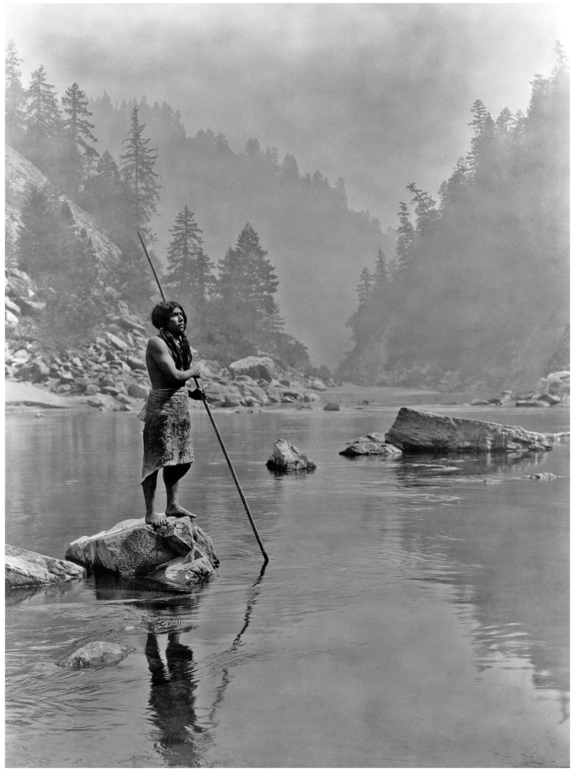
Make a good enough guess about the true population parameter

How do we know if the guess is good?

How confident are we that we captured the true population parameter?

CONFIDENCE INTERVALS

A plausible range of values for the true population parameter



VARIABILITY

Every sample statistic has some variability

You have an average, but how different might that average be if you take another sample?

LEFT-HANDEDNESS

You take a random sample of BYU students and 5 are left-handed.

If you take a different random sample of 50 BYU students, how many would you expect to be left-handed?

3 are left-handed. Is that surprising?

40 are left-handed. Is that surprising?

VARIABILITY

How much you expect the mean to vary from sample to sample

MEASURING VARIABILITY

2 ways to get at variability of sample statistic

Theory and math

Simulation

BOOTSTRAPPING

PERFECT KNOWLEDGE

With infinite resources, you could take thousands of simultaneous samples (or even conduct a census) and get exact population parameter and know its exact variability

This is impossible.

IMPROVISE!

Why bootstrap?

We can do something nearly impossible with limited resources

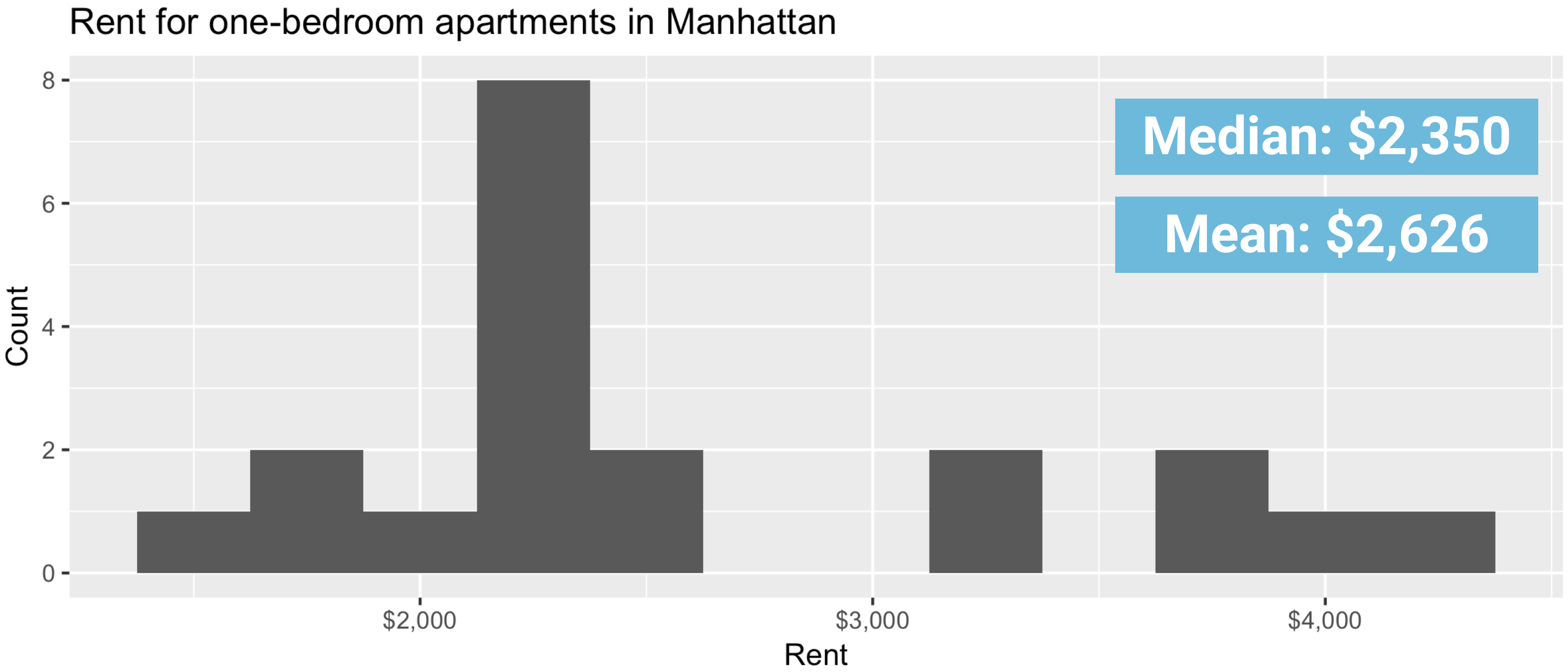
Use the sample you have to make new samples

How much does a typical 1-bedroom apartment in Manhattan rent for per month?

SAMPLE

Random sample of 20 apartments listed on Craigslist

SAMPLE DISTRIBUTION



MAIN QUESTION

Does this sample match the population? How well?

Median: $2,350

Median: ???

HOW TO BOOTSTRAP

Take a bootstrap sample

Sample with replacement; same size as original

Calculate a bootstrap statistic

Mean, median, proportion, difference, etc.

Repeat a lot

Calculate the bounds of an X% confidence interval as the middle X% of the bootstrap distribution

Sample

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20

Sample (arranged in order)

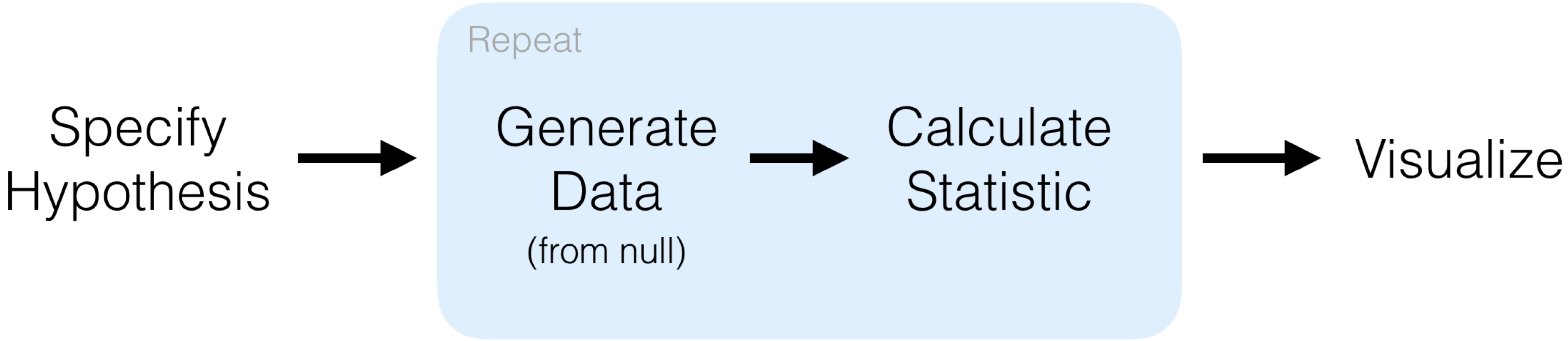
1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20

Bootstrap median

The number in between boxes 10 and 11 in the ordered section above

BOOTSTRAPPING WITH R

```
library(infer)
```



SEEDS

A seed ensures your random numbers are the same every time

```
set.seed(1234)
```

```
sample(1:100, 5)
```

BOOTSTRAPPING WITH R

```r
set.seed(1234)
manhattan <- read_csv("http://andhs.co/rents")

manhattan %>%
  # Specify the variable of interest
  specify(response = rent)
```

BOOTSTRAPPING WITH R

```
set.seed(1234)
manhattan <- read_csv("http://andhs.co/rents")

manhattan %>%
  # Specify the variable of interest
  specify(response = rent) %>%
  # Generate a bunch of bootstrap samples
  generate(reps = 1000, type = "bootstrap")
```

BOOTSTRAPPING WITH R

```
set.seed(1234)
manhattan <- read_csv("http://andhs.co/rents")

manhattan %>%
  # Specify the variable of interest
  specify(response = rent) %>%
  # Generate a bunch of bootstrap samples
  generate(reps = 1000, type = "bootstrap") %>%
  # Find the median of each sample
  calculate(stat = "median")
```

BOOTSTRAPPING WITH R

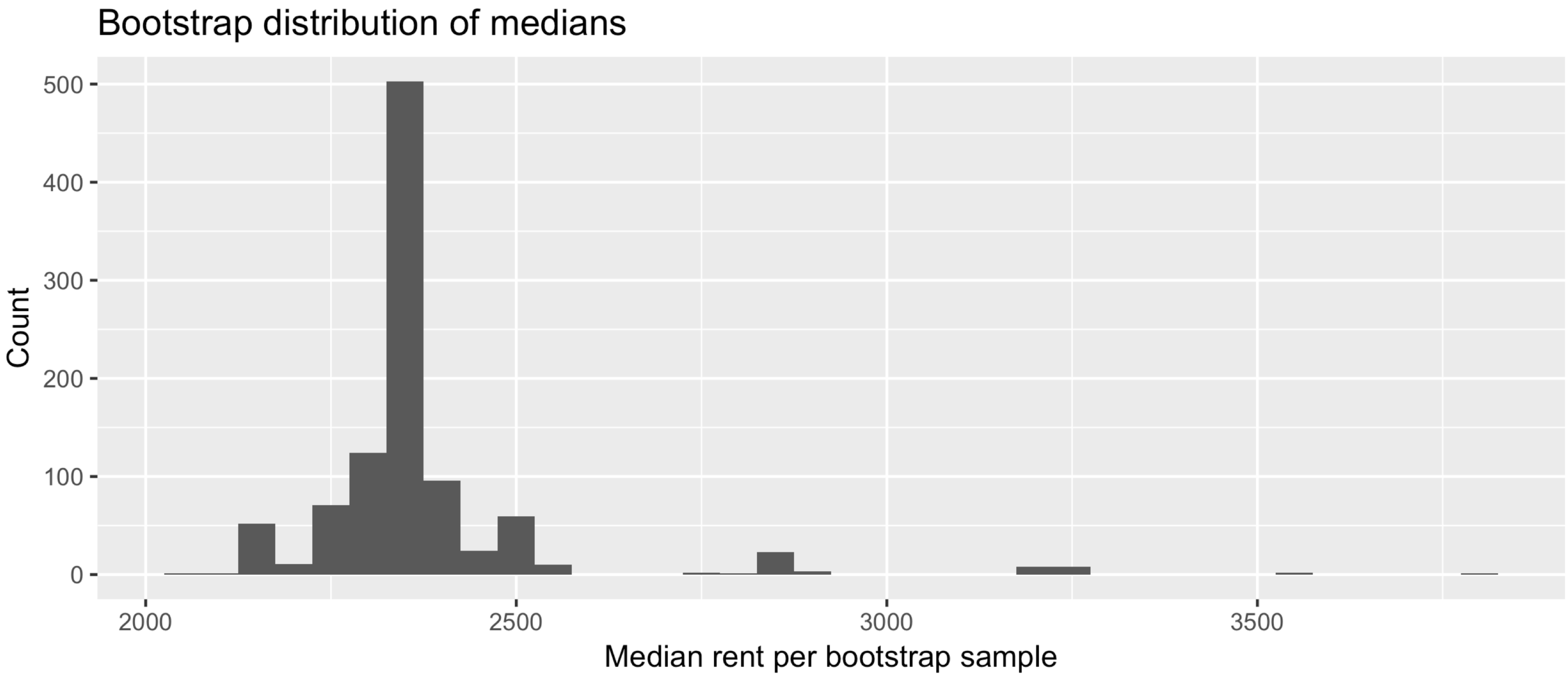
```r
set.seed(1234)
manhattan <- read_csv("http://andhs.co/rents")

# Save resulting bootstrap distribution
boot_rent <- manhattan %>%
  # Specify the variable of interest
  specify(response = rent) %>%
  # Generate a bunch of bootstrap samples
  generate(reps = 1000, type = "bootstrap") %>%
  # Find the median of each sample
  calculate(stat = "median")
```

SEE BOOTSTRAP MEDIANS

	replicate	stat
1	1	2150.0
2	2	2495.0
3	3	2237.5
4	4	2495.0
5	5	2350.0
6	6	2350.0
7	7	2160.0
8	8	2324.5
9	9	2495.0
10	10	2350.0
11	11	2300.0
12	12	2350.0
13	13	2325.0
14	14	2350.0
15	15	2349.5
16	16	2350.0

VISUALIZE BOOTSTRAP DISTRIBUTION

```
ggplot(boot_rent, aes(x = stat)) +
  geom_histogram(binwidth = 50)
```



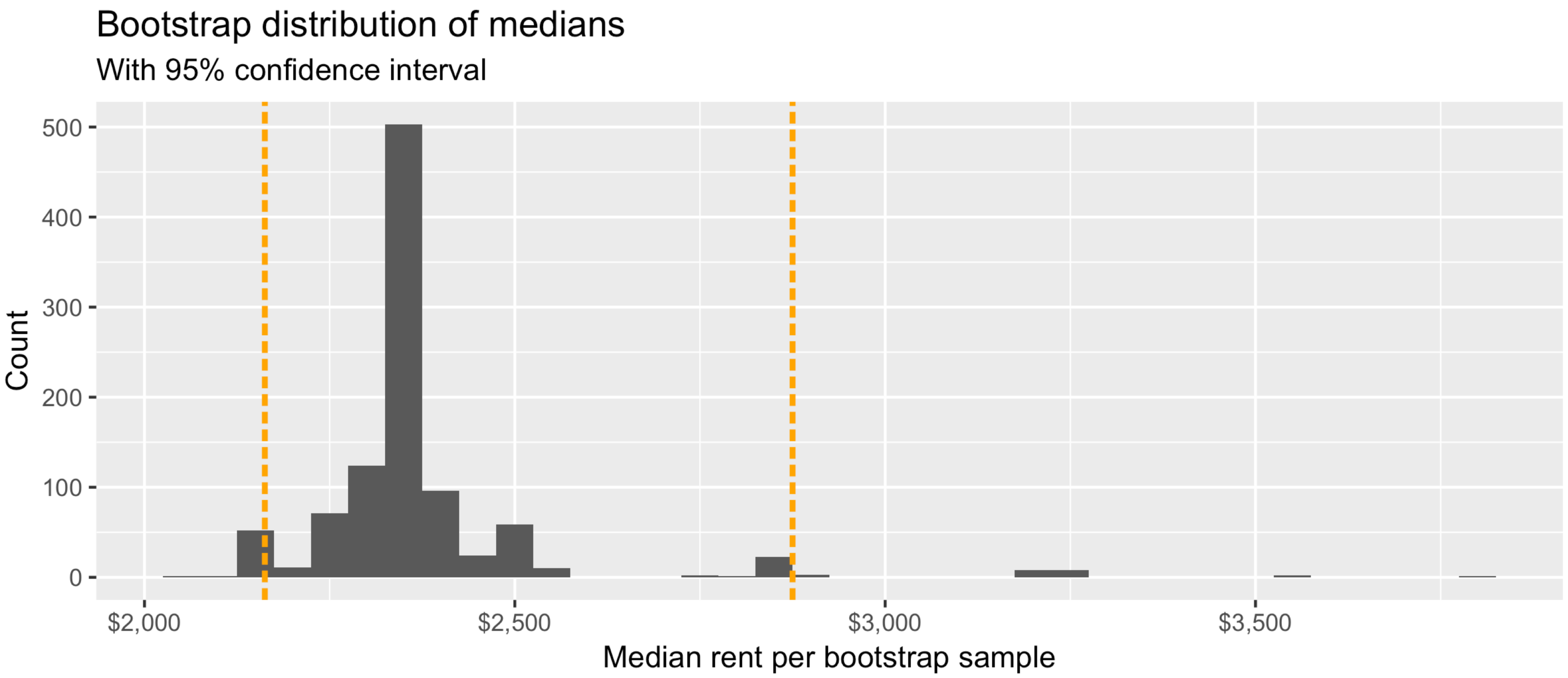
CALCULATE CONFIDENCE INTERVAL

95% confidence interval is the middle 95% of the the distribution

From 2.5% to 97.5%

```
boot_rent %>%
  get_ci(level = 0.95, type = "percentile")
```

```
# A tibble: 1 x 2
  `2.5%` `97.5%`
   <dbl>   <dbl>
1  2162.    2875
```


Bootstrap distribution of medians
With 95% confidence interval
Count
500
400
300
200
100
0
$2,000
$2,500
$3,000
$3,500
Median rent per bootstrap sample

INTERPRET CONFIDENCE INTERVAL

The 95% confidence interval for the median rent of one bedroom apartments in Manhattan was calculated as (2162.5, 2875). Which of the following is the correct interpretation of this interval?

95% of the time the median rent one bedroom apartments in this sample is between $2,162.5 and $2,875.

95% of all one bedroom apartments in Manhattan have rents between $2,162.5 and $2,875.

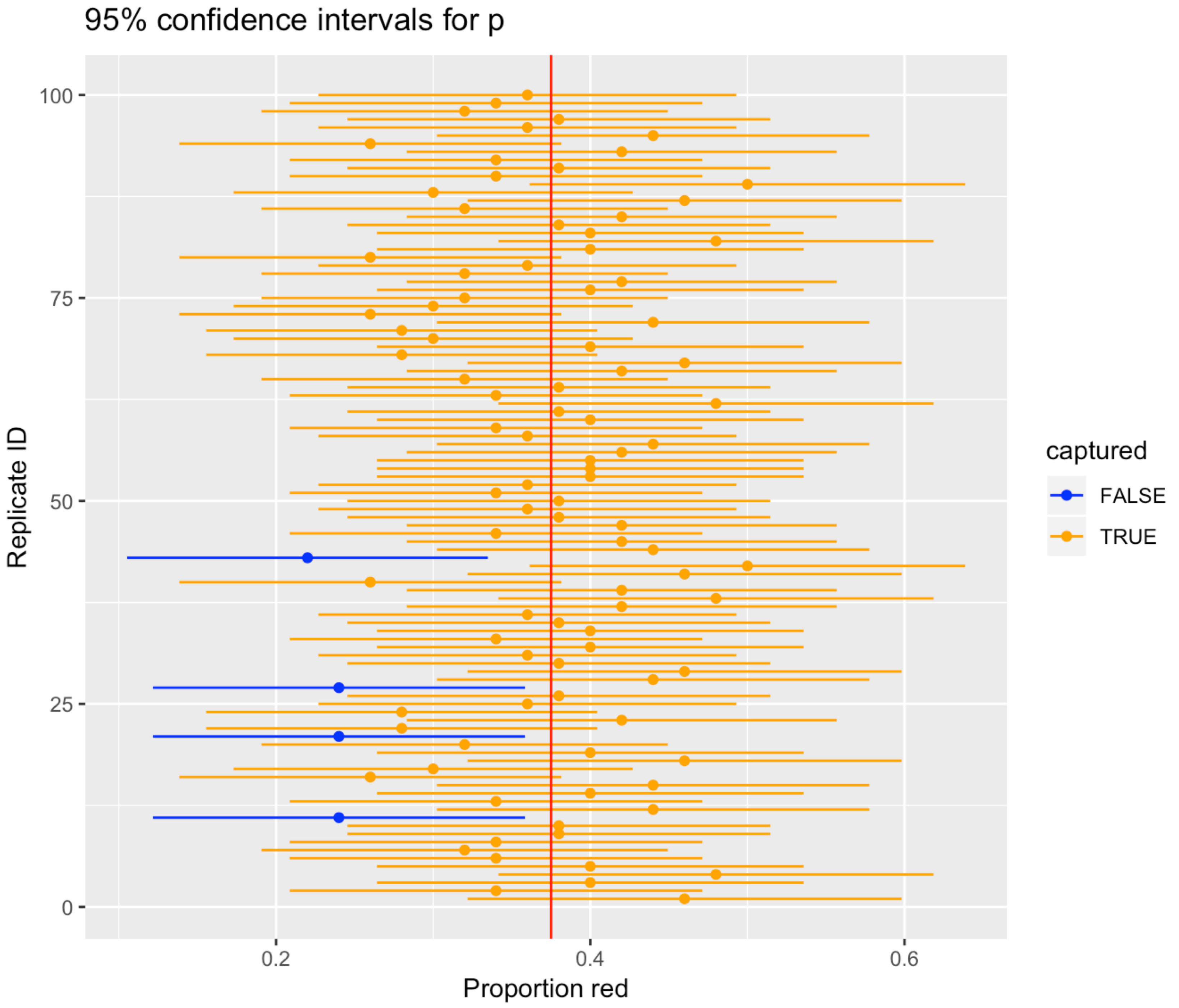
★ We are 95% confident that the median rent of all one bedroom apartments is between $2162.5 and $2875.

We are 95% confident that the median rent of one bedroom apartments in this sample is between $2162.5 and $2875.

MORE ON CONFIDENCE INTERVALS

Confidence intervals are a net

If we took 100 samples, at least 95 of of them would have the true population parameter in their 95% confidence intervals


95% confidence intervals for p
Replicate ID
100
75
50
25
0
0.2
0.4
0.6
Proportion red
captured
FALSE
TRUE

DON'T BE TEMPTED!

It is way too tempting to say "We're 95% sure that the population parameter is X"

People do this all the time! People with PhDs!

YOU will try to do this too

ONLY LEGAL INTERPRETATION

“There is a 95% chance that when I compute a confidence interval from this data, the true population value will be in it.”

CNN conducts a poll among a random sample of 800 voters about whether they approve of the president's performance. CNN analysts create a 90% confidence interval for the true proportion of all voters in the US who approve of the president's performance.

If CNN conducts many identical polls on the same night, about 90% of the intervals produced will capture the true proportion of voters who approve of the president

About 90% of people who support the president will respond to the poll

If CNN repeats this poll 20 times on the same night and calculates 90% confidence intervals for each poll, we can expect that around 18 of those intervals will contain the true proportion of voters who approve of the president.

There's a 90% chance that the actual population proportion is in the confidence interval

A city manager wants to know the true average property value of single-value homes in her city. She takes a random sample of 200 houses and builds a 95% confidence interval through bootstrapping. The interval is ($180,000, $300,000).

If the city manager took another random sample of 200 houses, there's a 95% chance *that* sample mean would be between $180,000 and $300,000

About 95% of houses in the sample are valued between $180,000 and $300,000

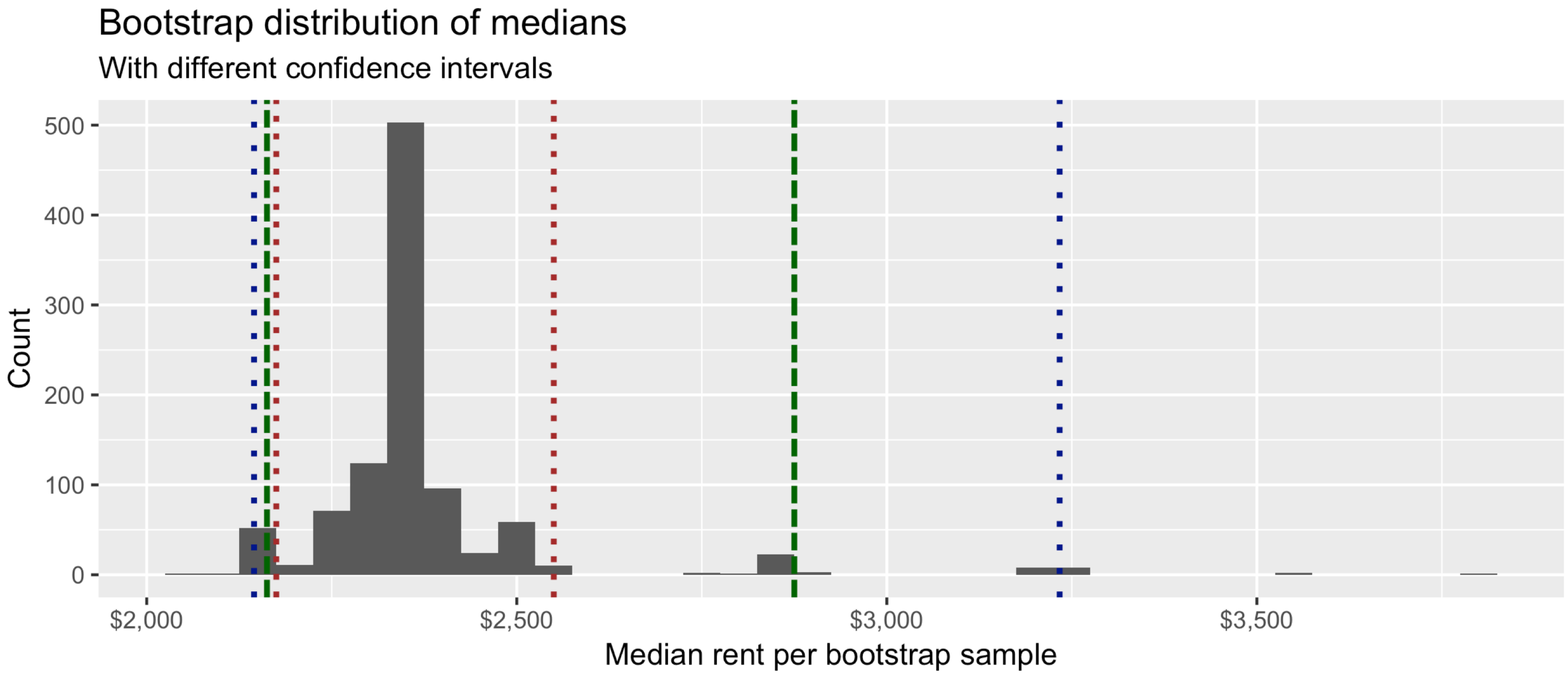
★ We're 95% confident that the interval ($180,000, $300,000) captured the true mean value

☆ There's a 95% chance that the true mean is between $180,000 and $300,000

PRECISION VS. ACCURACY

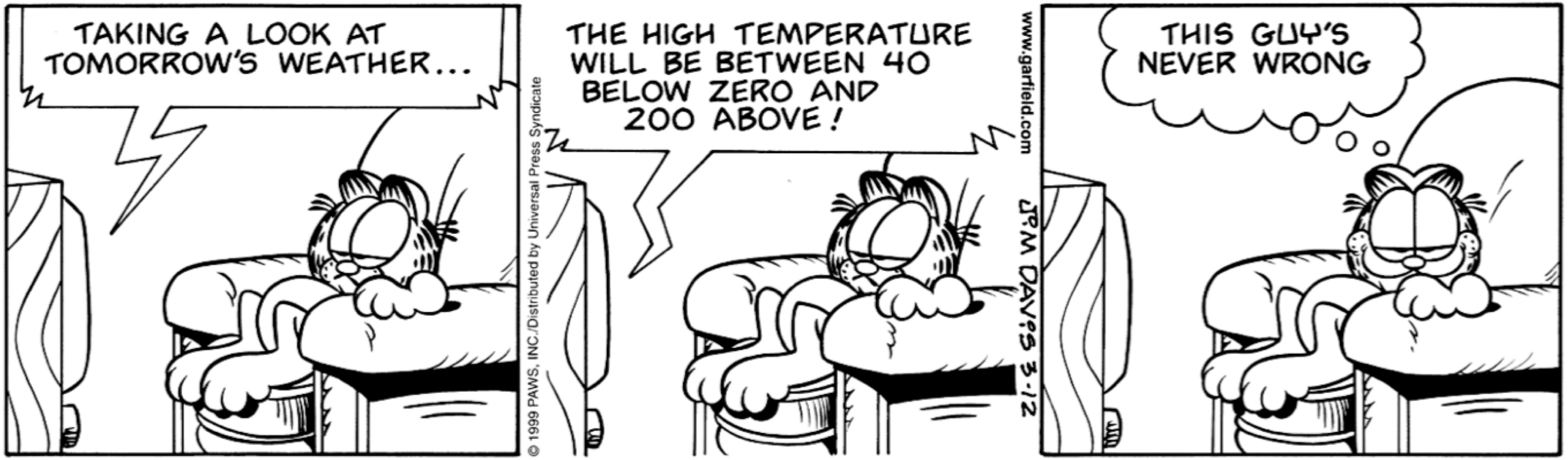
COMMON LEVELS

90%, 95%, 99%



PRECISION VS. ACCURACY

If we want to be very certain that we capture the population parameter, should we use a wider interval or a narrower interval?



The city manager was to be more confident about home prices in her city, so she uses the same sample data and uses bootstrapping techniques to calculate a 99% confidence interval.

What will happen to the the interval when she changes the confidence level from 95% to 99%?

It's impossible to say without seeing the sample data

★ Increasing the confidence to 99% will increase the margin of error and result in a wider interval

Increasing the confidence to 99% will decrease the margin of error and result in a narrower interval

MORAL OF THE STORY

Sample statistic ≠ population parameter

But if the sample is good, it can be a good estimate

Report estimate with confidence interval

Width of interval depends on how variable sample statistics would be from different samples

We can't keep sampling from the population, so bootstrap

This lets us measure the variability