Welcome to the course!

FOUNDATIONS OF INFERENCE

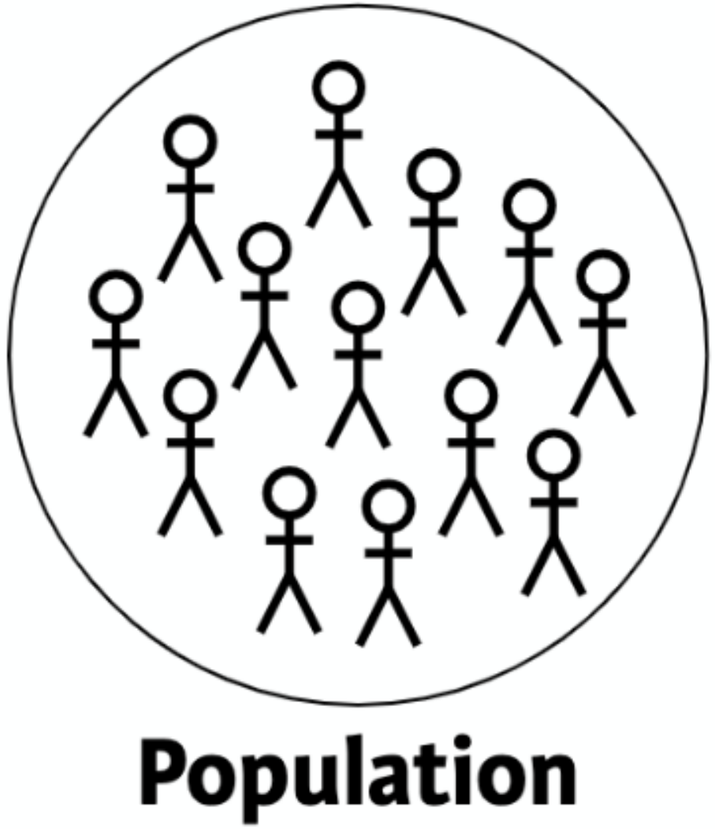
Jo Hardin
Instructor

datacamp

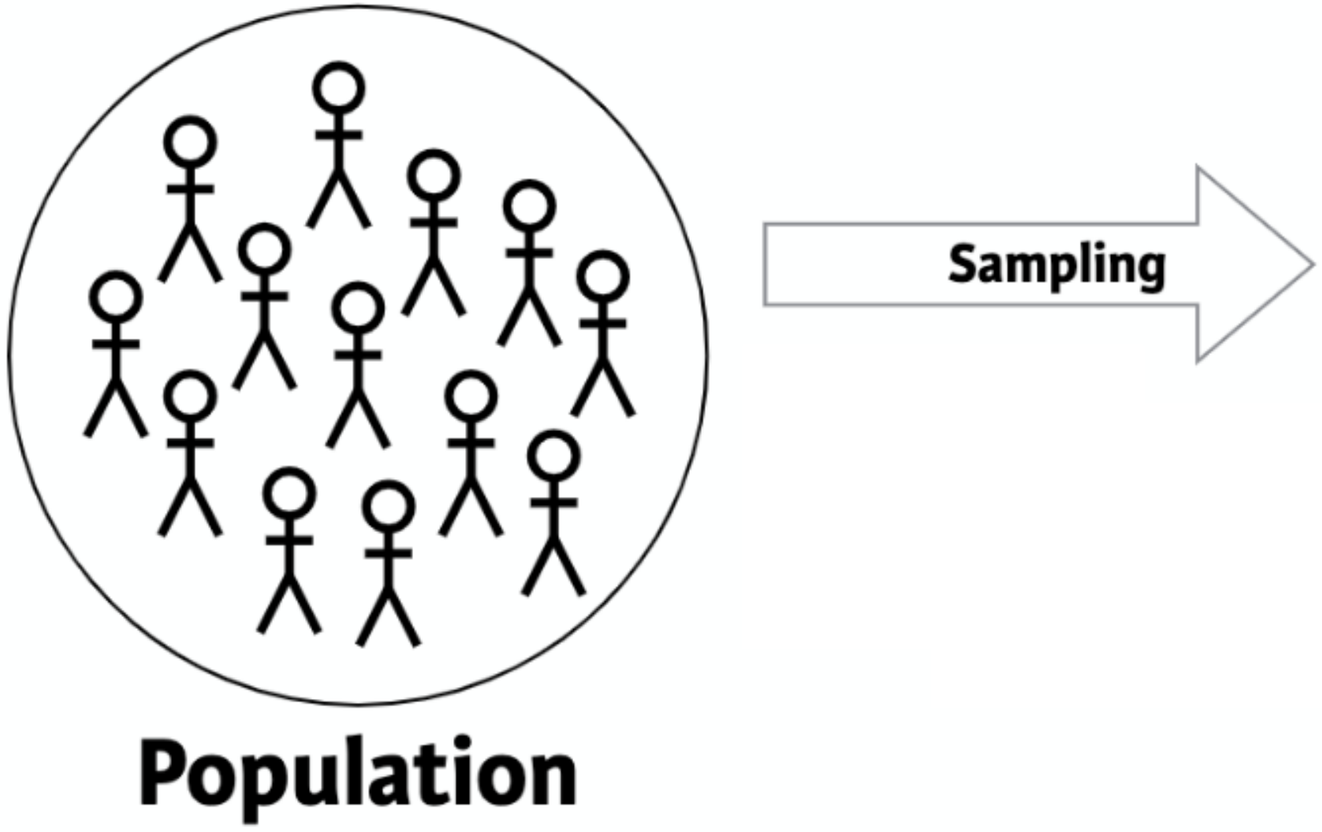
What is statistical inference?

The process of making claims about a population based on information from a sample

What is statistical inference?

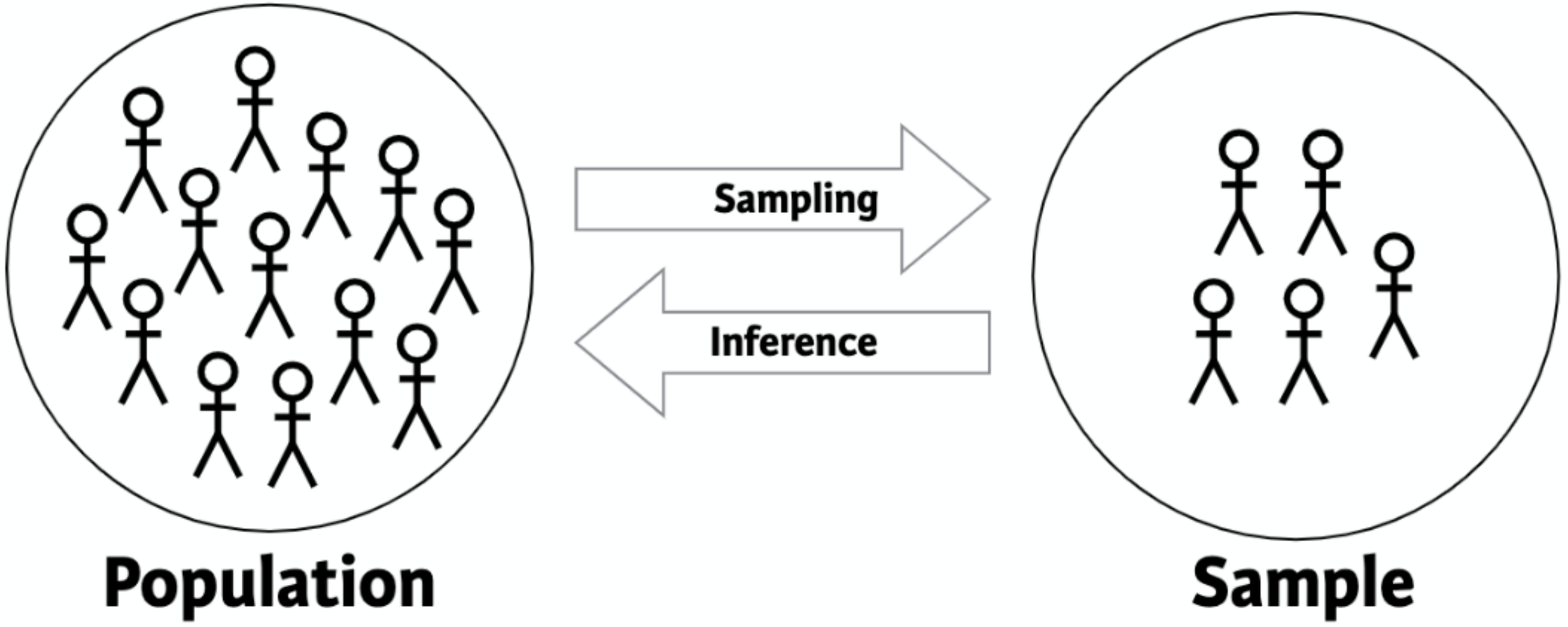


What is statistical inference?

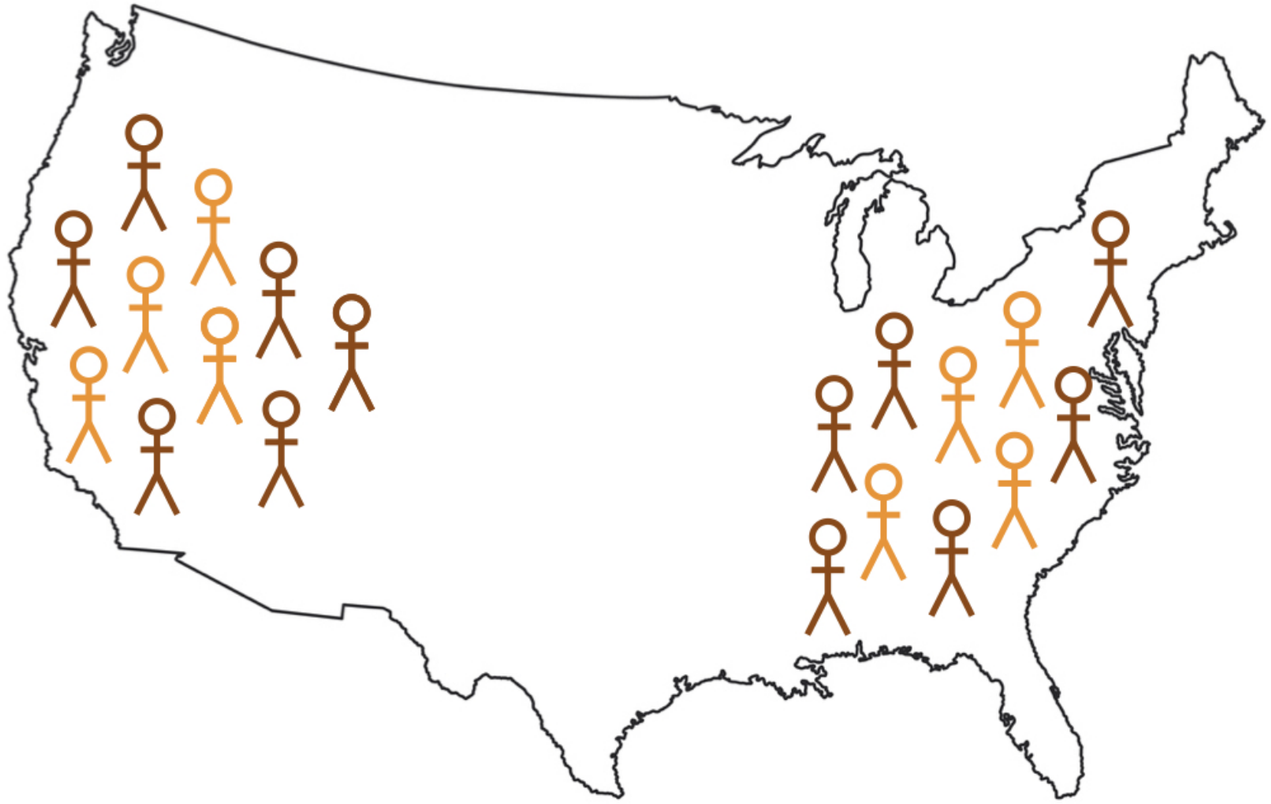


What is statistical inference?

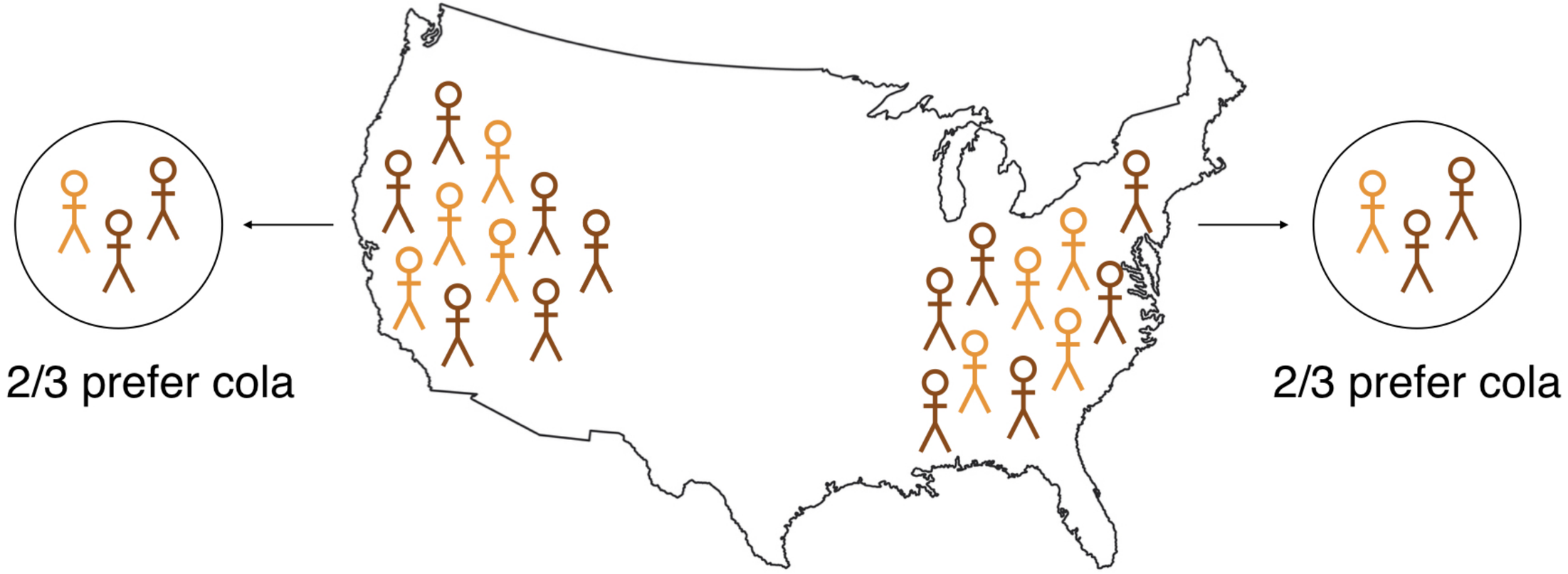
What is statistical inference?



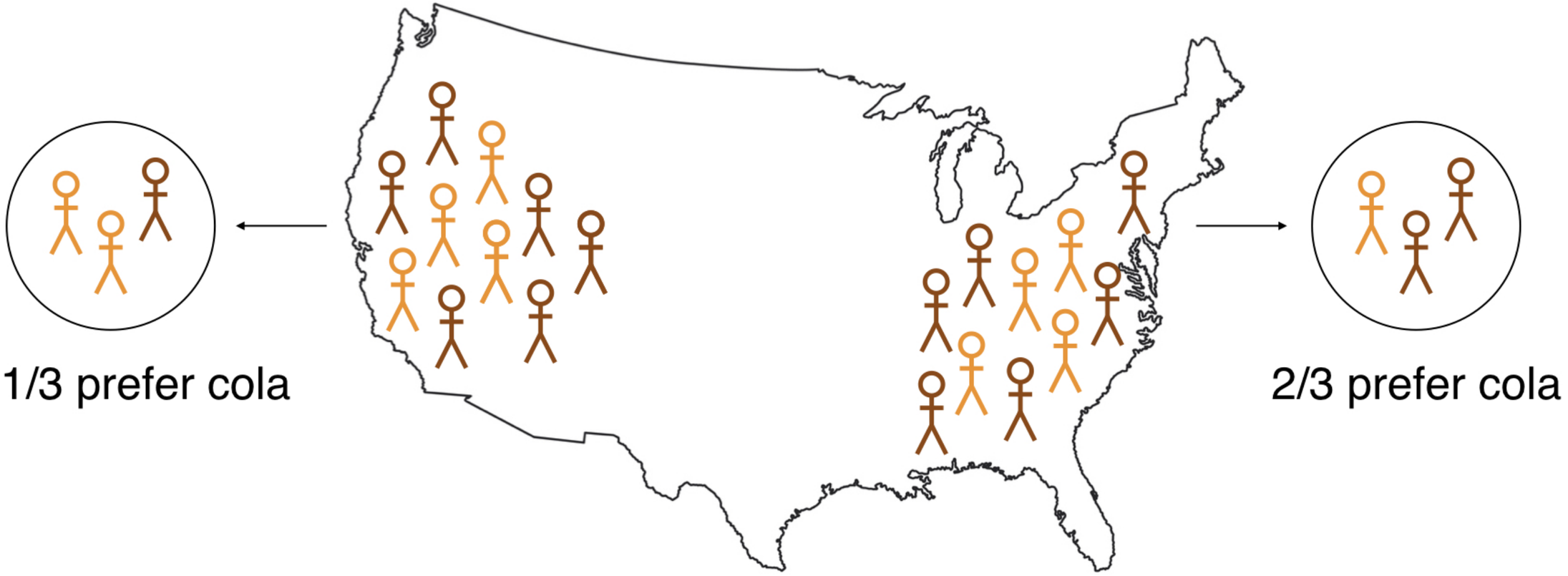
Assume two populations prefer cola at same rate



The sample data



The sample data (take 2)



Vocabulary

Null hypothesis (H_0): The claim is not that interesting

Alternative hypothesis (H_A): The claim corresponding to the research hypothesis

The "goal" is to disprove the null hypothesis

Example: cheetah speed

Compare speed of two different subspecies of cheetah

H_0: Asian and African cheetahs run the same speed, on average

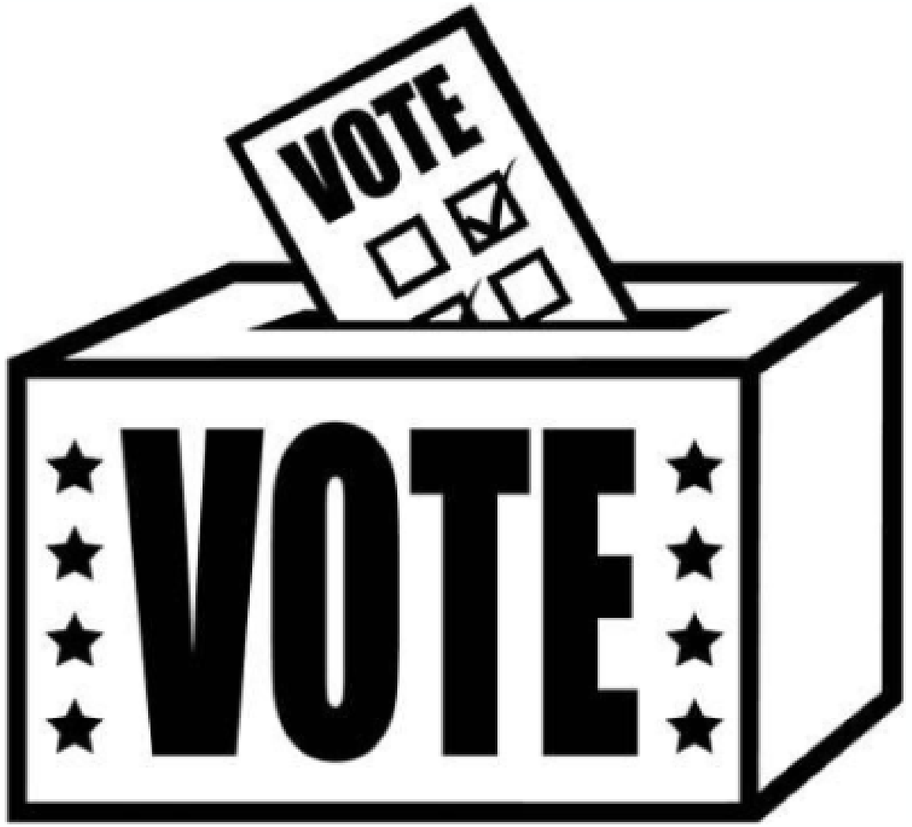
H_A: African cheetahs are faster than Asian cheetahs, on average

Example: election

From a sample, the researchers would like to claim that Candidate X will win

H_0: Candidate X will get half the votes

H_A: Candidate X will get more than half the votes



Let's practice!

FOUNDATIONS OF INFERENCE

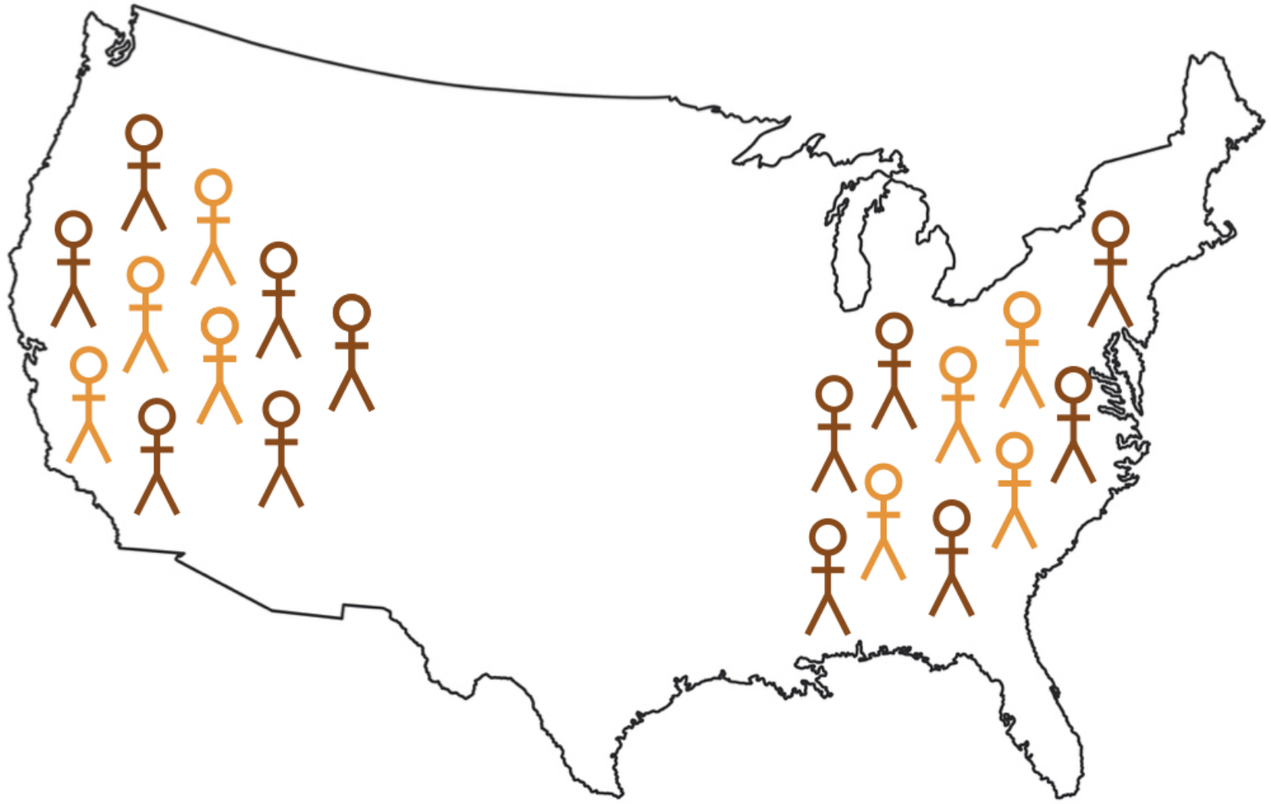
Randomized distributions

FOUNDATIONS OF INFERENCE

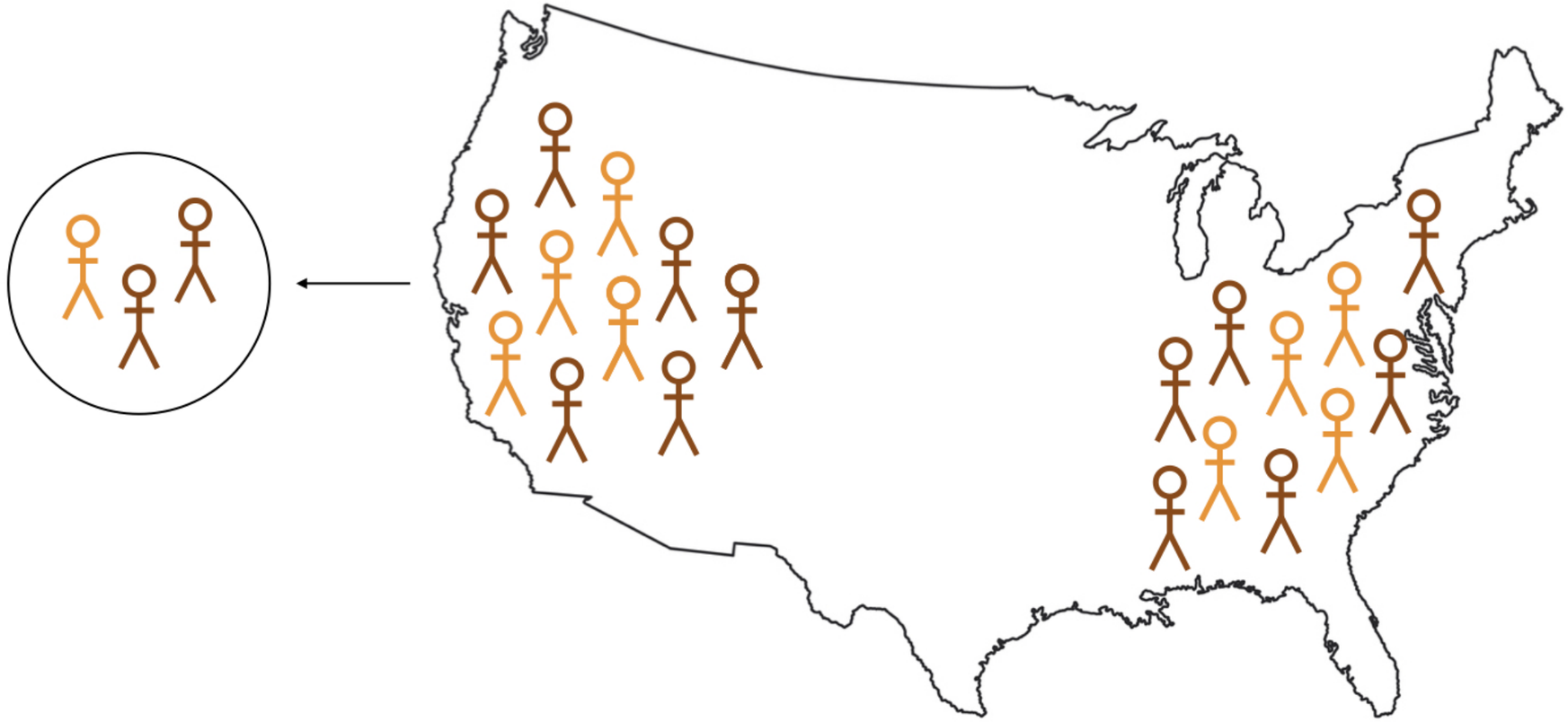
Jo Hardin
Instructor

datacamp

Logic of inference

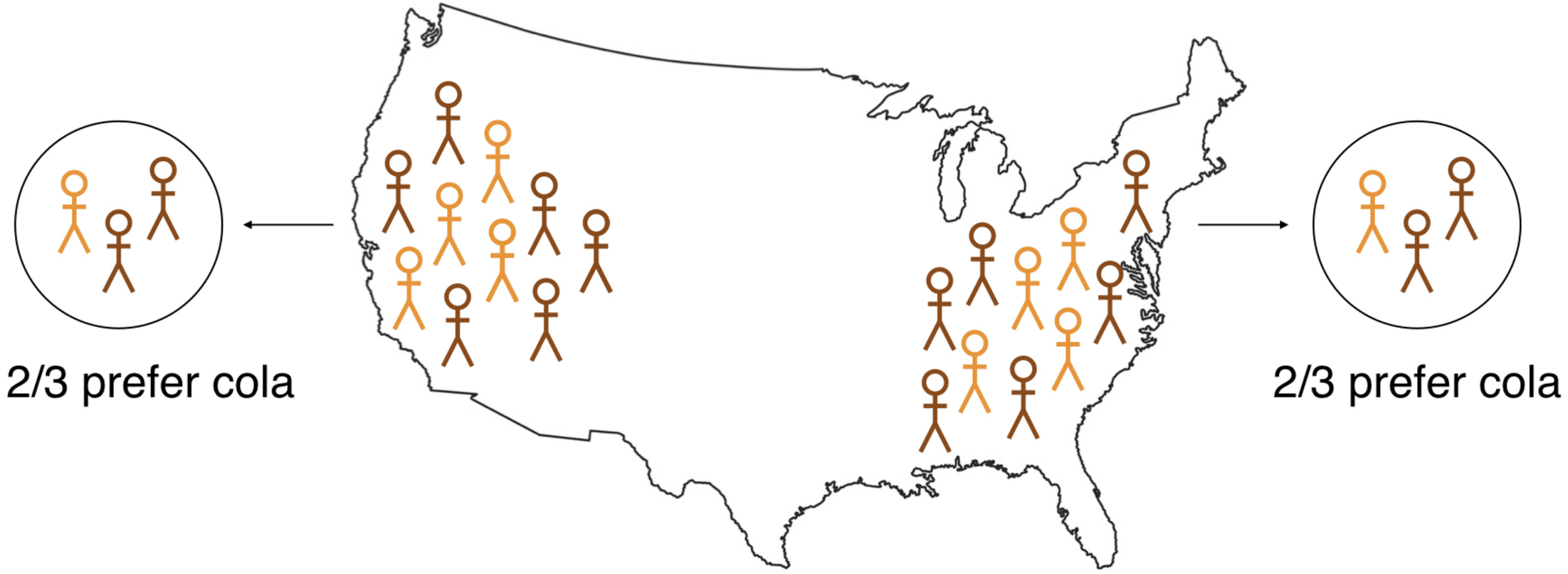


Logic of inference

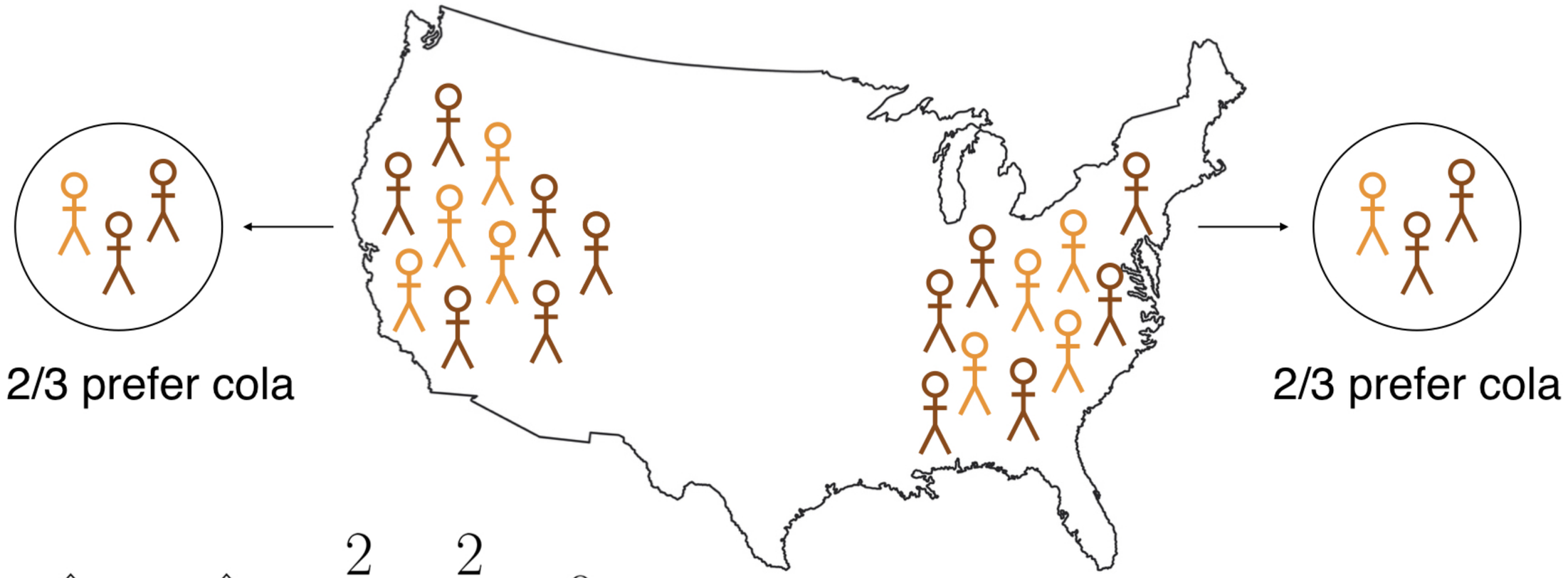


Logic of inference

Logic of inference

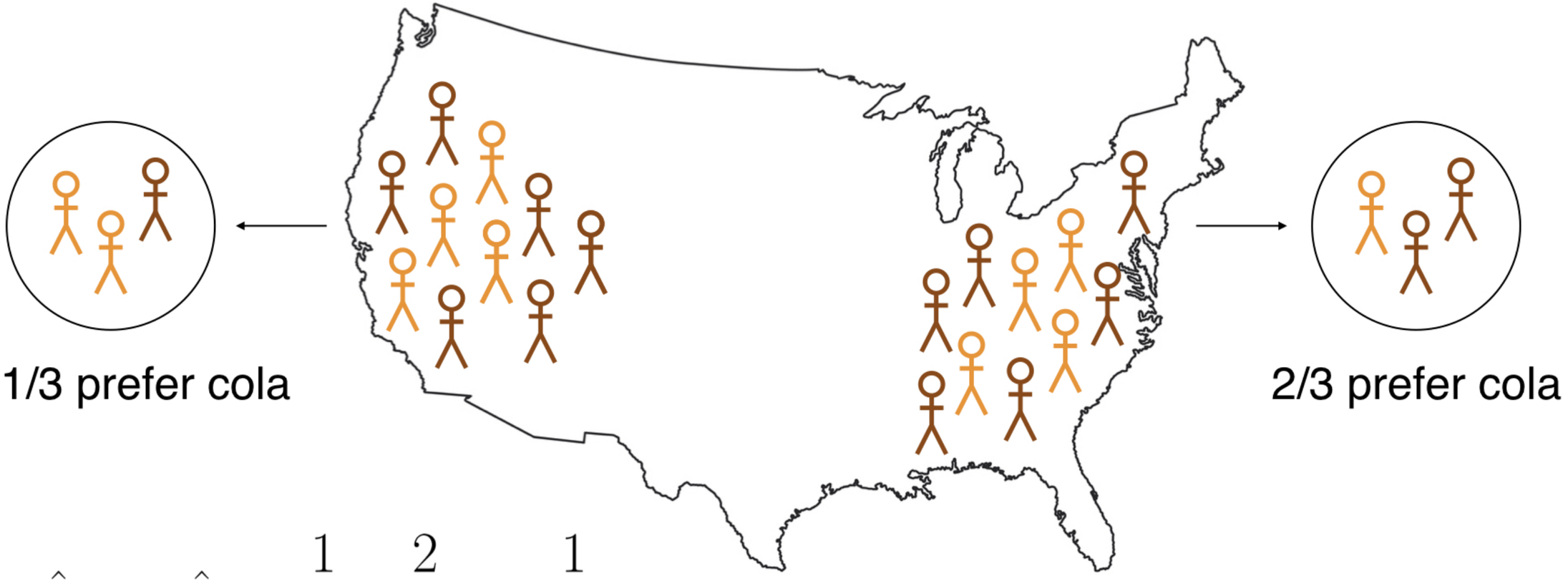


Logic of inference



$$p\hat{W} - p\hat{E} = \frac{2}{3} - \frac{2}{3} = 0$$

Logic of inference



$$\hat{p_W} - \hat{p_E} = \frac{1}{3} - \frac{2}{3} = -\frac{1}{3}$$

Understanding the null distribution

Generating a distribution of the statistic from the null population gives information about whether the observed data are inconsistent with the null hypothesis

Understanding the null distribution

Original data

Location	Cola	Orange
East	28	6
West	19	7

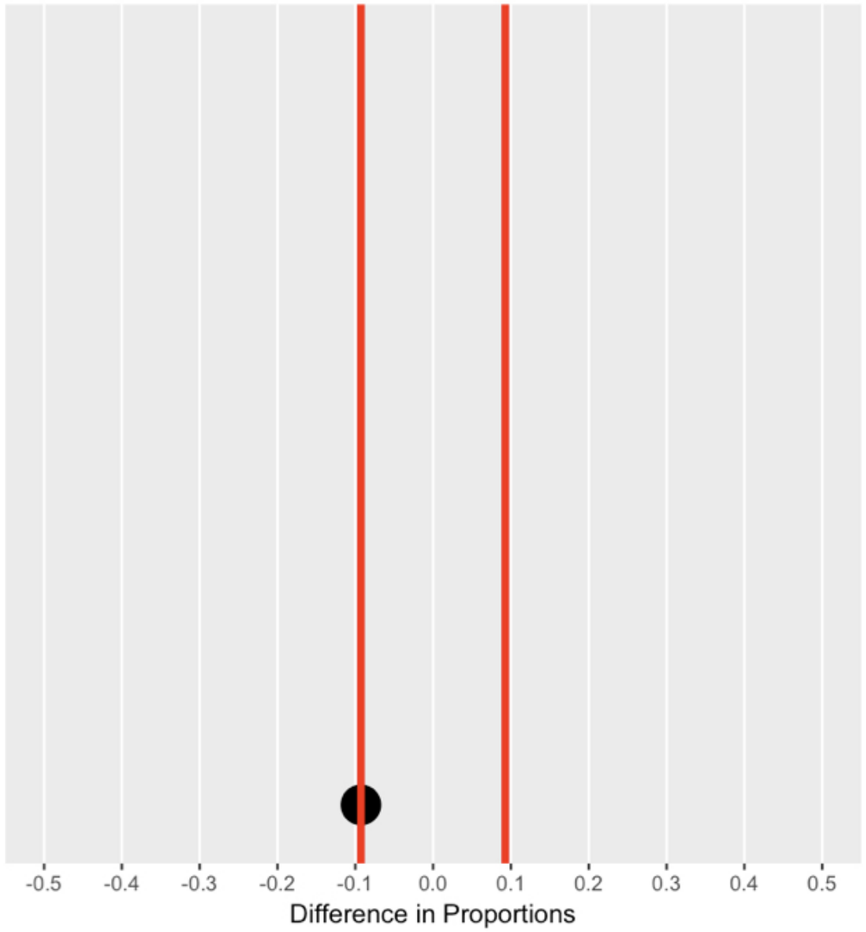
$$\hat{p}_{\text{east}} = 28/(28+6) = 0.82$$

$$\hat{p}_{\text{west}} = 19/(19+7) = 0.73$$

Understanding the null distribution

First shuffle, same as original

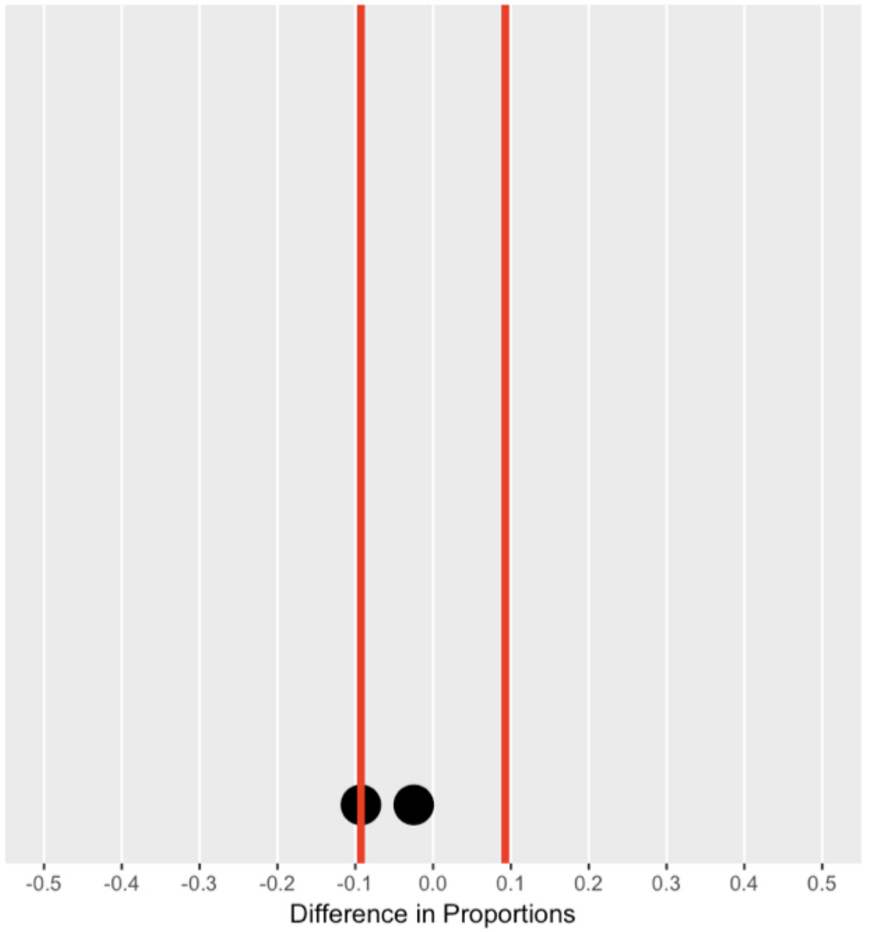
Location	Cola	Orange
East	28	6
West	19	7



Understanding the null distribution

Second shuffle

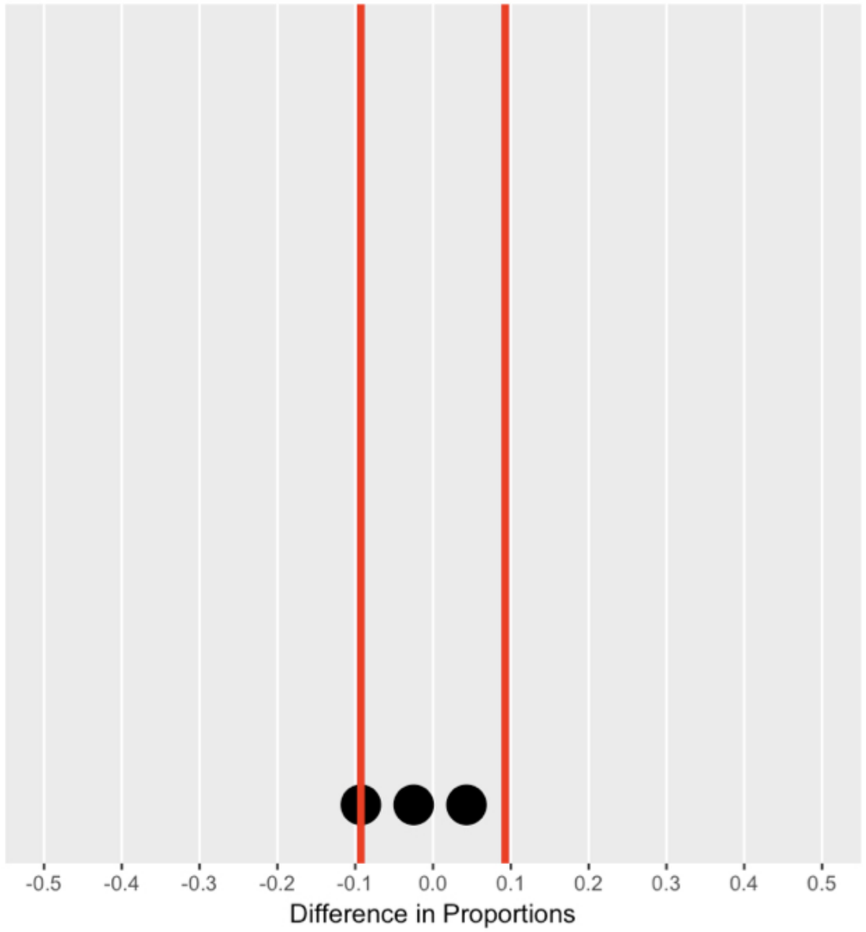
Location	Cola	Orange
East	27	7
West	20	6



Understanding the null distribution

Third shuffle

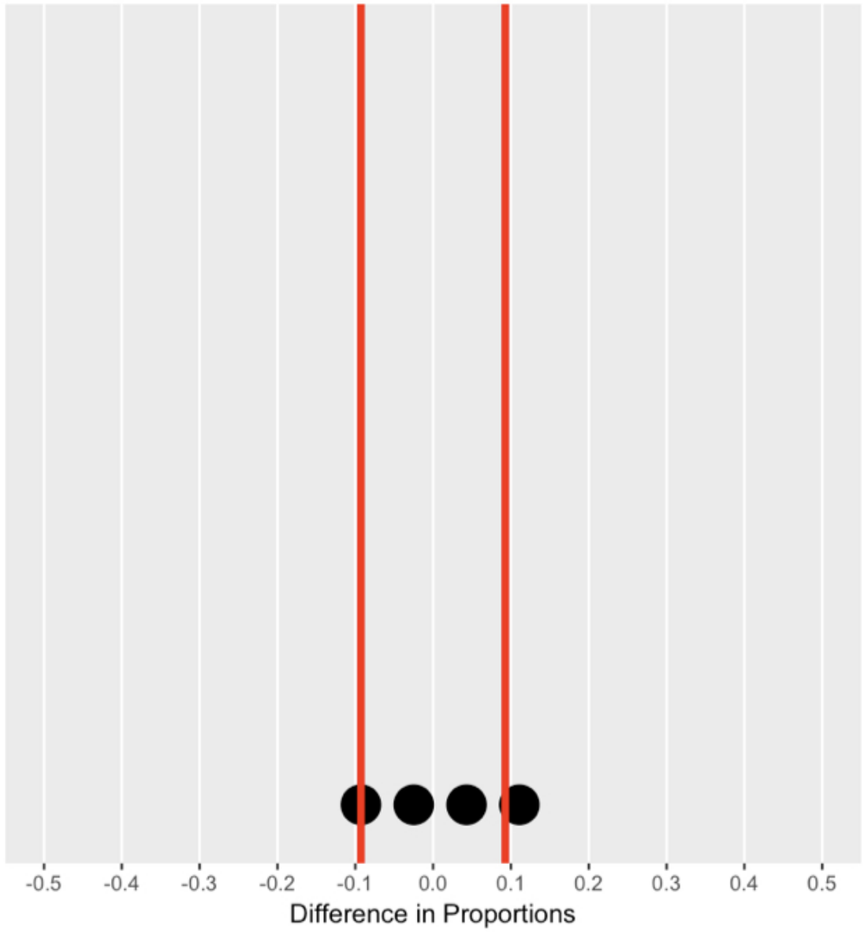
Location	Cola	Orange
East	28	8
West	21	5



Understanding the null distribution

Fourth shuffle

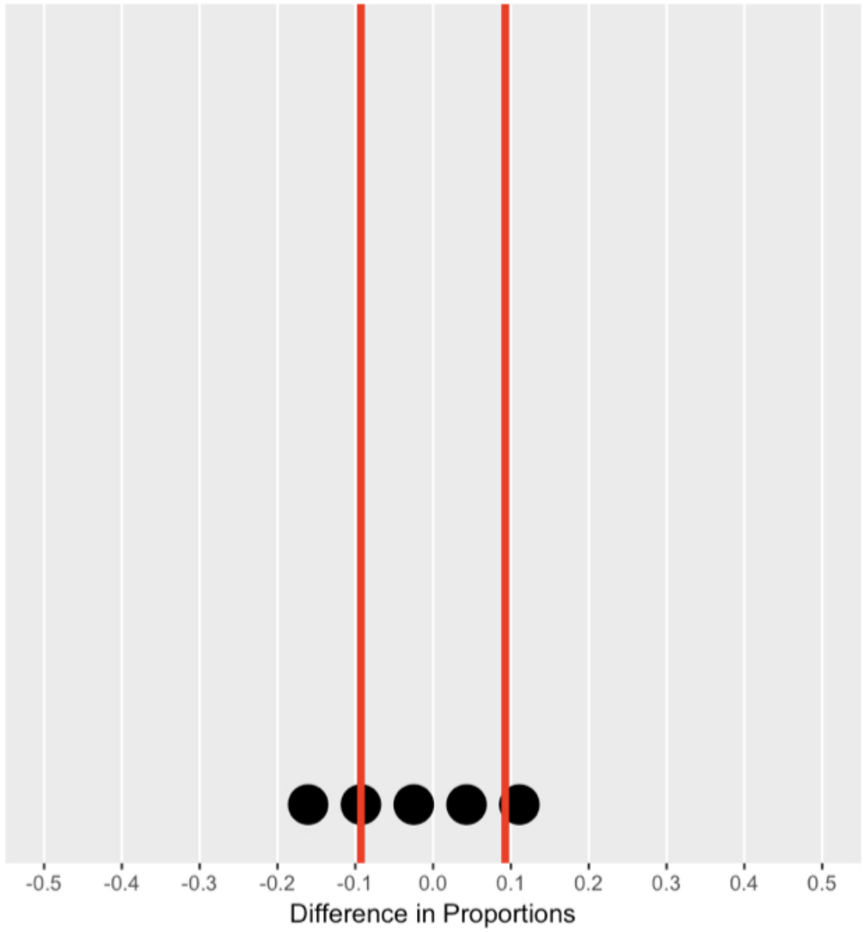
Location	Cola	Orange
East	25	9
West	22	4



Understanding the null distribution

Fifth shuffle

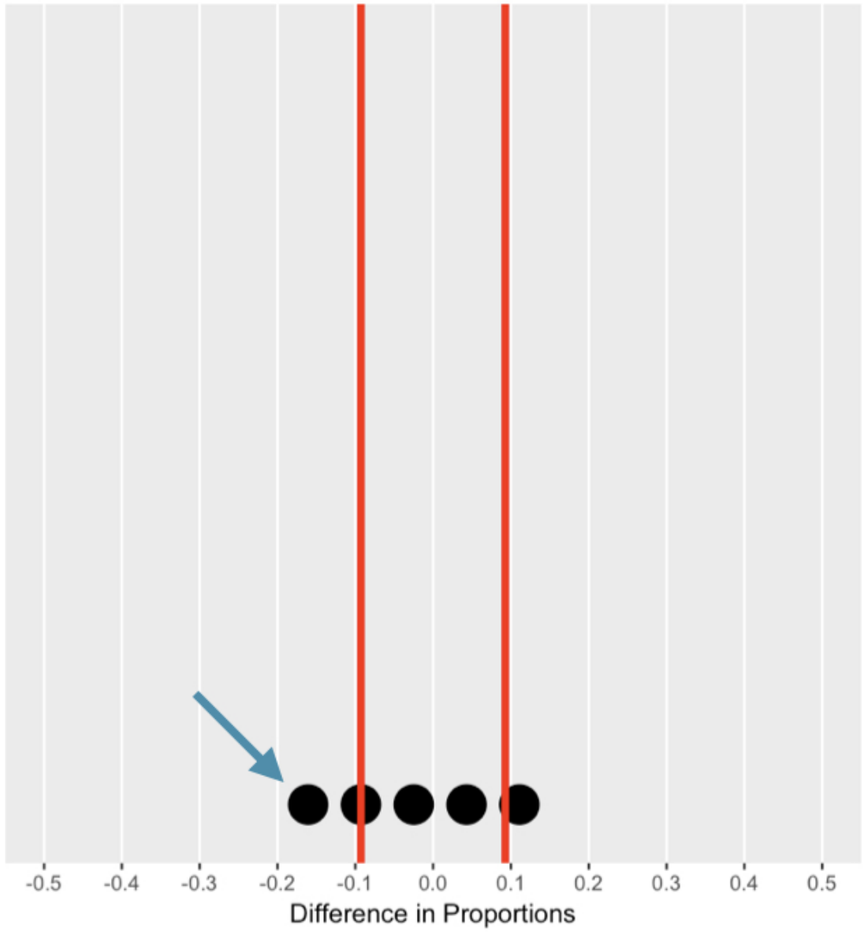
Location	Cola	Orange
East	29	5
West	18	8



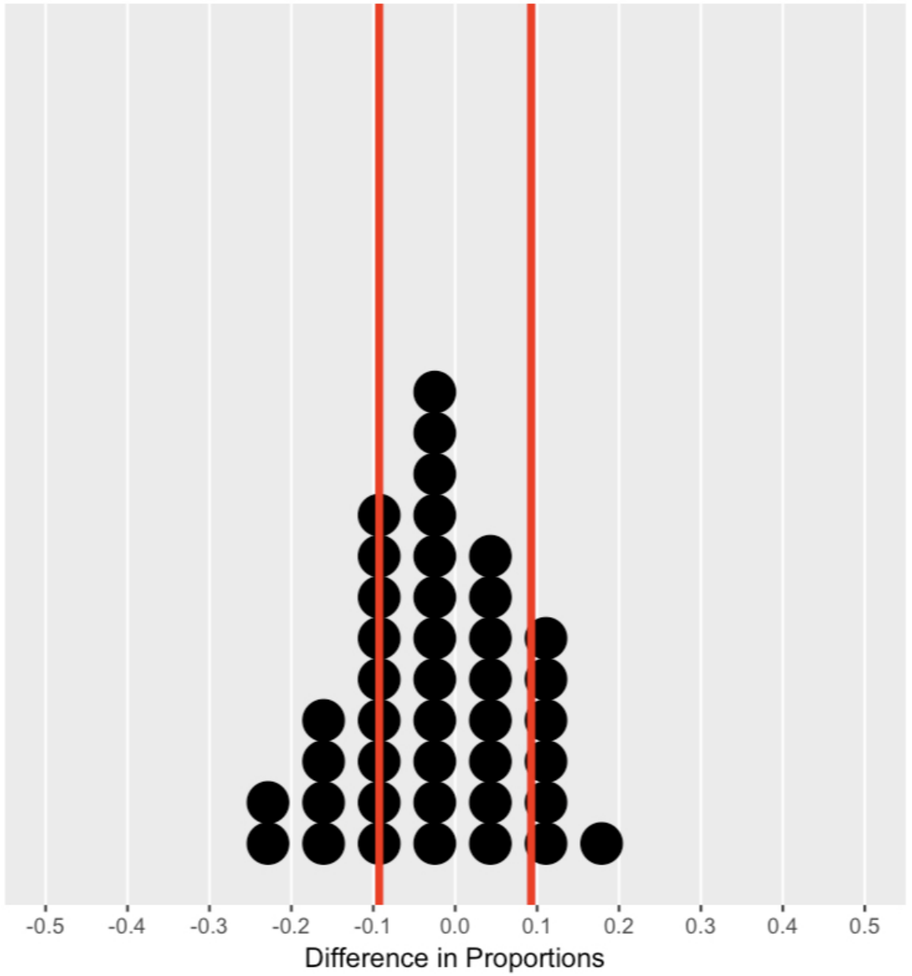
Understanding the null distribution

Fifth shuffle

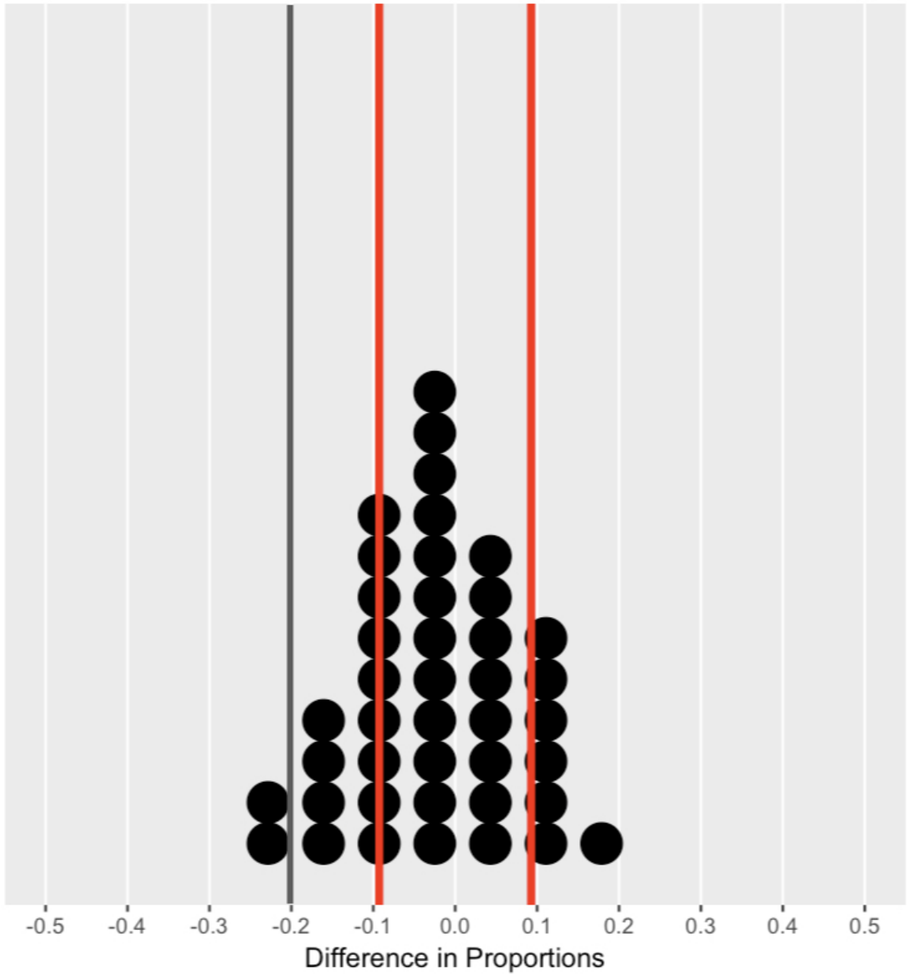
Location	Cola	Orange
East	29	5
West	18	8



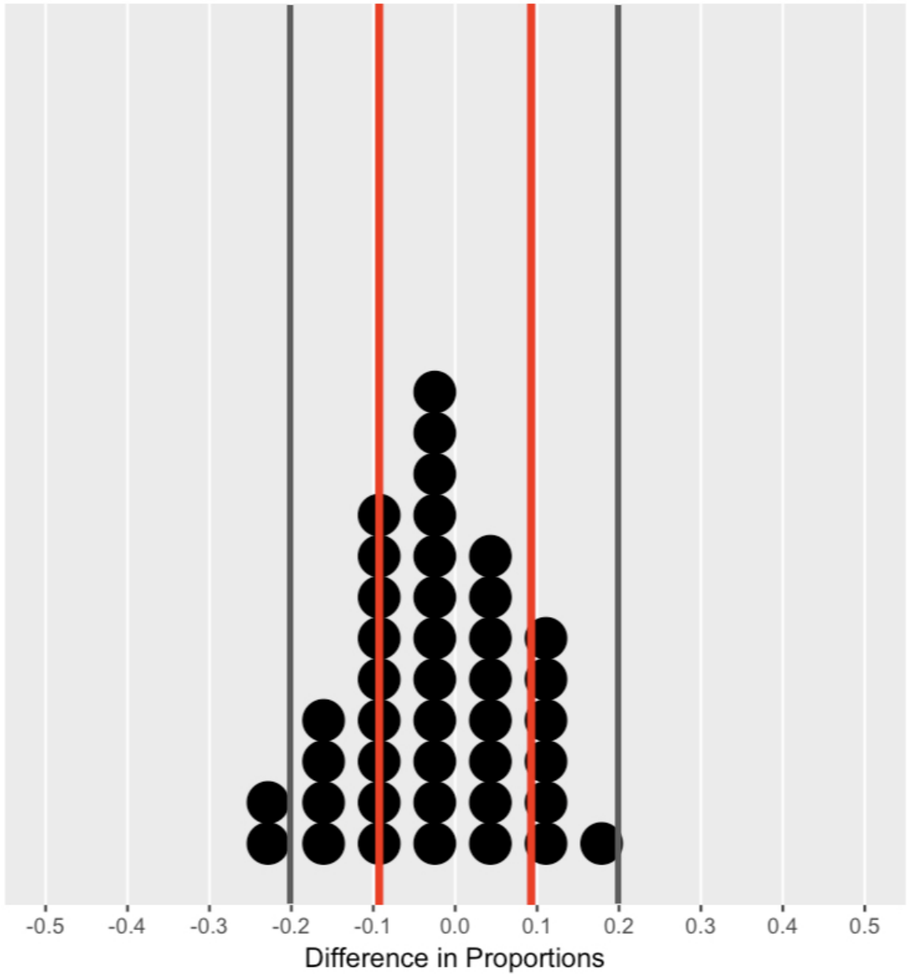
Understanding the null distribution



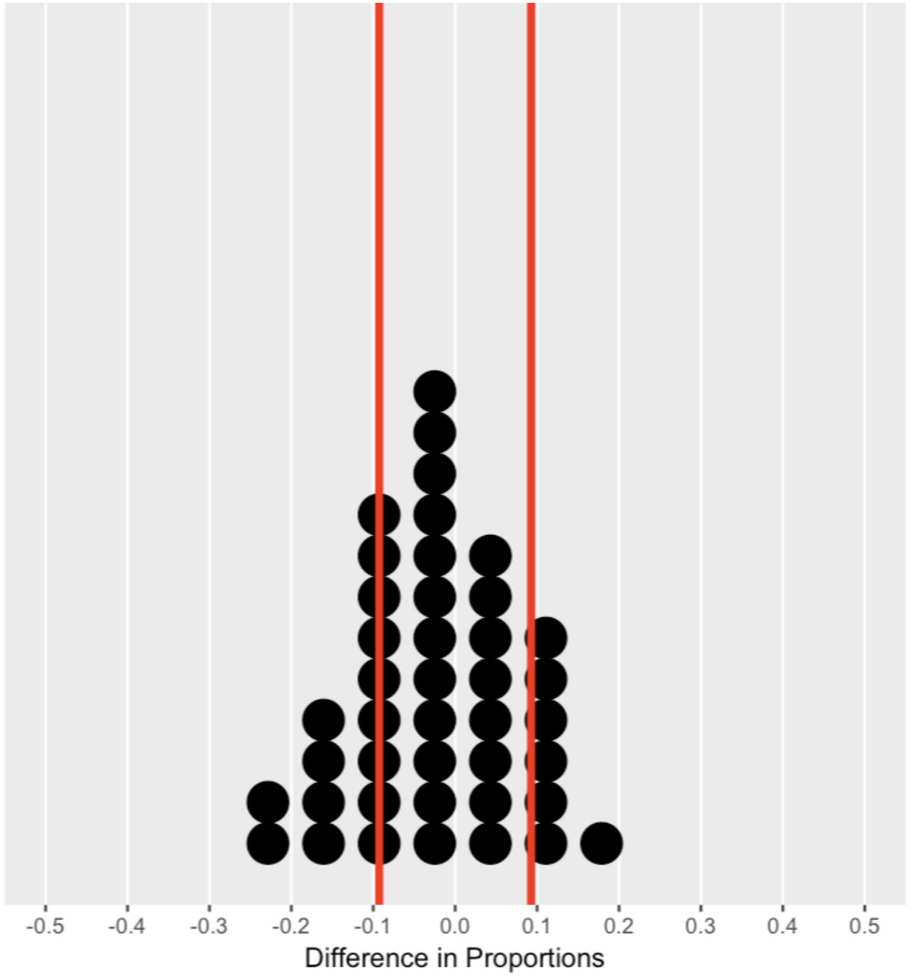
Understanding the null distribution



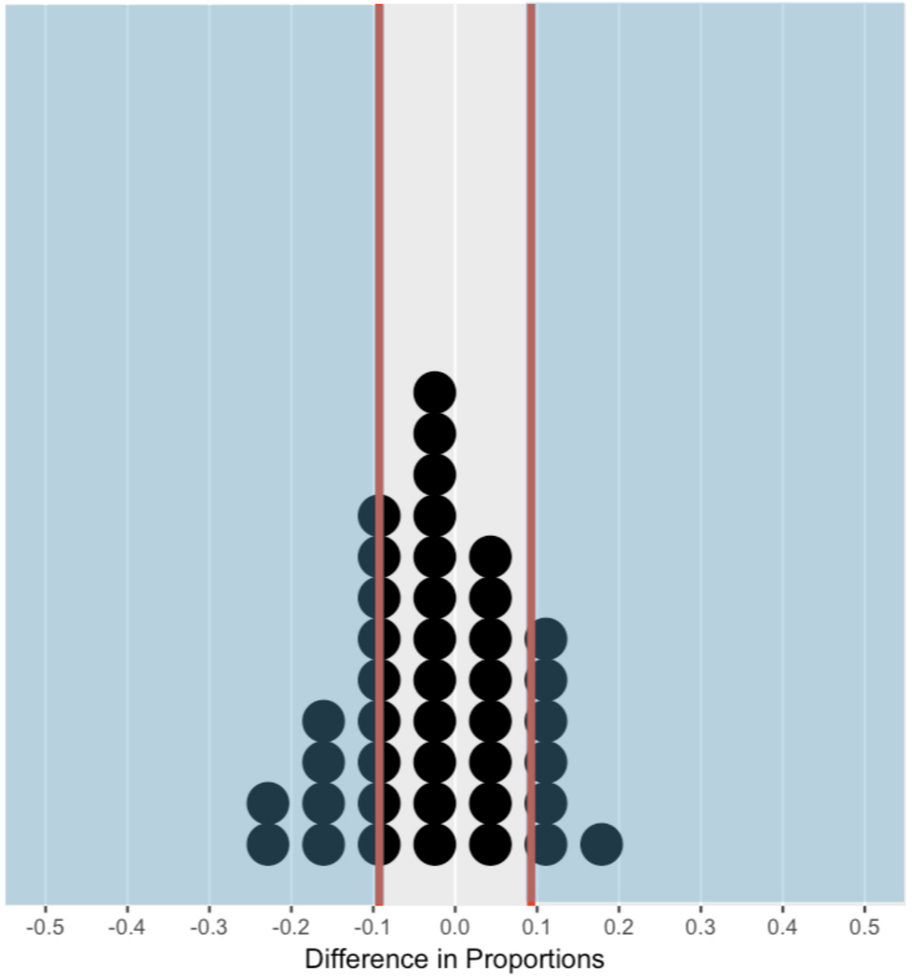
Understanding the null distribution



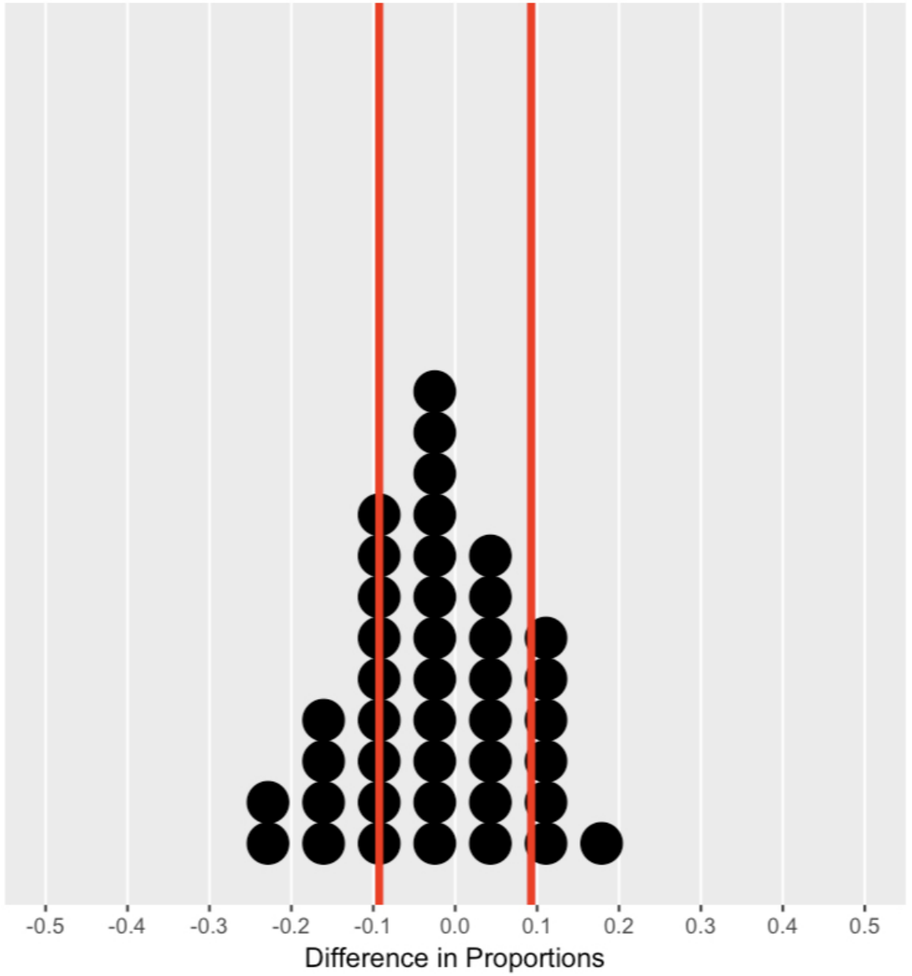
Understanding the null distribution



Understanding the null distribution



Understanding the null distribution



One random permutation

```
soda %>%
group_by(location) %>%
summarize(prop_cola =
    mean(drink == "cola")) %>%
summarize(diff(prop_cola))
```

```
# A tibble: 1 x 1
  `diff(prop_cola)`
              <dbl>
1        -0.09276018
```

```
library(infer)
soda %>% specify(drink ~ location,
                 success = "cola") %>%
hypothesize(null = "independence") %>%
generate(reps = 1, type = "permute") %>%
calculate(stat = "diff in props",
            order = c("west","east"))
```

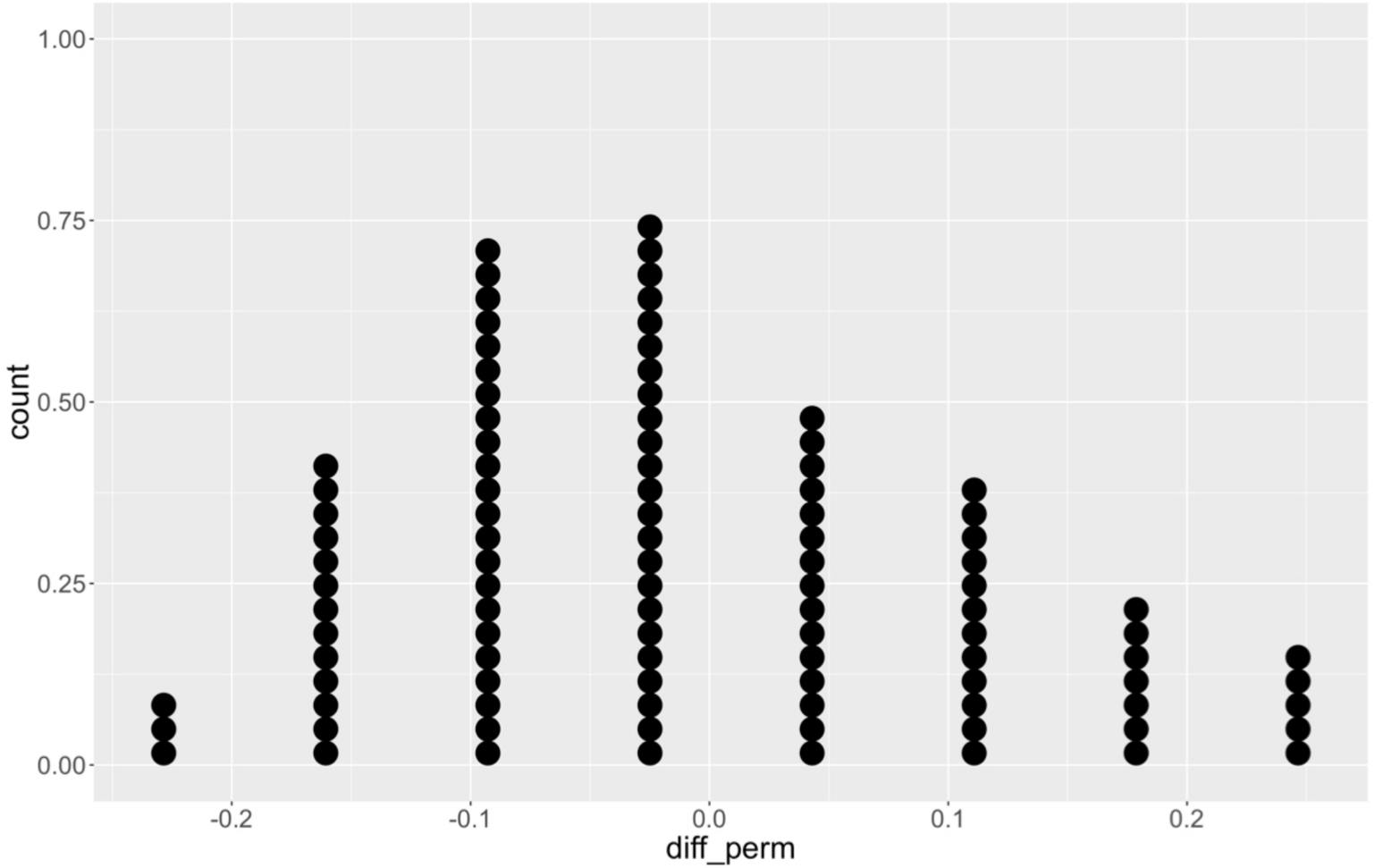
```
# A tibble: 1 x 2
  replicate        stat
      <int>       <dbl>
1         1   -0.02488688
```

Many random permutations

```
soda %>%
  specify(drink ~ location, success = "cola") %>%
  hypothesize(null = "independence") %>%
  generate(reps = 5, type = "permute") %>%
  calculate(stat = "diff in props", order = c("west", "east"))
```

```
# A tibble: 5 x 2
  replicate        stat
      <int>       <dbl>
1         1  0.04298643
2         2 -0.09276018
3         3  0.11085973
4         4  0.17873303
5         5 -0.16063348
```

Random distribution



Let's practice!

FOUNDATIONS OF INFERENCE

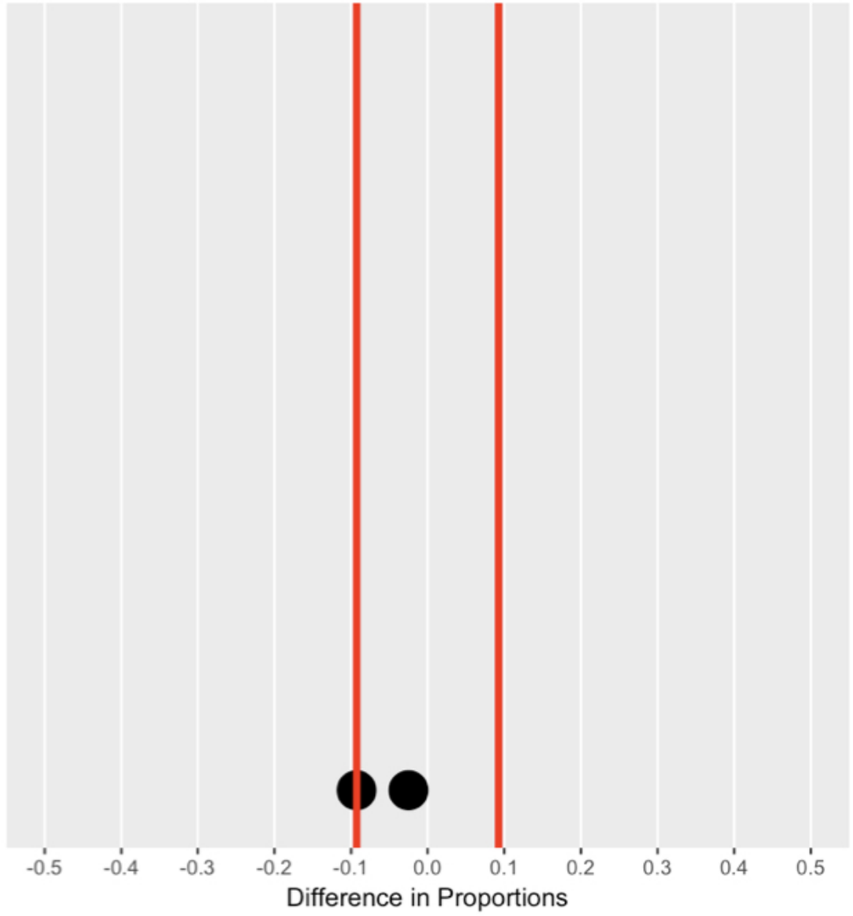
Using the randomization distribution

FOUNDATIONS OF INFERENCE

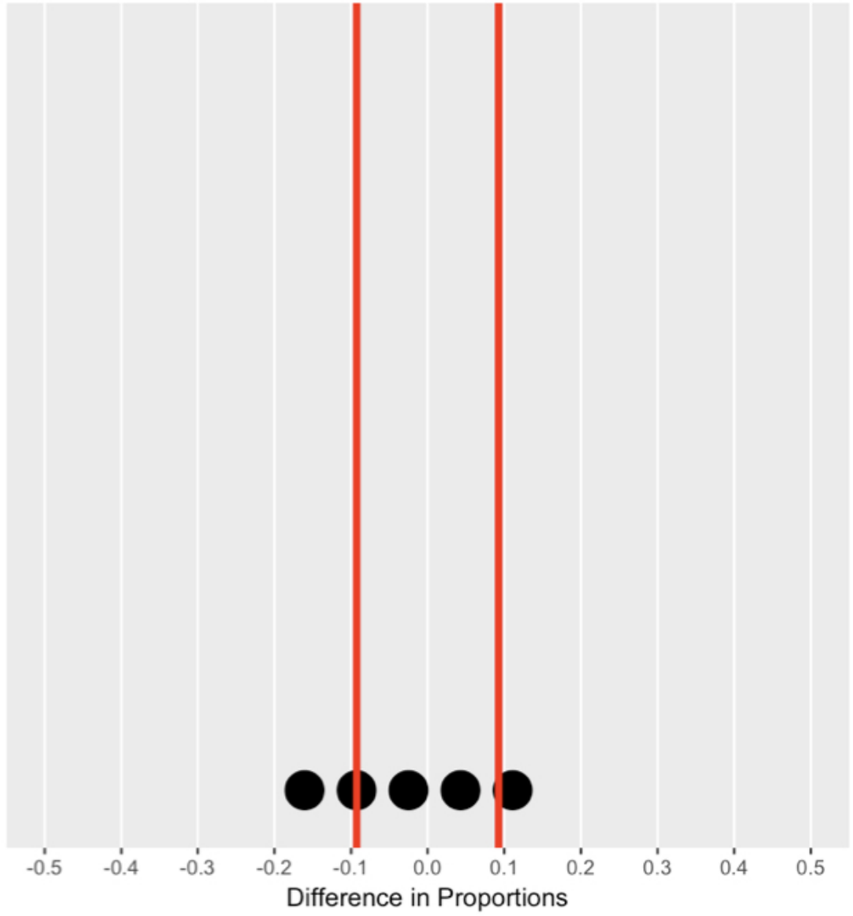
Jo Hardin
Instructor

datacamp

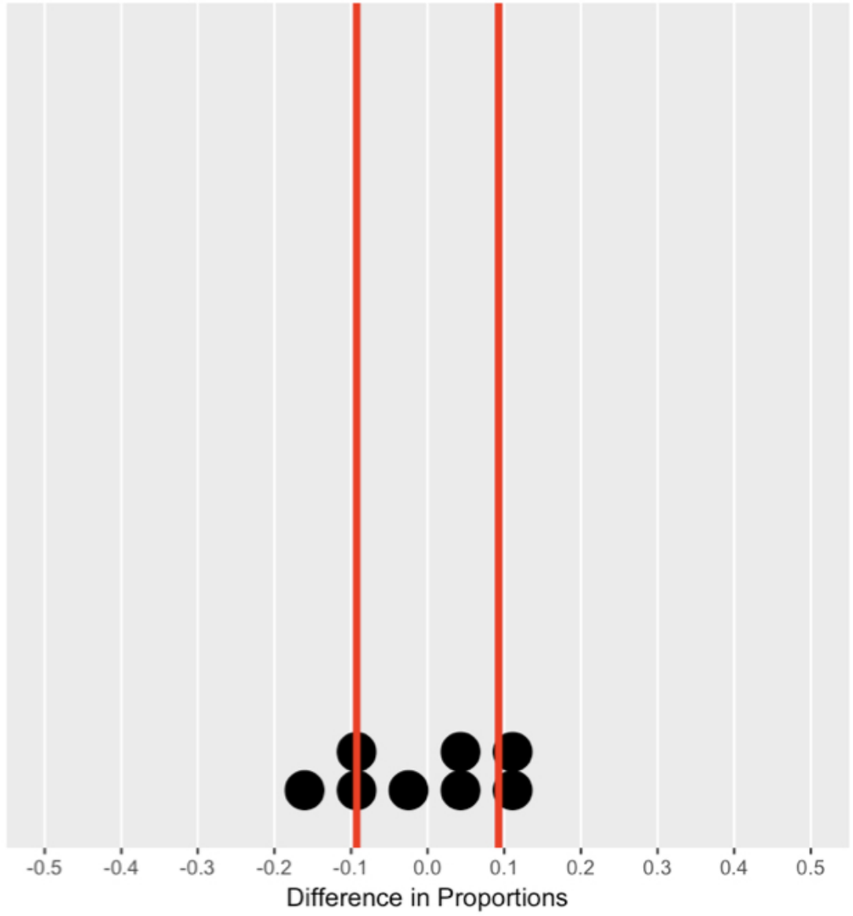
Understanding the null distribution



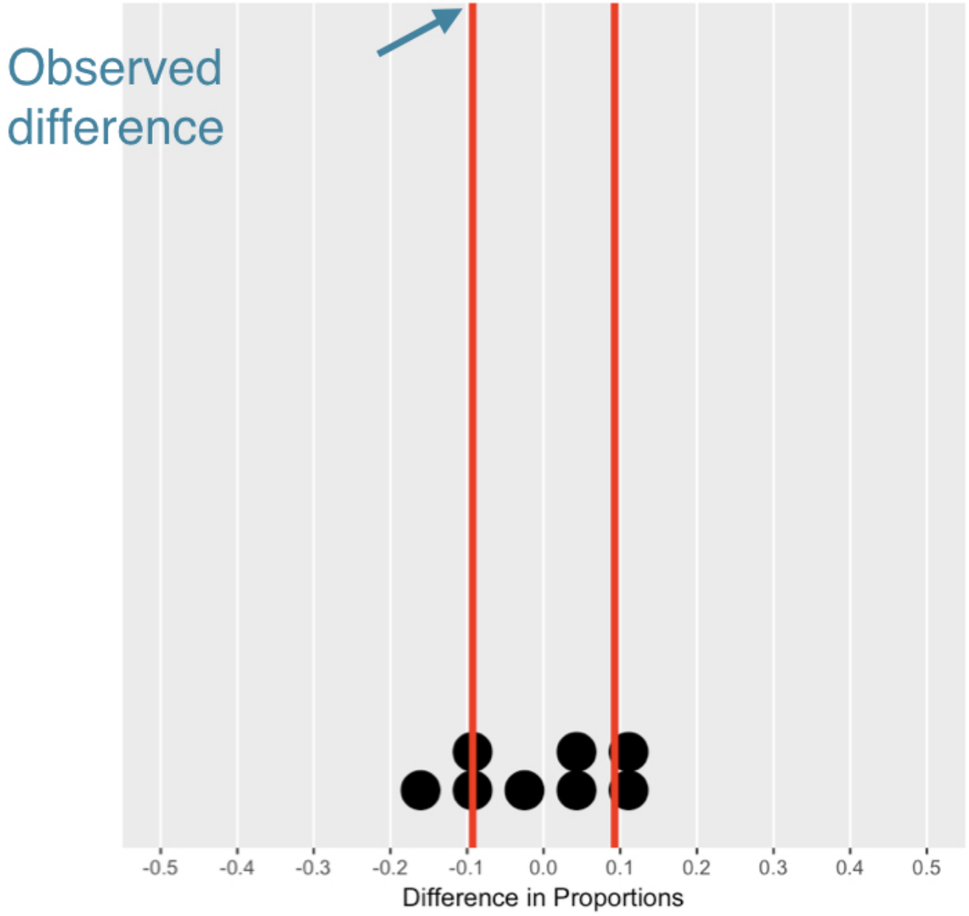
Understanding the null distribution



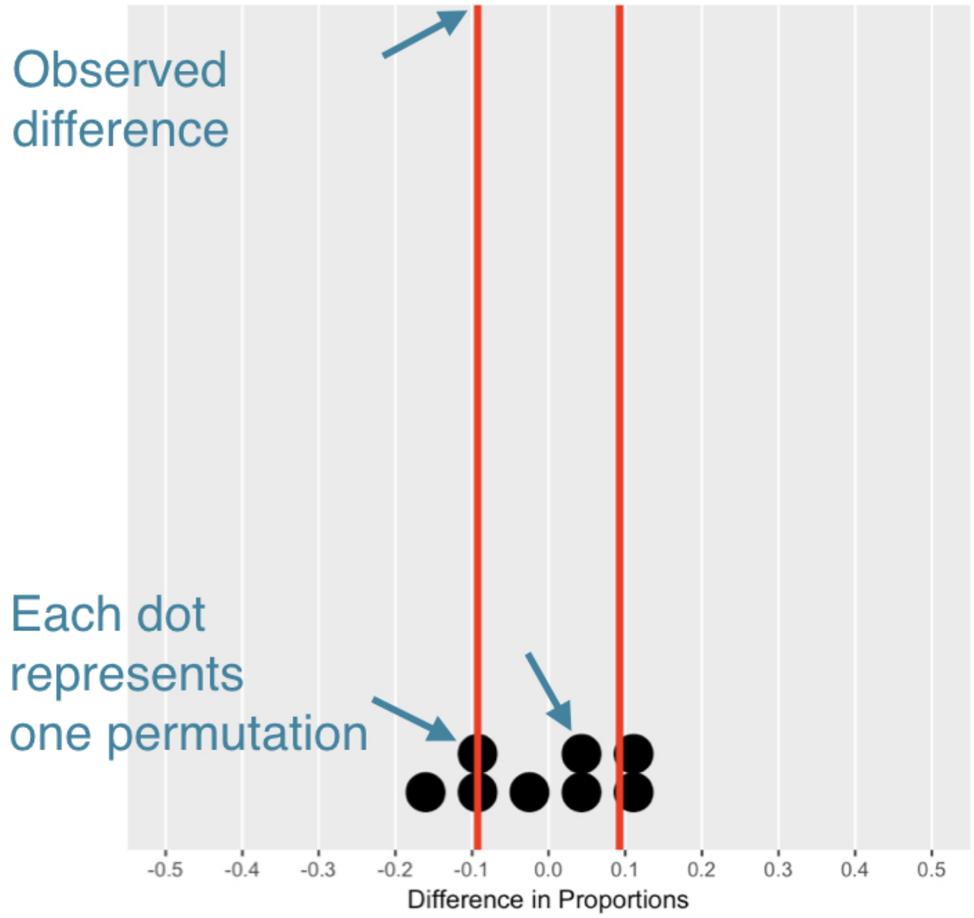
Understanding the null distribution



Understanding the null distribution



Understanding the null distribution



Data consistent with null?

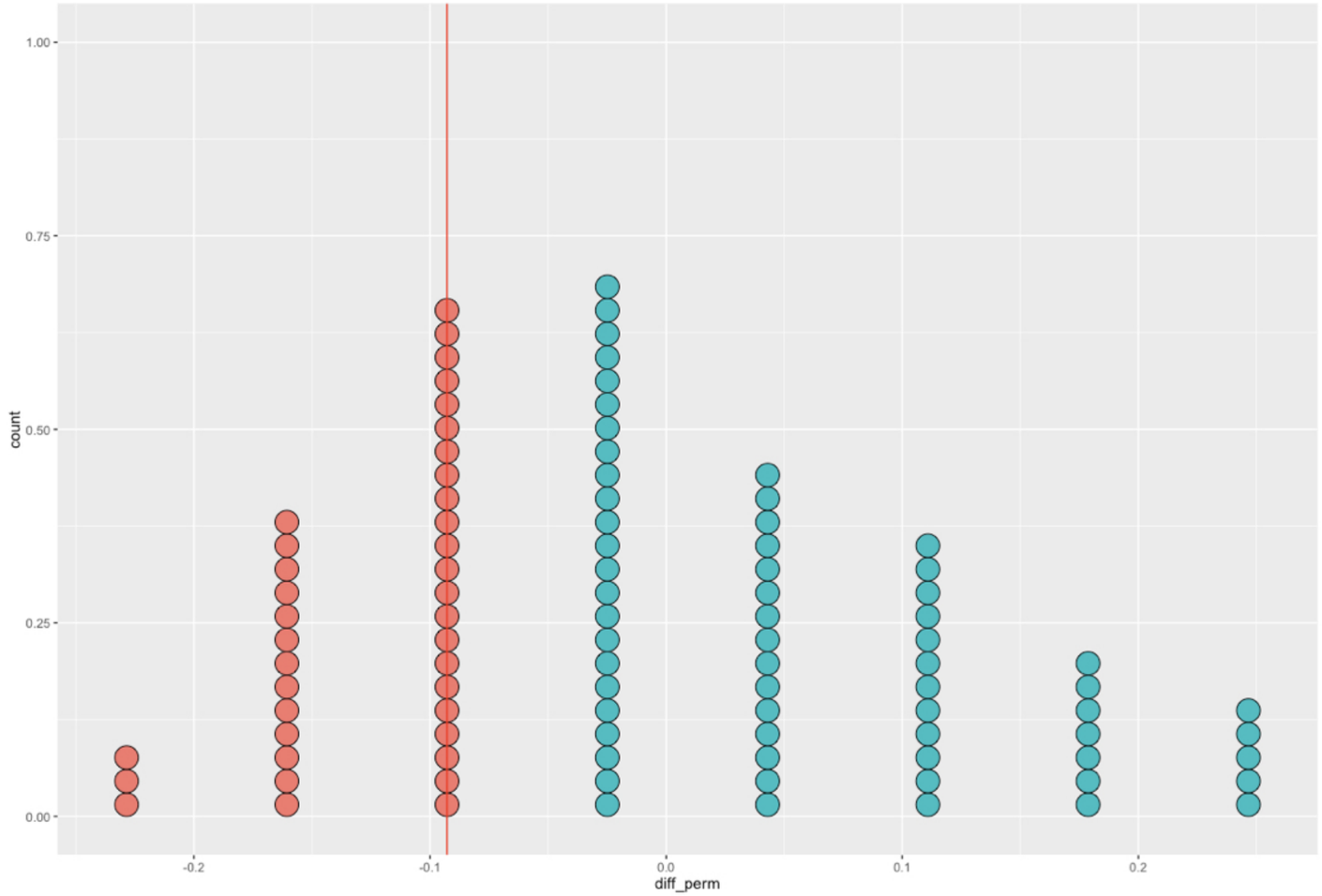
```
table(soda)
```

```
          location
drink     East West
cola      28   19
orange    6    7
```

```
soda %>% group_by(location) %>%
    summarize(mean(drink == "cola"))
```

```
# A tibble: 2 × 2
  location `mean(drink == "cola")`
    <fctr>                    <dbl>
1     East                0.8235294
2     West                0.7307692
```

Significance



How extreme are the observed data?

```
diff_orig <- soda %>%
  group_by(location) %>%
  summarize(prop_cola = mean(drink == "cola")) %>%
  summarize(diff(prop_cola)) %>%
  pull()

 soda_perm <- soda %>%
  specify(drink ~ location, success = "cola") %>%
  hypothesize(null = "independence") %>%
  generate(reps = 100, type = "permute") %>%
  calculate(stat = "diff in props",
              order = c("west", "east"))

soda_perm %>%
    summarize(proportion = mean(diff_orig >= stat))
```

```
 # A tibble: 1 x 1
  proportion
       <dbl>
1      0.380
```

Let's practice!

FOUNDATIONS OF INFERENCE

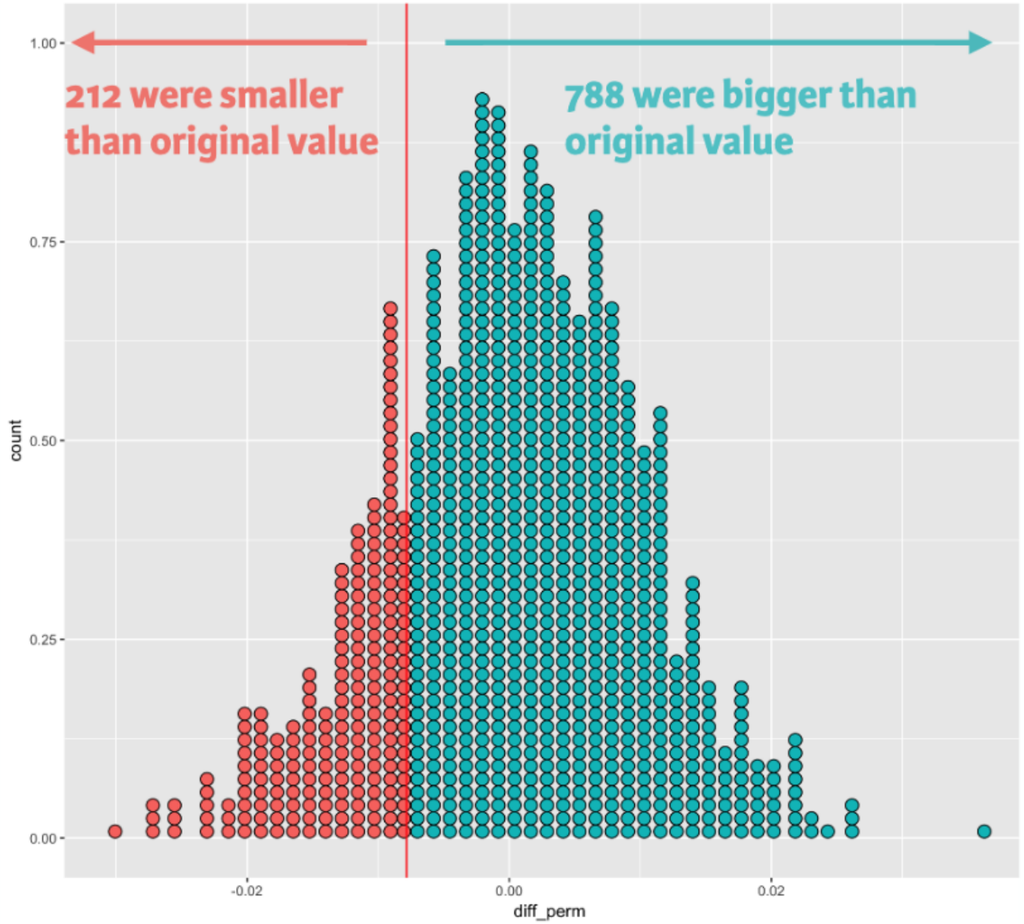
Study conclusions

FOUNDATIONS OF INFERENCE

Jo Hardin
Instructor

datacamp

Significance



We fail to reject the null hypothesis:

There is no evidence that our data are inconsistent with the null hypothesis

NHANES: random sample

- Representative sample of US population
- Conclusions from sample may apply to population
- Nothing to report in this case

Let's practice!

FOUNDATIONS OF INFERENCE