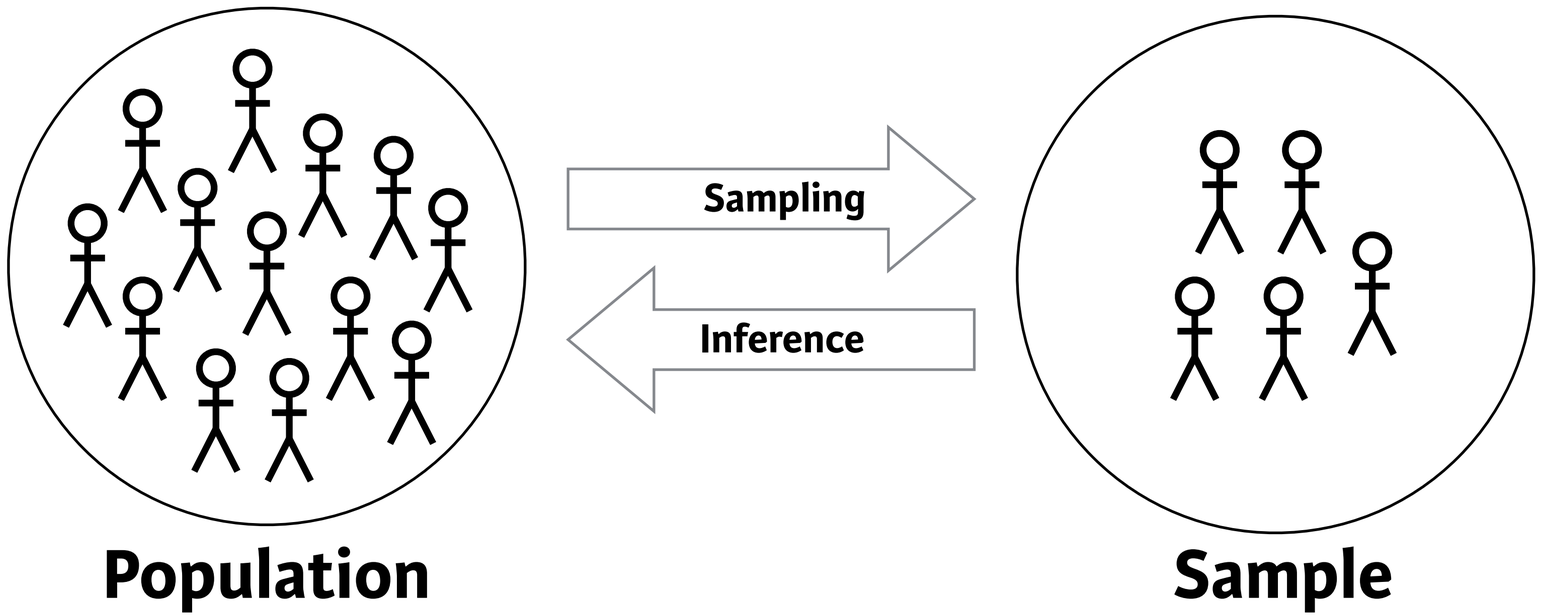
FOUNDATIONS OF INFERENCE

Welcome to the course!

What is statistical inference?

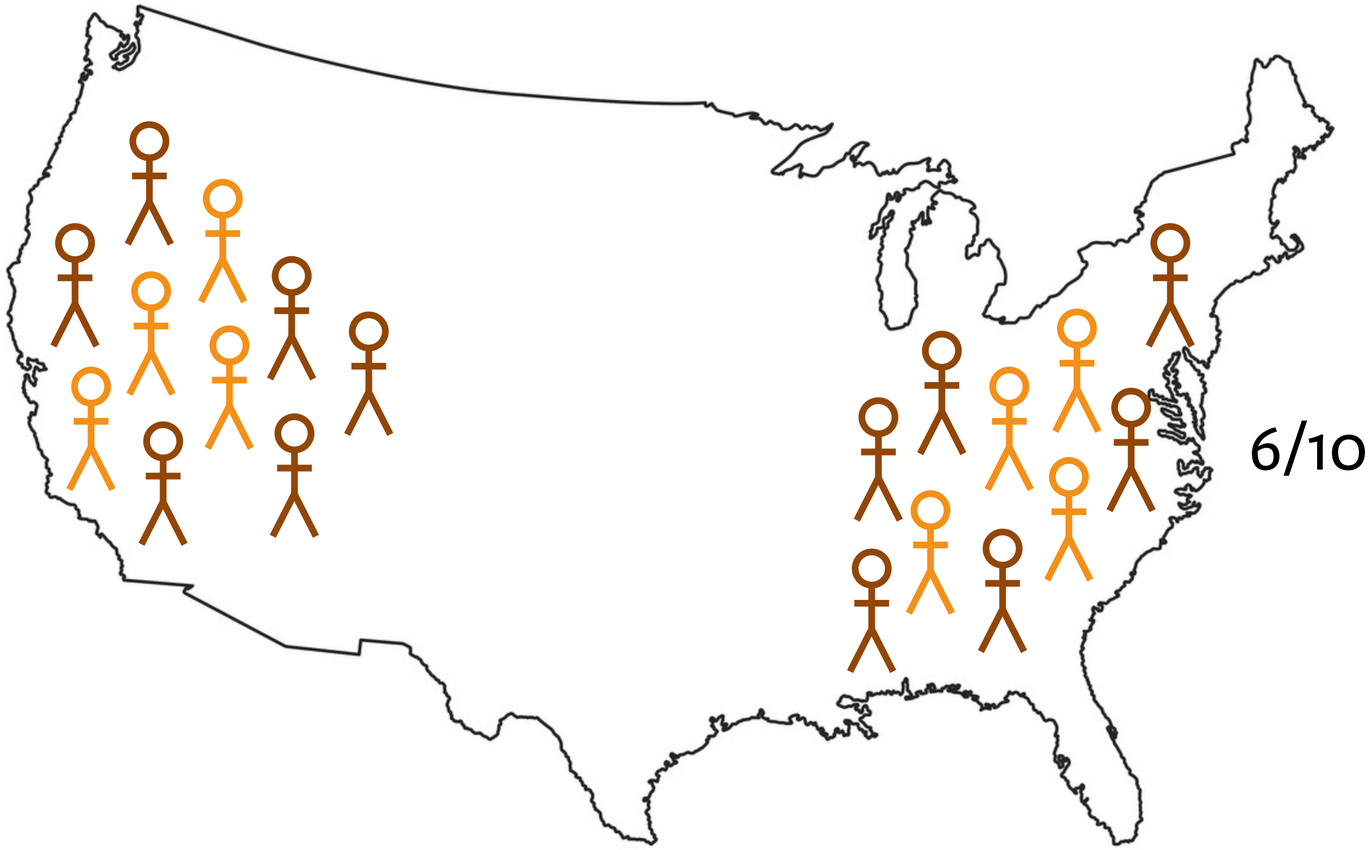
The process of making claims about a population based on information from a sample

What is statistical inference?



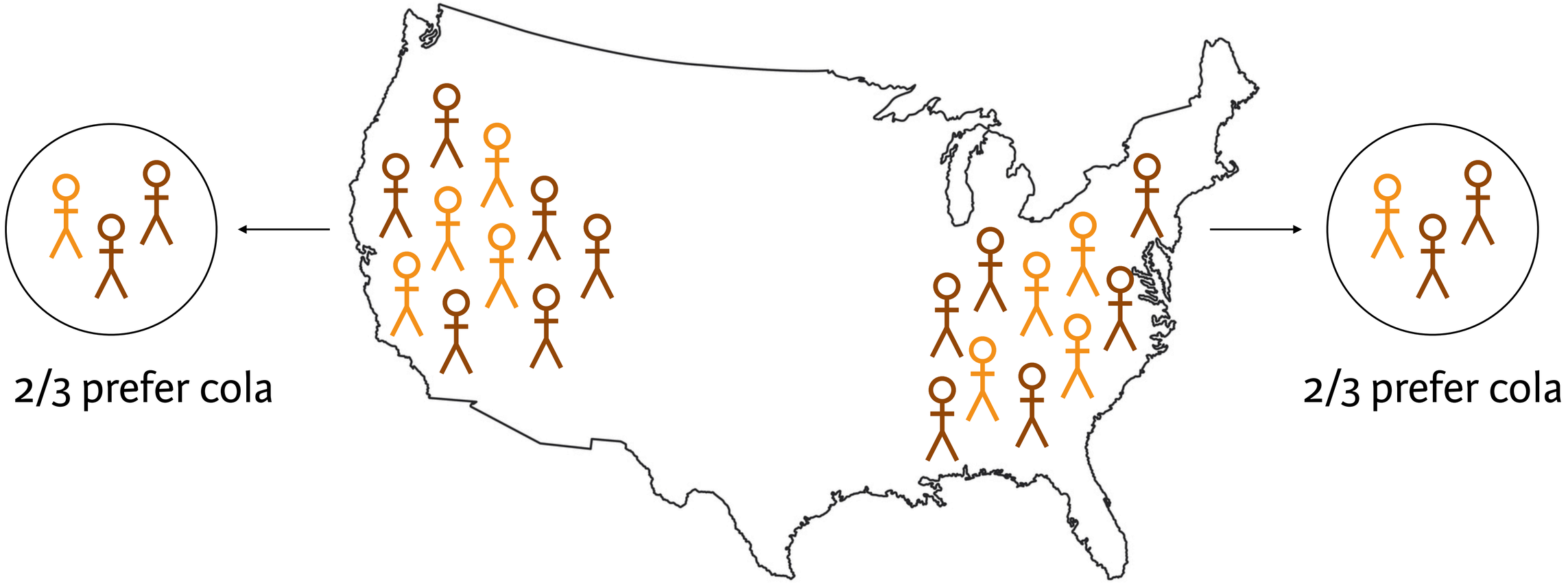
Assume two populations prefer cola at same rate

6/10 prefer cola

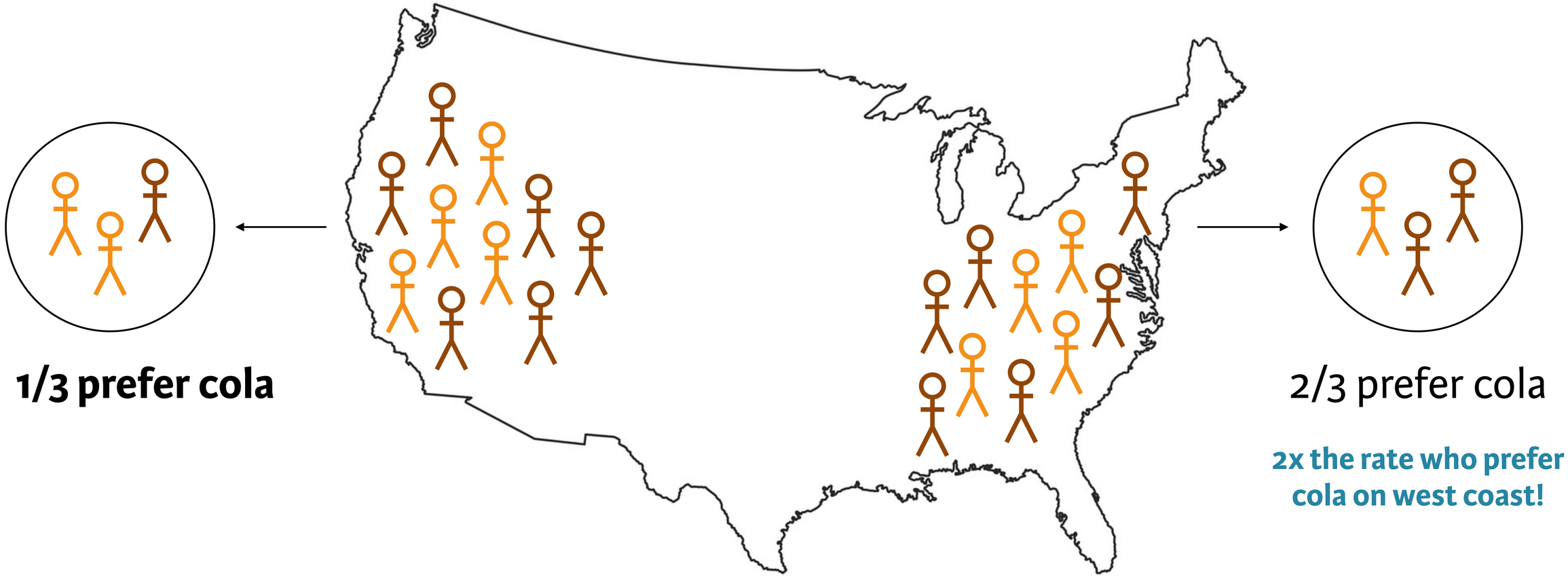


6/10 prefer cola

The sample data



The sample data (take 2)



1/3 prefer cola

2/3 prefer cola

2x the rate who prefer cola on west coast!

Vocabulary

- **Null hypothesis (H_0)**: The claim that is not interesting
- **Alternative hypothesis (H_A)**: The claim corresponding to the research hypothesis

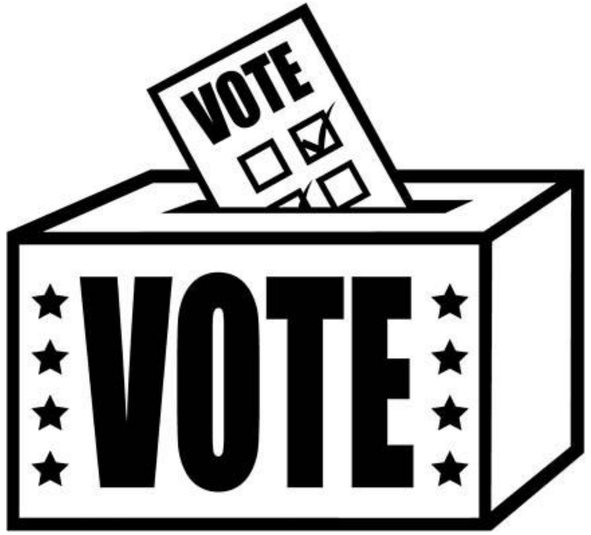
The "goal" is to disprove the null hypothesis

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 **A population measure!**
- H_0: Candidate X will get half the votes
- H_A: Candidate X will get more than half the votes



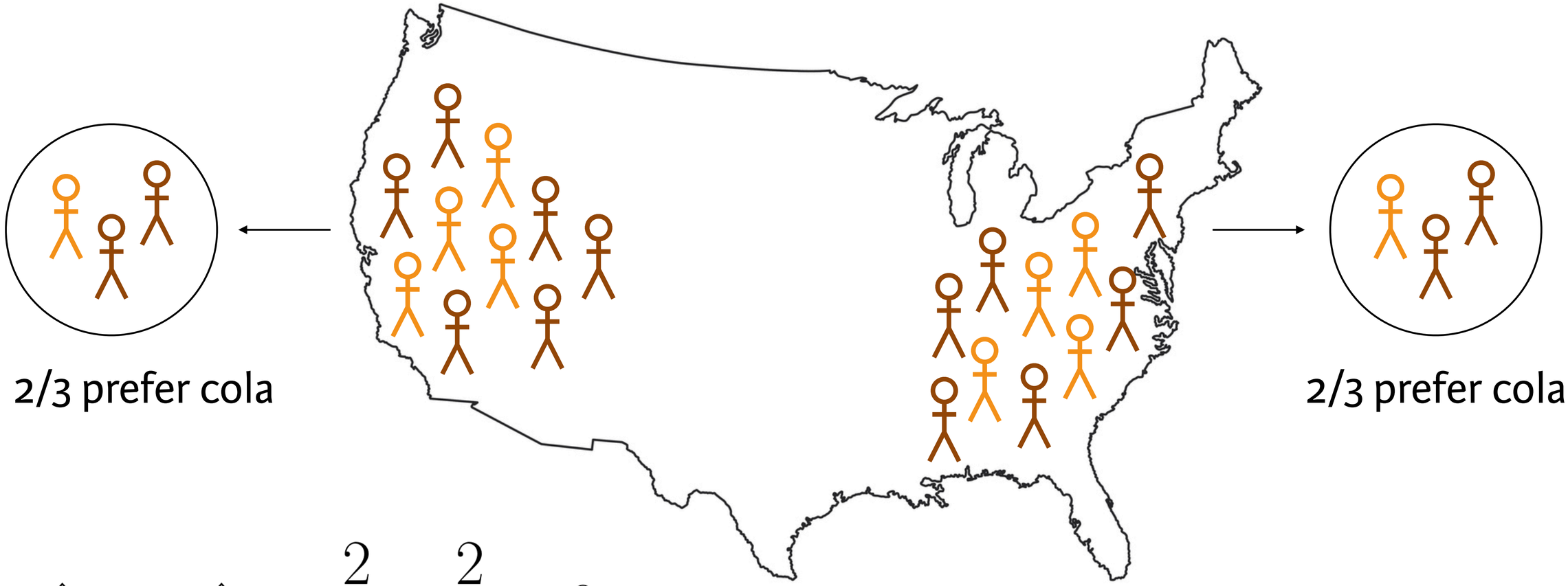
FOUNDATIONS OF INFERENCE

Let’s practice!

FOUNDATIONS OF INFERENCE

Randomized distributions

Logic of inference

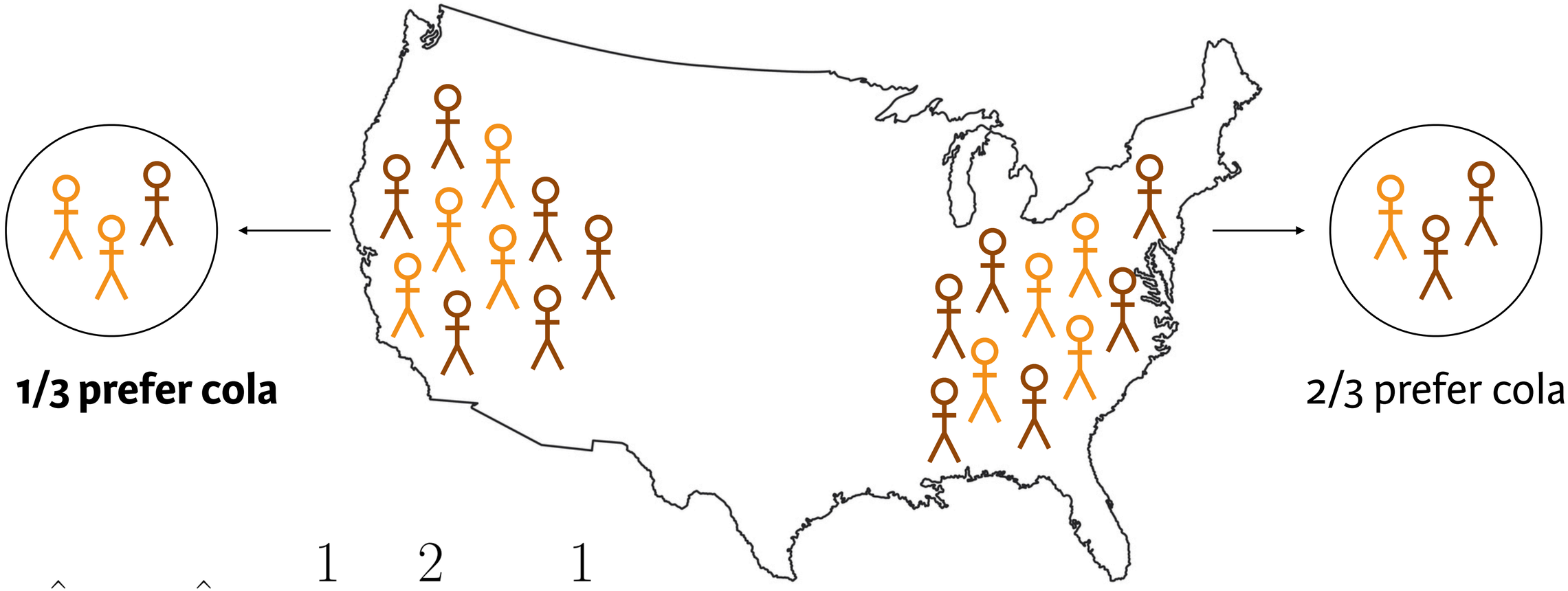


2/3 prefer cola

2/3 prefer cola

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

Logic of inference



1/3 prefer cola

2/3 prefer cola

$$p\hat{W} - p\hat{E} = \frac{1}{3} - \frac{2}{3} = -\frac{1}{3}$$

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

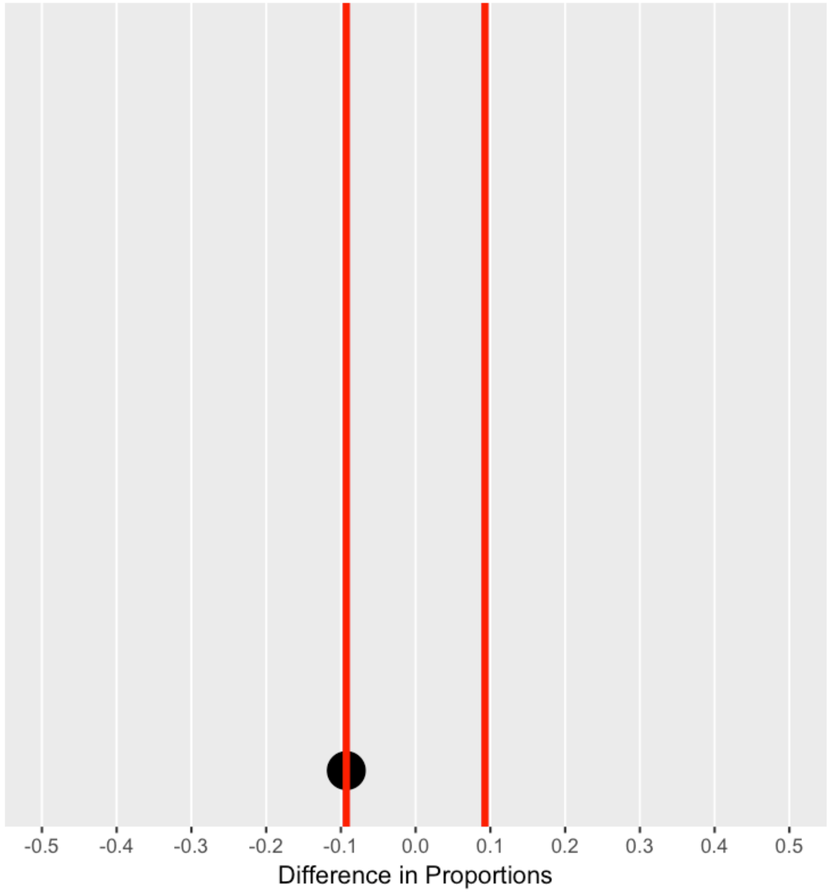
28/(28 + 6) = 0.82

19/(19 + 7) = 0.73

Understanding the null distribution

First shuffle, same as original

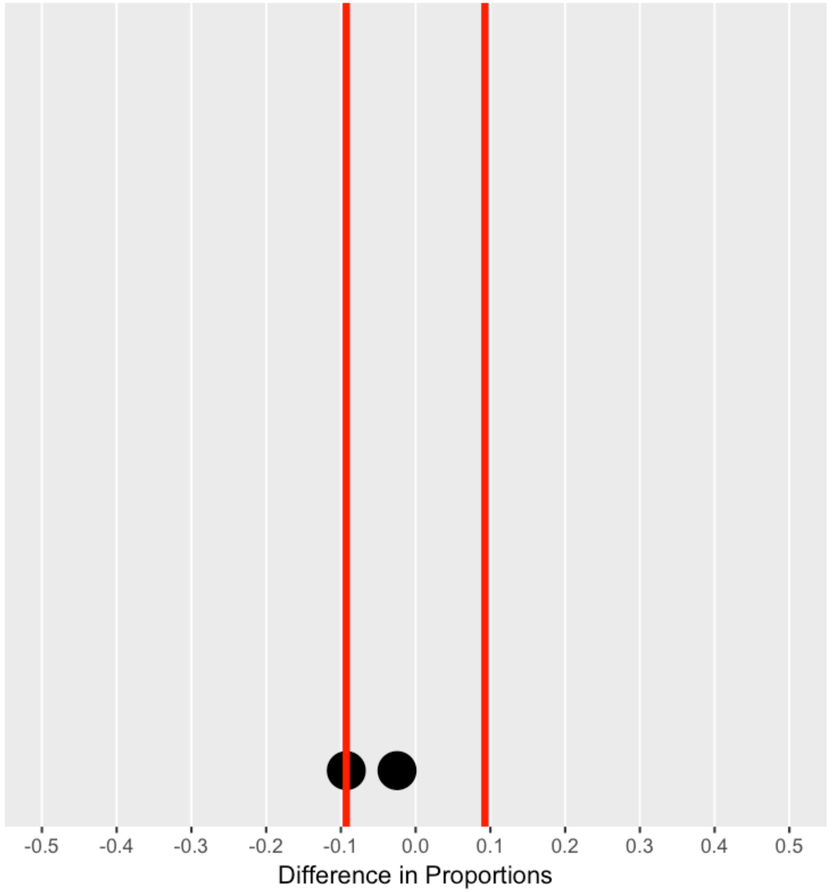
Location	Cola	Orange
East	28	6
West	19	7



Understanding the null distribution

Second shuffle

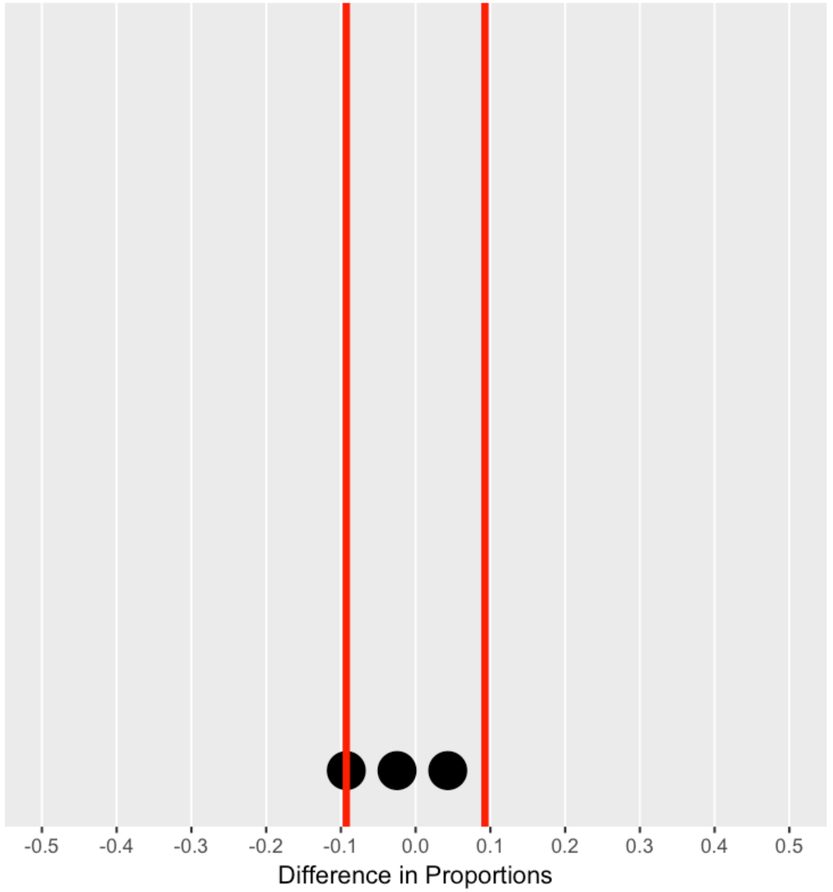
Location	Cola	Orange
East	27	7
West	20	6



Understanding the null distribution

Third shuffle

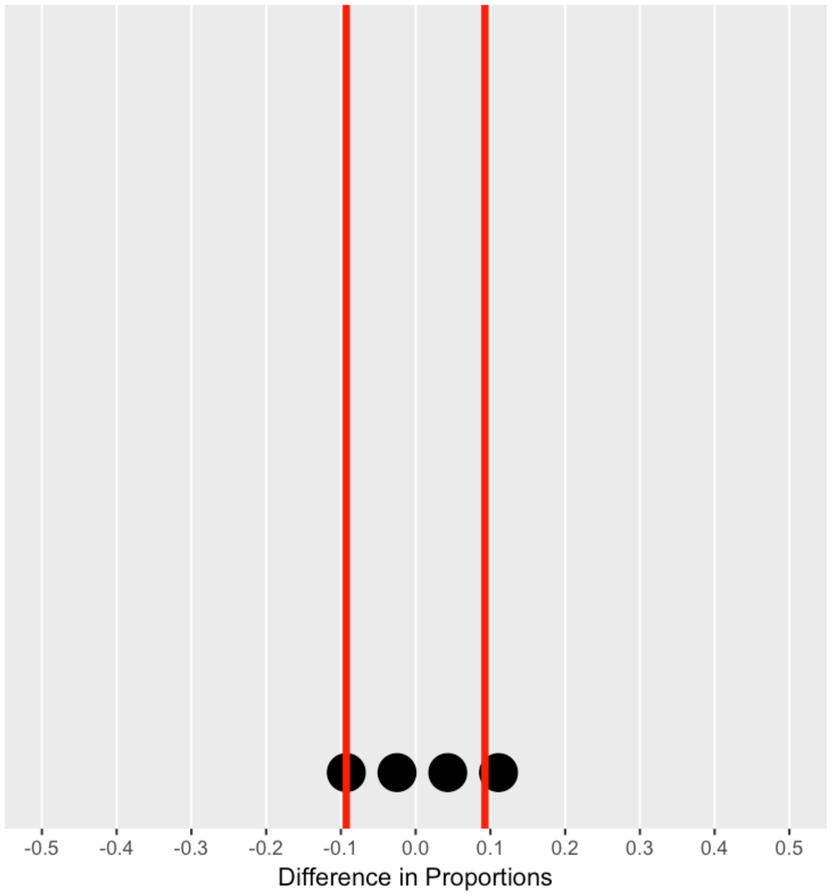
Location	Cola	Orange
East	28	8
West	21	5



Understanding the null distribution

Fourth shuffle

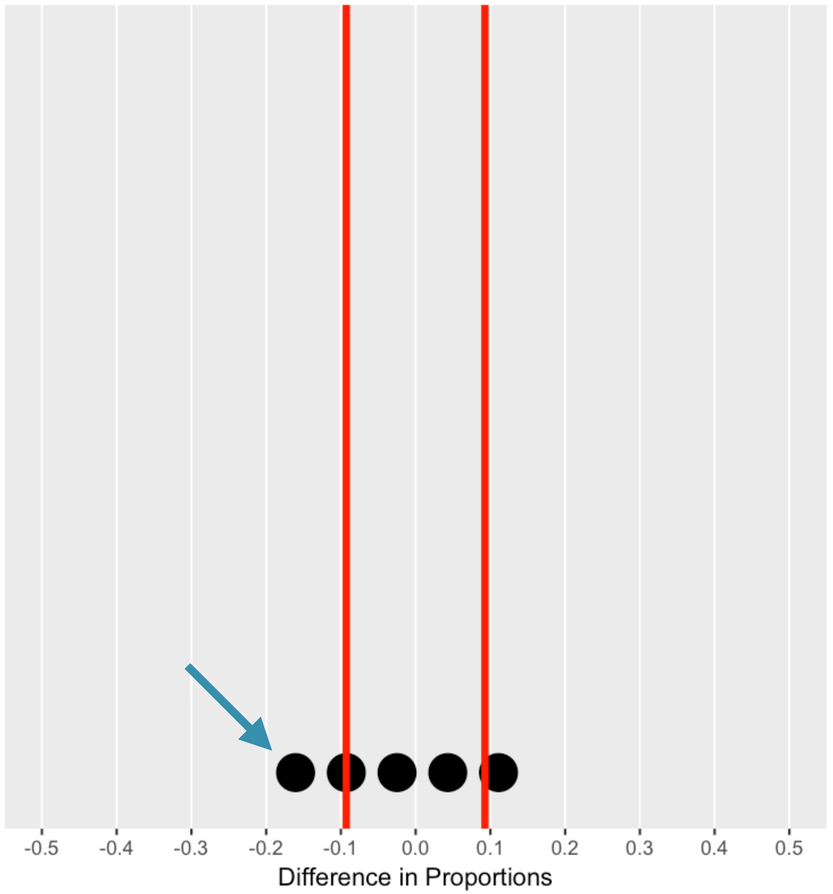
Location	Cola	Orange
East	25	9
West	22	4



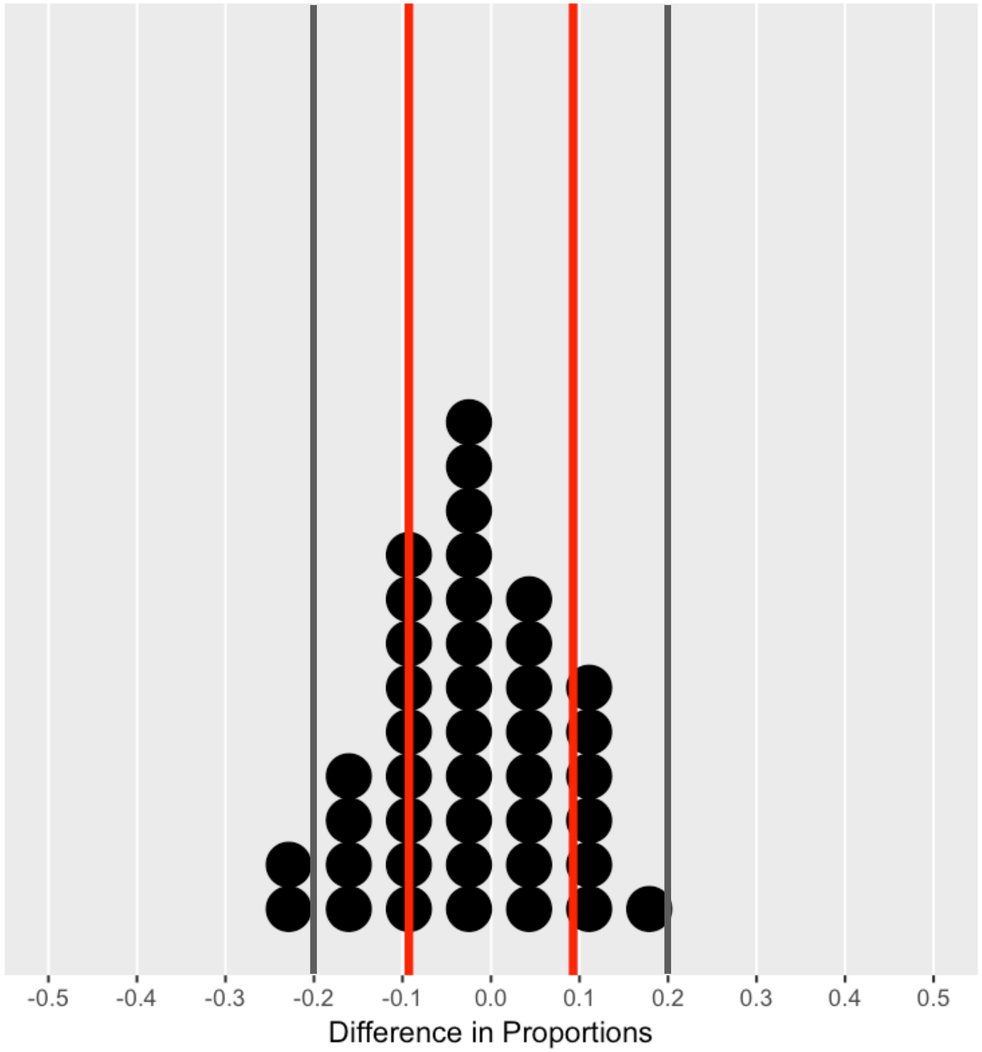
Understanding the null distribution

Fifth shuffle

Location	Cola	Orange
East	29	5
West	18	8



Understanding the null distribution



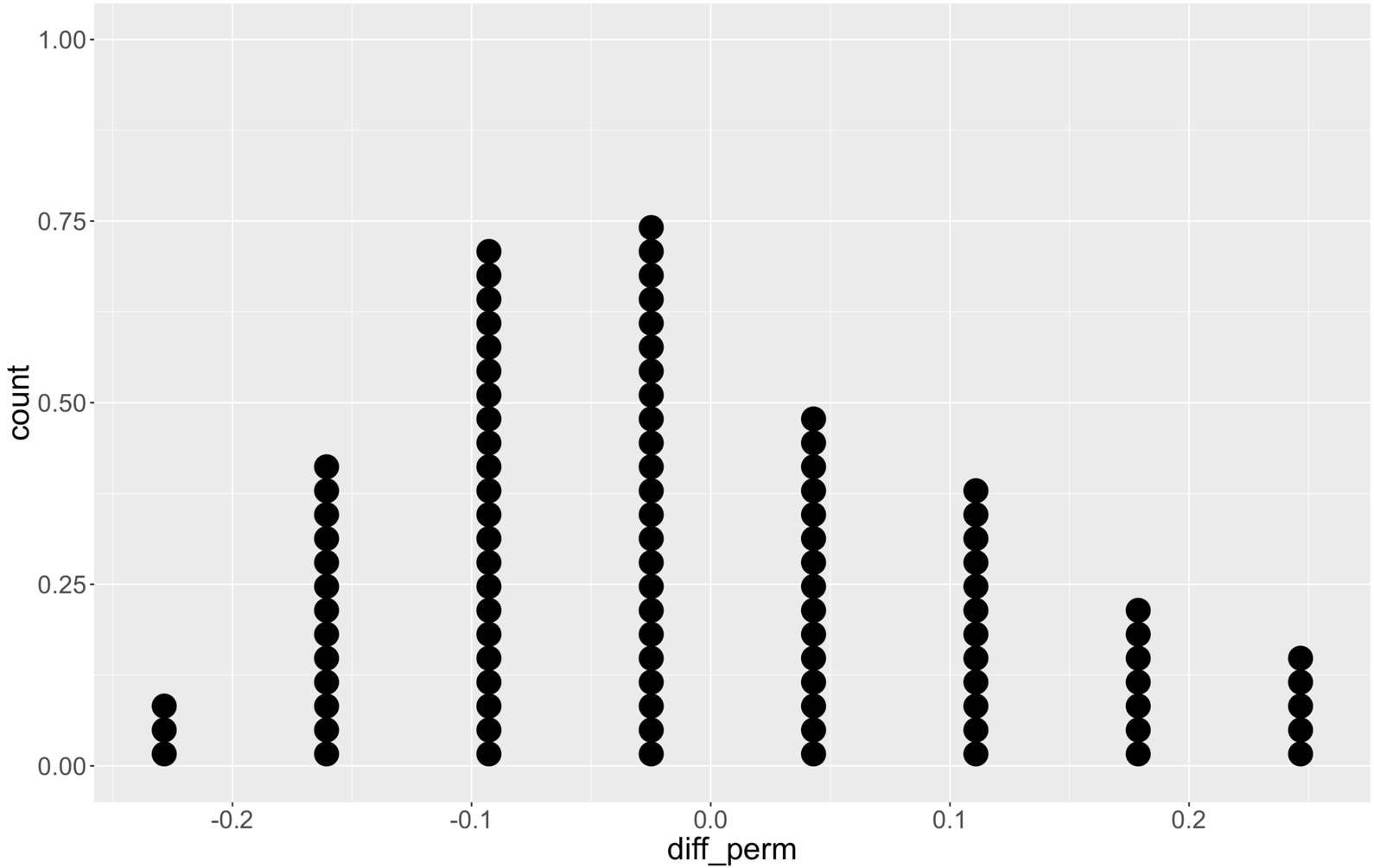
One random permutation

```
> soda %>%
    mutate(drink_perm = sample(drink)) %>%
    group_by(location) %>%
    summarize(prop_cola_perm = mean(drink_perm == "cola"),
              prop_cola = mean(drink == "cola")) %>%
    summarize(diff_perm = diff(prop_cola_perm),
              diff_orig = diff(prop_cola))
# A tibble: 1 × 2
    diff_perm   diff_orig
        <dbl>       <dbl>
1 -0.02488688 -0.09276018
```

Many random permutations

```
> soda %>%
    rep_sample_n(size = nrow(soda), reps = 5) %>%
    mutate(drink_perm = sample(drink)) %>%
    group_by(replicate, location) %>%
    summarize(prop_cola_perm = mean(drink_perm == "cola"),
              prop_cola = mean(drink == "cola")) %>%
    summarize(diff_perm = diff(prop_cola_perm),
              diff_orig = diff(prop_cola)) # West - East
# A tibble: 5 × 3
  replicate   diff_perm   diff_orig
      <int>       <dbl>       <dbl>
1         1 -0.09276018 -0.09276018
2         2 -0.09276018 -0.09276018
3         3  0.11085973 -0.09276018
4         4  0.17873303 -0.09276018
5         5  0.11085973 -0.09276018
```

Randomization distribution



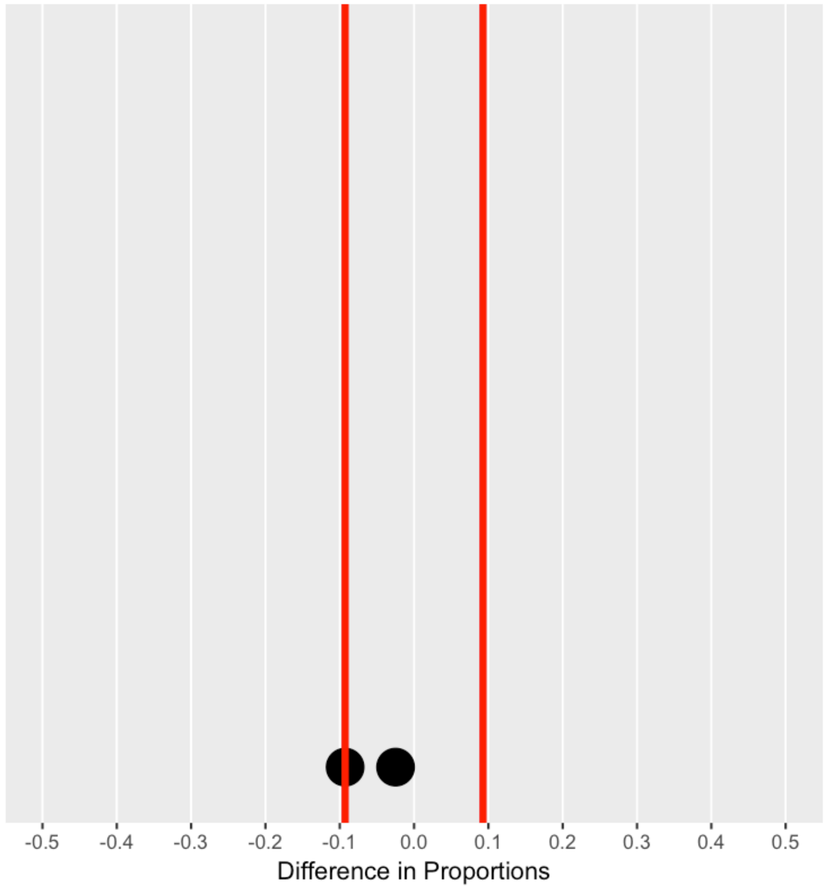
FOUNDATIONS OF INFERENCE

Let’s practice!

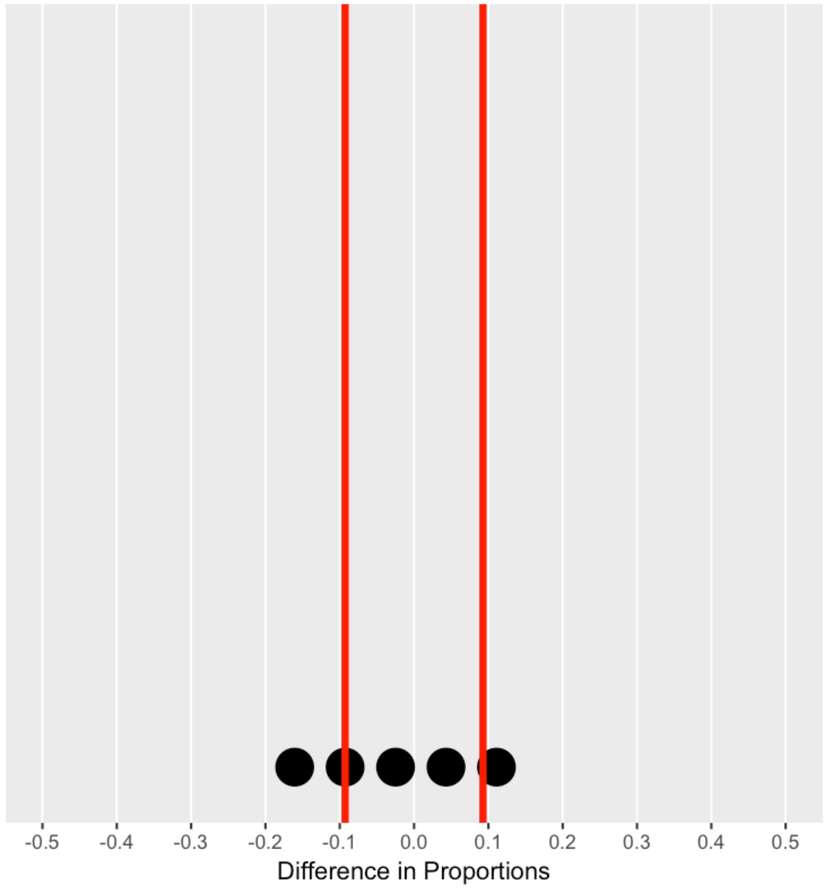
FOUNDATIONS OF INFERENCE

Using the randomization distribution

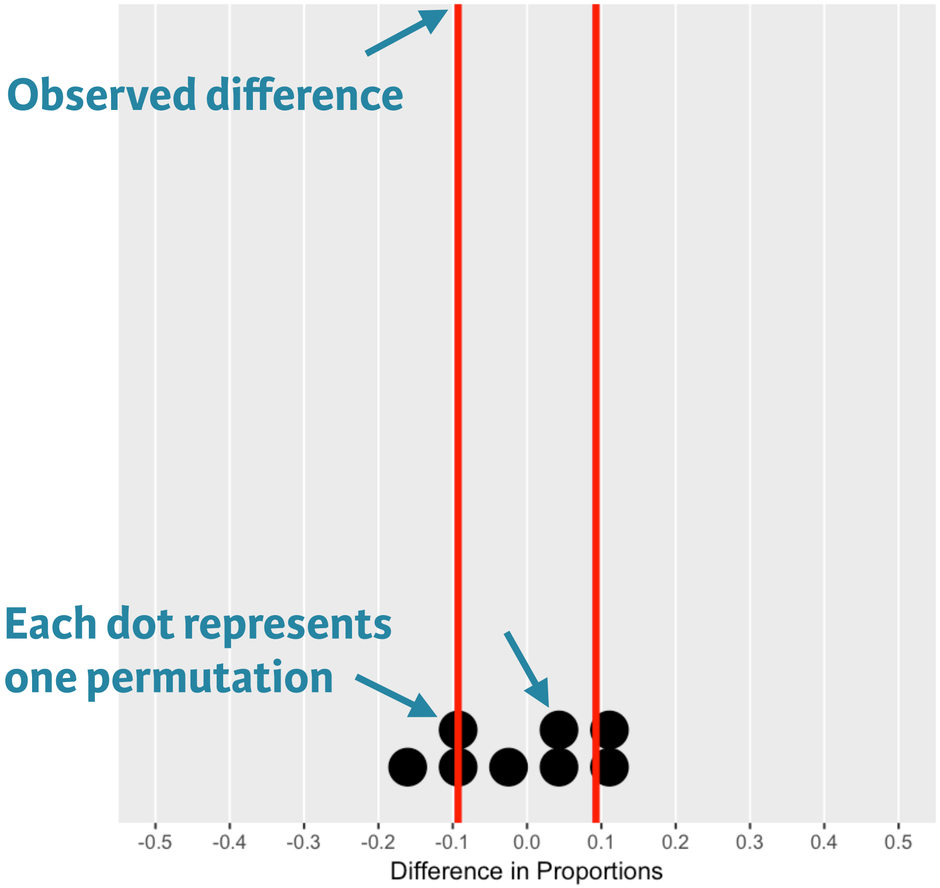
Understanding the null distribution



Understanding the null distribution



Understanding the null distribution

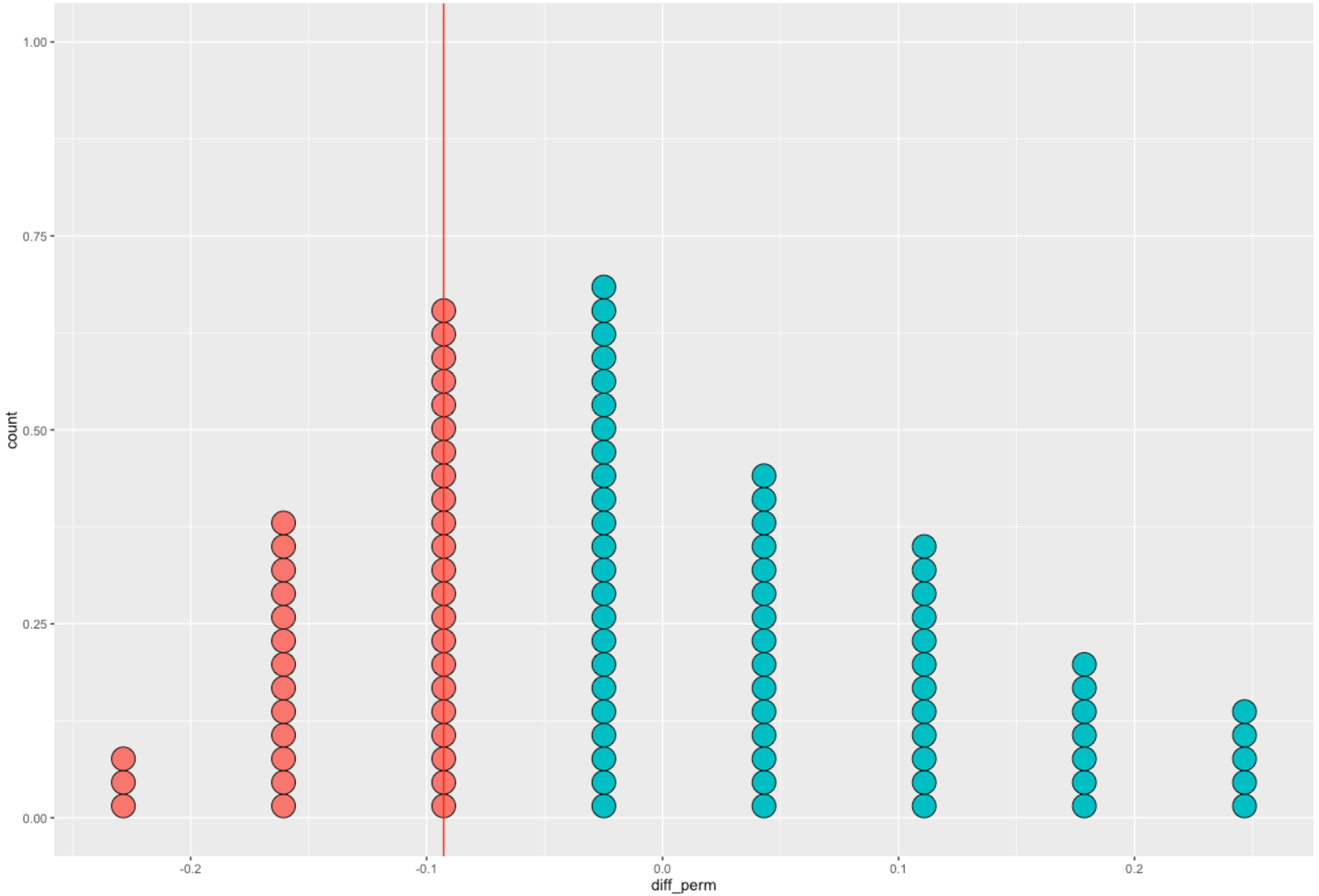


Data consistent with null?

```
> 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?

```
> set.seed(470)
> soda_perm <- soda %>%
    rep_sample_n(size = nrow(soda), reps = 100) %>%
    mutate(drink_perm = sample(drink)) %>%
    group_by(replicate, location) %>%
    summarize(prop_cola_perm = mean(drink_perm == "cola"),
              prop_cola = mean(drink == "cola")) %>%
    summarize(diff_perm = diff(prop_cola_perm),
              diff_orig = diff(prop_cola)) # West - East

> soda_perm %>%
    summarize(count = sum(diff_orig >= diff_perm),
              proportion = mean(diff_orig >= diff_perm))
# A tibble: 1 × 2
  count proportion
  <int>      <dbl>
1    38       0.38
```

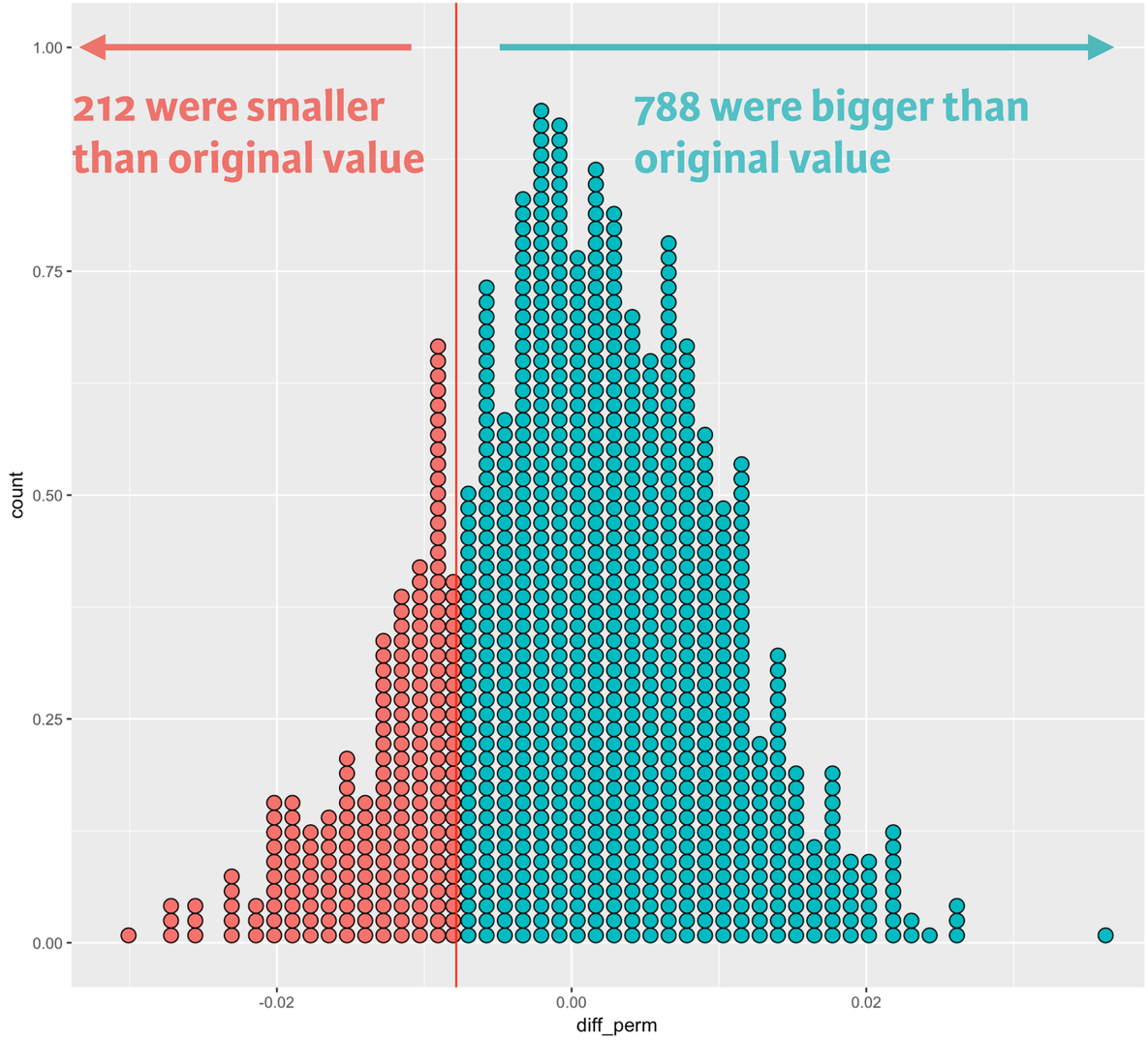
FOUNDATIONS OF INFERENCE

Let’s practice!

FOUNDATIONS OF INFERENCE

Study conclusions

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

FOUNDATIONS OF INFERENCE

Let’s practice!