INFERENCE FOR NUMERICAL DATA

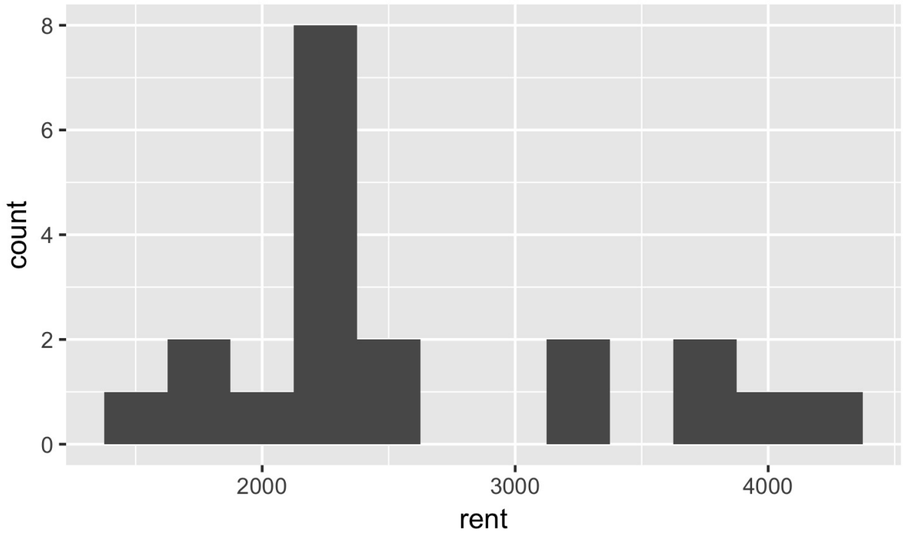
Welcome to the course!

Mine Cetinkaya-Rundel
Associate Professor of the Practice, Duke University

Rent in Manhattan

On a given day, twenty 1 BR apartments were randomly selected on Craigslist Manhattan from apartments listed as "by owner" (as opposed to by a rental agency).

Is the mean or the median a better measure of typical rent in Manhattan?



Bootstrapping techniques

- Assume the data is representative
- Pulling oneself up by one's bootstraps

Observed sample

sample median = $2,350

Bootstrap population

Bootstraping scheme

1. Take a bootstrap sample - a random sample taken with replacement from the original sample, of the same size as the original sample.

2. Calculate the bootstrap statistic - a statistic such as mean, median, proportion, etc. computed on the bootstrap samples.

3. Repeat steps (1) and (2) many times to create a bootstrap distribution - a distribution of bootstrap statistics.

Bootstraping scheme, in R

```
library(infer)

___ %>%                                  # start with data frame
  specify(response = ___) %>%            # specify the variable of interest
```

Bootstraping scheme, in R

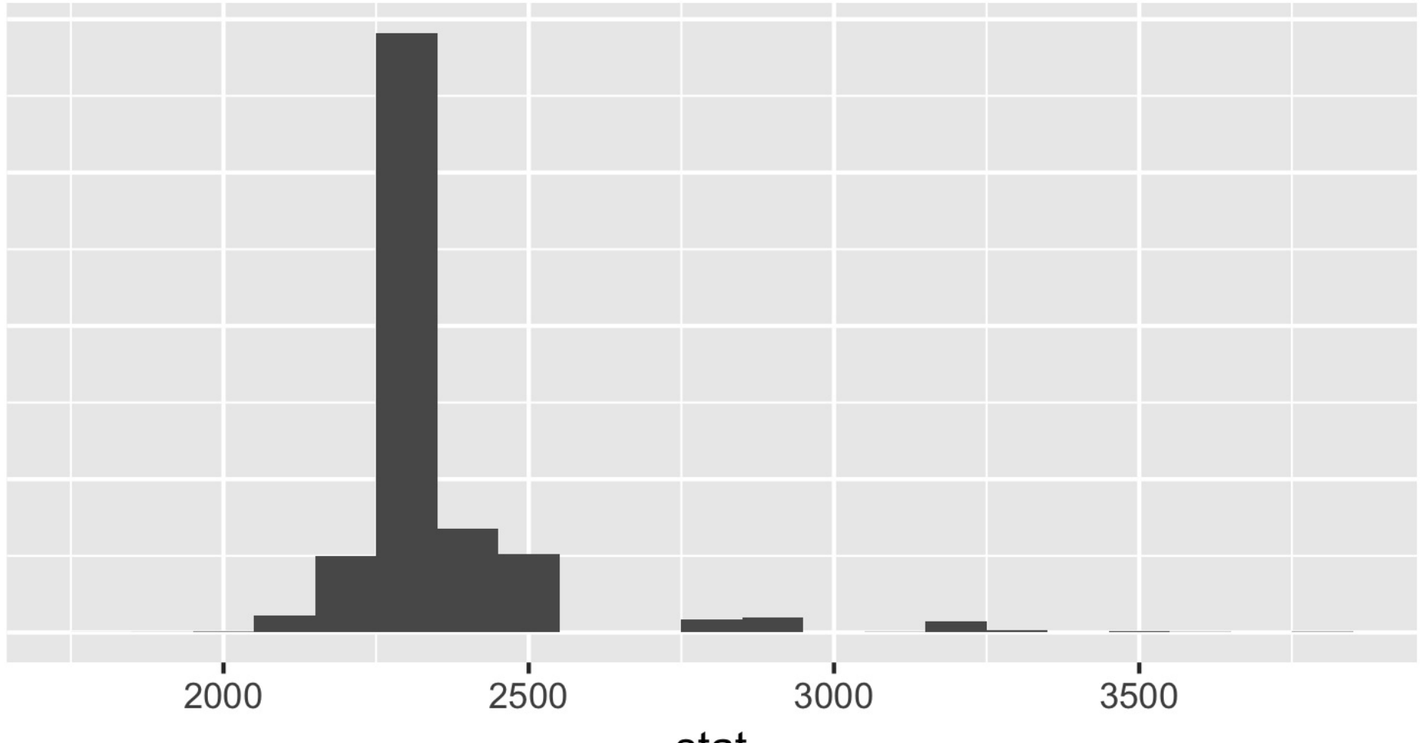
```
library(infer)

___ %>%                                  # start with data frame
  specify(response = ___) %>%            # specify the variable of interest
  generate(reps = ___, type = "bootstrap") %>%  # generate bootstrap samples
```

Bootstraping scheme, in R

```
library(infer)

___ %>%                                  # start with data frame
  specify(response = ___) %>%            # specify the variable of interest
  generate(reps = ___, type = "bootstrap") %>%  # generate bootstrap samples
  calculate(stat = "___")                # calculate bootstrap statistic
```

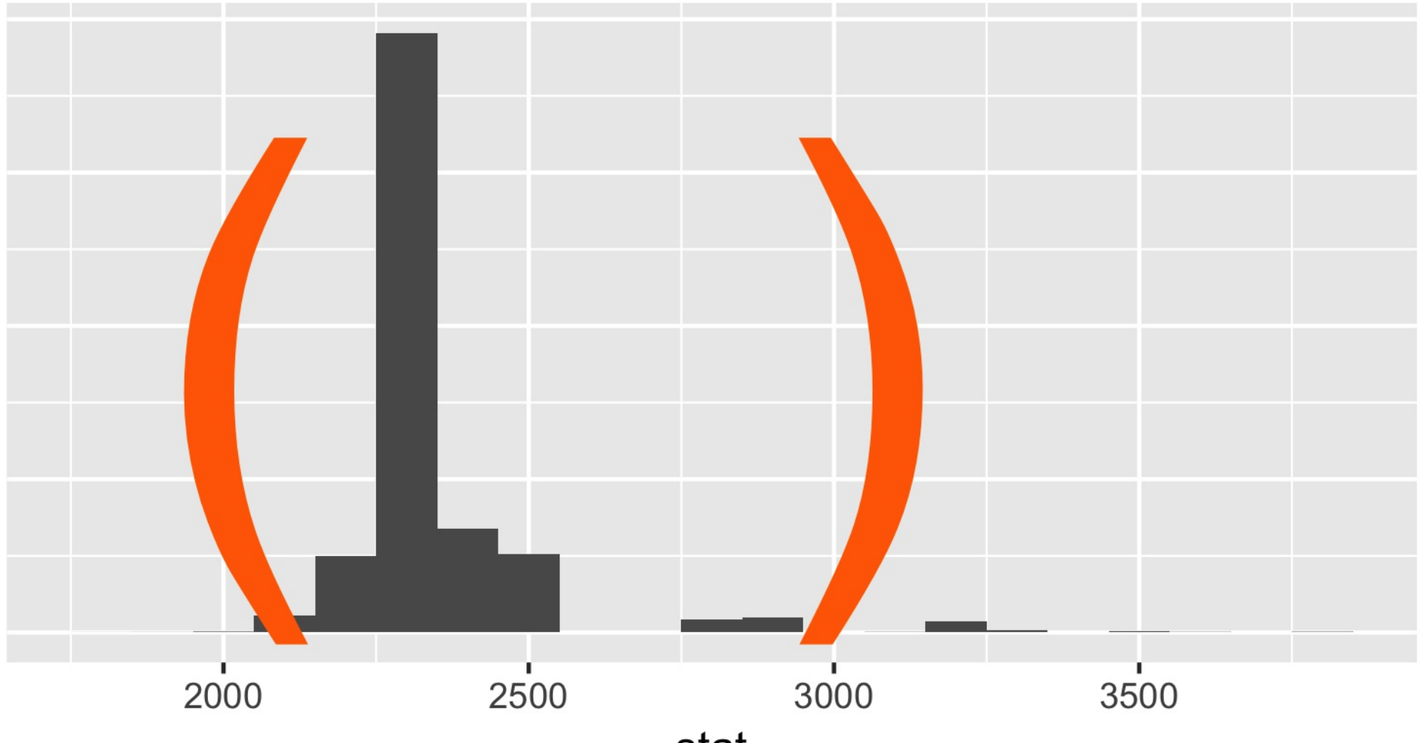
Constructing the bootstrap interval

```
library(infer)

___ %>%                                  # start with data frame
  specify(response = ___) %>%            # specify the variable of interest
  generate(reps = ___, type = "bootstrap") %>%  # generate bootstrap samples
  calculate(stat = "___")                # calculate bootstrap statistic
```



Constructing the bootstrap interval

```
library(infer)

___ %>%                                    # start with data frame
  specify(response = ___) %>%              # specify the variable of interest
  generate(reps = ___, type = "bootstrap") %>%  # generate bootstrap samples
  calculate(stat = "___")                  # calculate bootstrap statistic
```



INFERENCE FOR NUMERICAL DATA

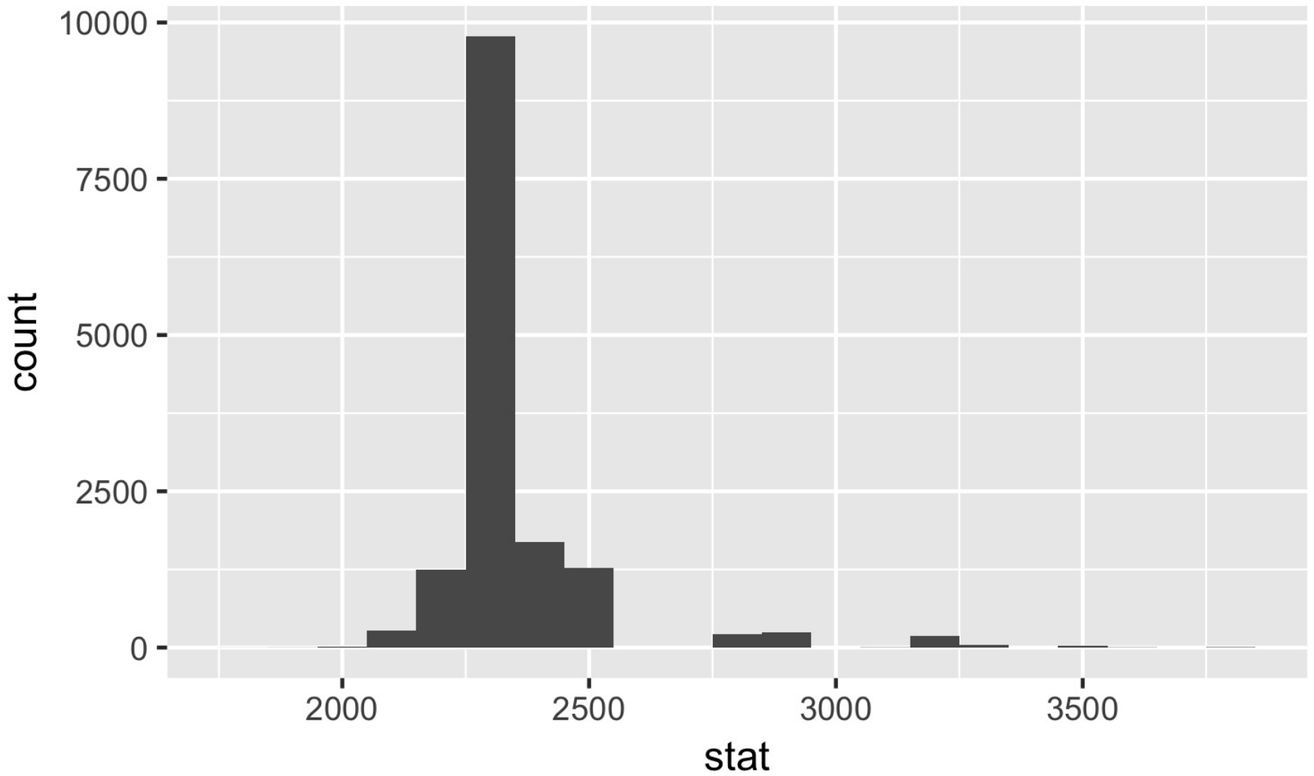
Let's practice!

INFERENCE FOR NUMERICAL DATA

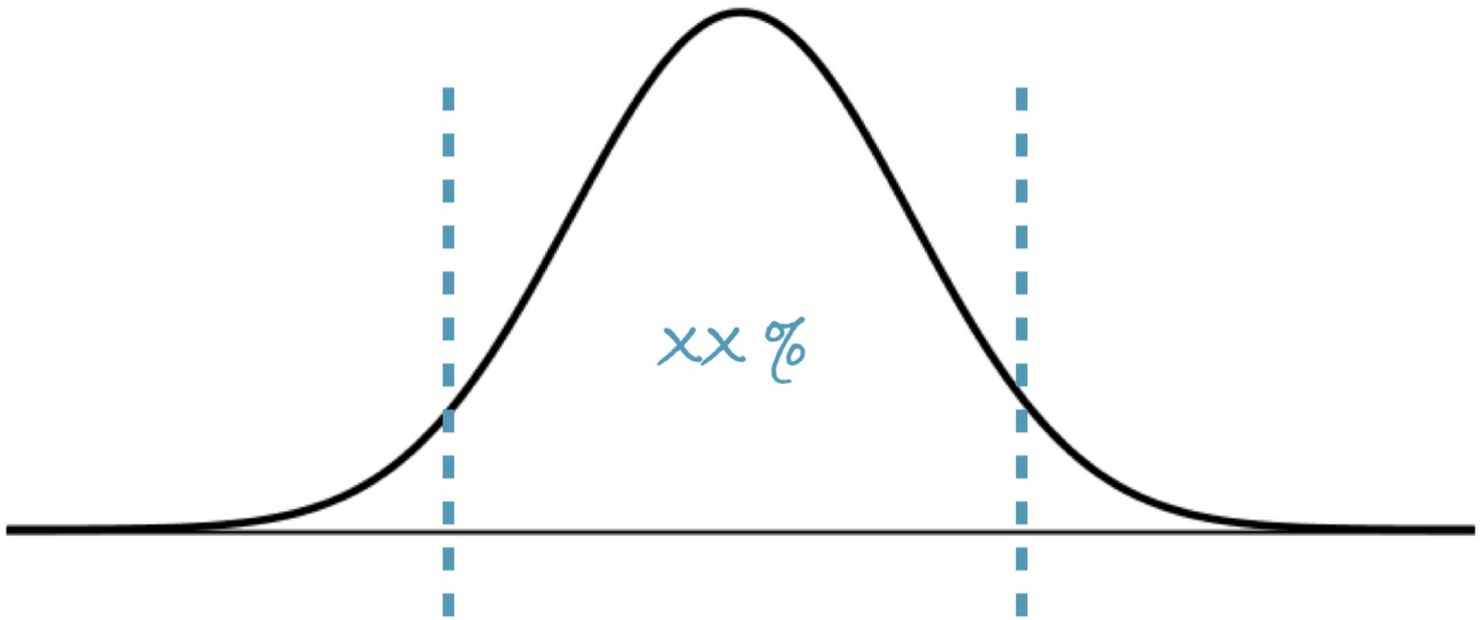
Review: Percentile and standard error methods

Mine Cetinkaya-Rundel
Associate Professor of the Practice, Duke University

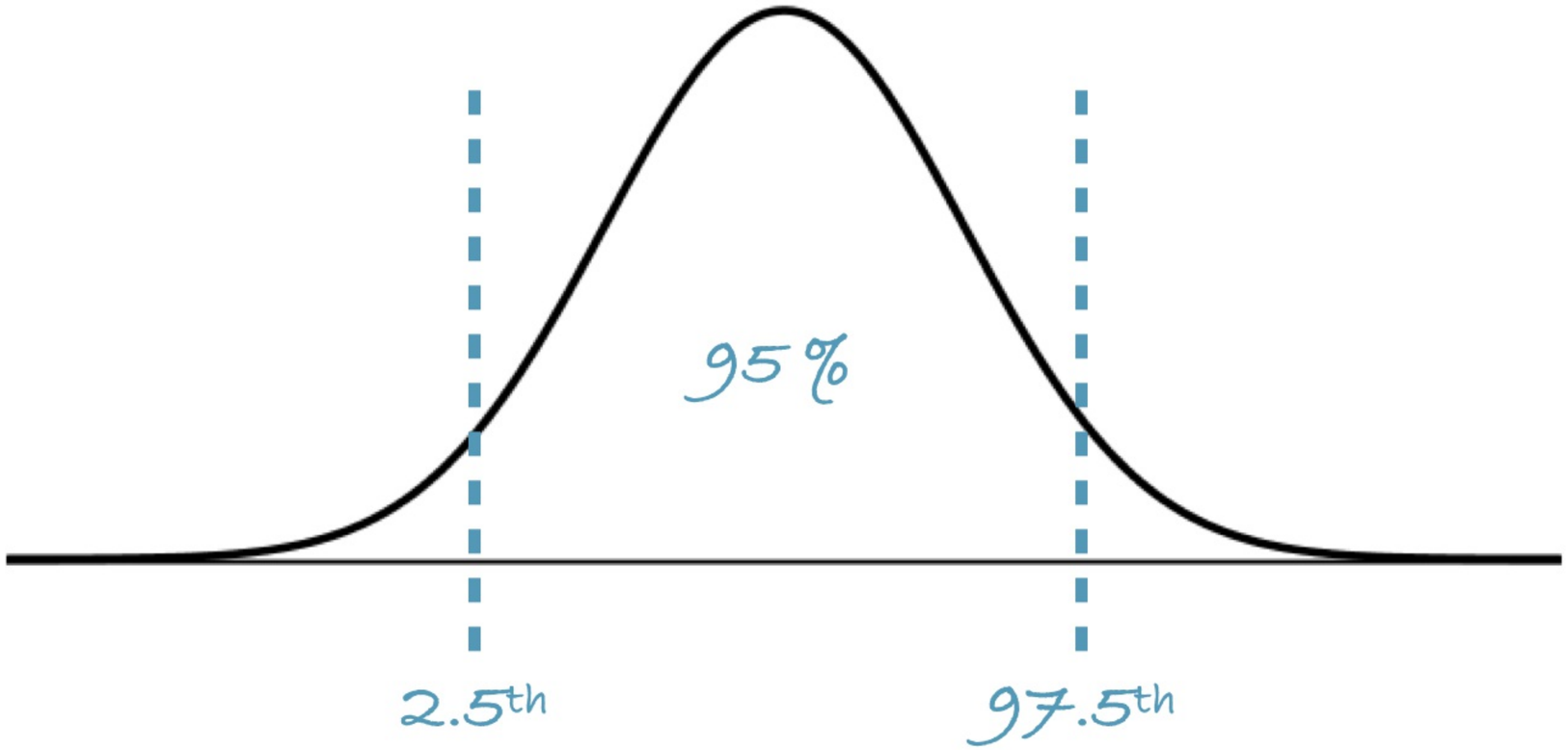
Bootstrap distribution



Percentile method



Percentile method



Standard error method

sample statistic $\pm\ t^*_{df=n-1} \times SE_{boot}$

- df for t^* is $n-1$, where n is the sample size
- SE_{boot} is the standard deviation of the bootstrap distribution

distribution

INFERENCE FOR NUMERICAL DATA

Let's practice!

INFERENCE FOR NUMERICAL DATA

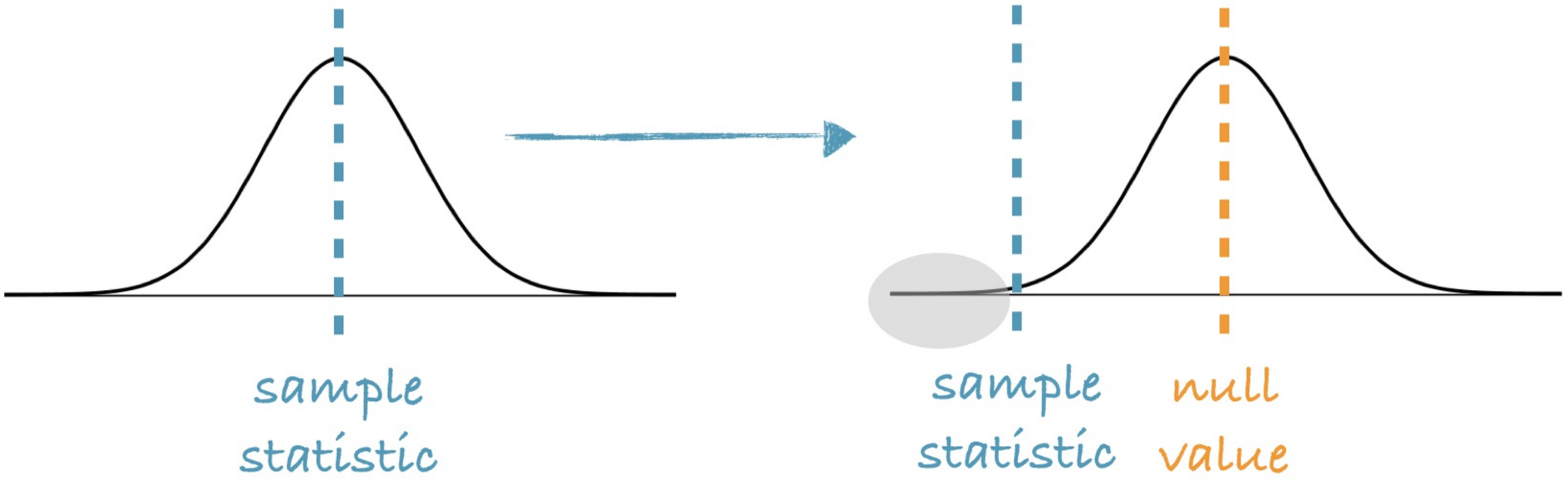
Re-centering a bootstrap distribution for hypothesis testing

Mine Cetinkaya-Rundel
Associate Professor of the Practice, Duke University

Re-centering a bootstrap distribution for hypothesis testing

- Bootstrap distributions are by design centered at the observed sample statistic.
- However since in a hypothesis test we assume that H_0 is true, we shift the bootstrap distribution to be centered at the null value.
- p-value = The proportion of simulations that yield a sample statistic at least as favorable to the alternative hypothesis as the observed sample statistic.

Re-centering the bootstrap distribution - sketch



INFERENCE FOR NUMERICAL DATA

Let's practice!