# Randomization distributions & p-values

**Stat 120**

**Day 12**

# Statistical Hypotheses

***Null Hypothesis ($H_0$):*** Claim that there is no effect or difference.

***Alternative Hypothesis ($H_a$):*** Claim for which we seek evidence.

- Always claims about population parameters.

# Statistical Significance

When results as extreme as the observed sample statistic are *unlikely* to occur by random chance alone (assuming the null hypothesis is true), we say the sample results are ***statistically significant***

- If our sample is **statistically significant**, we have convincing evidence against $H_0$, in **favor of $H_a$**
- If our sample is **not statistically significant**, our test is **inconclusive. The null hypothesis may be true (or maybe not).**

## Key Question

***How unusual is it to see a sample statistic as extreme as that observed, if $H_0$ is true?***

# Extrasensory Perception (Example 1)

*p* = Proportion of correct guesses

$H_0$: $\boldsymbol{p = 1/5}$

$H_a$: $\boldsymbol{p > 1/5}$

- Suppose we try this n=10 times and get 3 correct guesses.
- What kinds of statistics (sample proportions) would we observe just by chance, if the null were true and ESP does not exist?
- How can we generate this distribution?

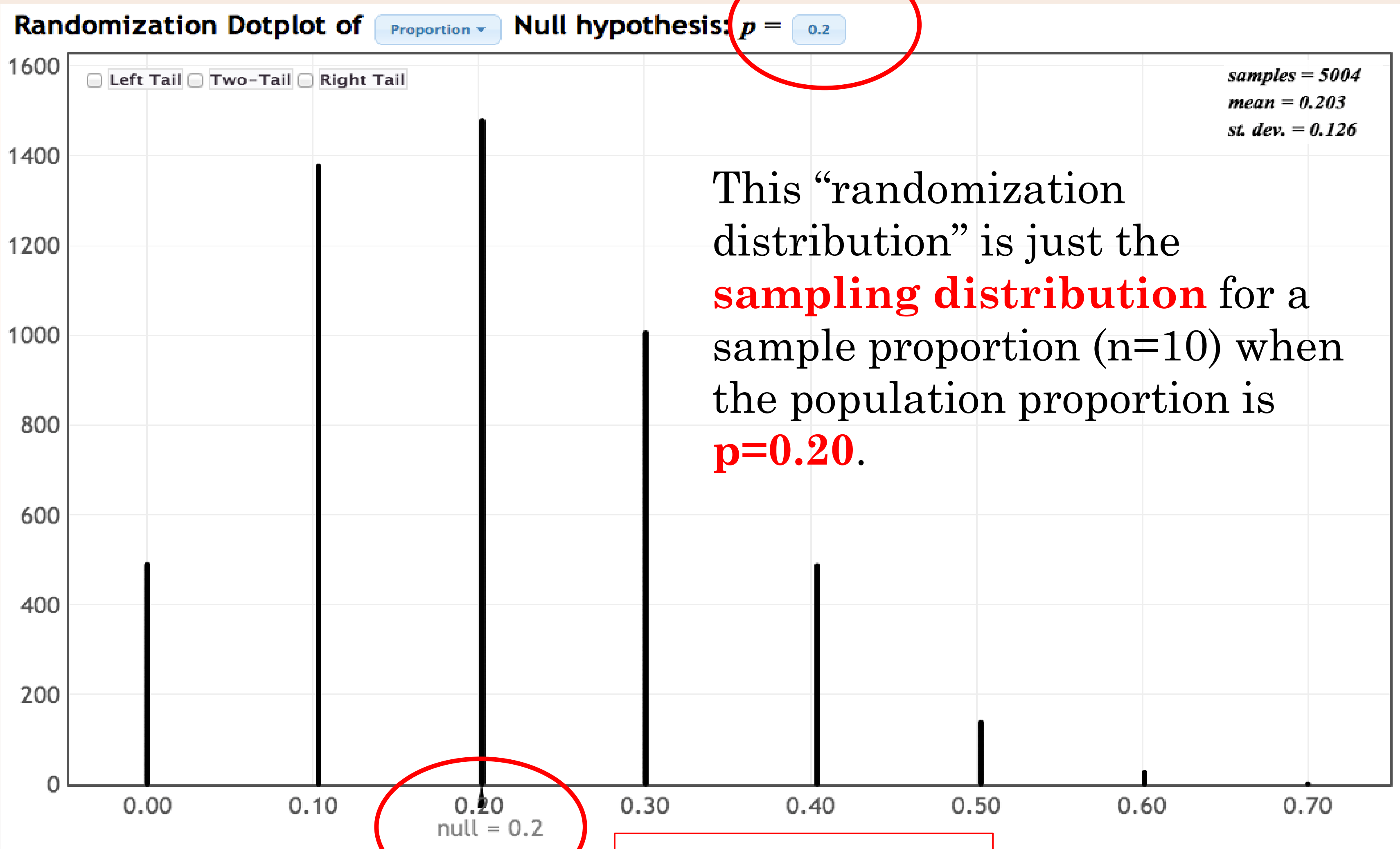
**Simulate many samples of size n=10 with p=0.2 and look at the distribution of sample proportions.**

# Randomization Distribution

A ***randomization distribution*** is a collection of statistics from samples simulated assuming the **null hypothesis is true**

- Also known as a **permutation distribution**.
- A randomization distribution is **centered** at the value of the **parameter given in the null hypothesis**.

# Randomization Distribution for ESP




This "randomization distribution" is just the **sampling distribution** for a sample proportion (n=10) when the population proportion is **p=0.20**.

$H_0$: $p = 1/5$

## Key Question

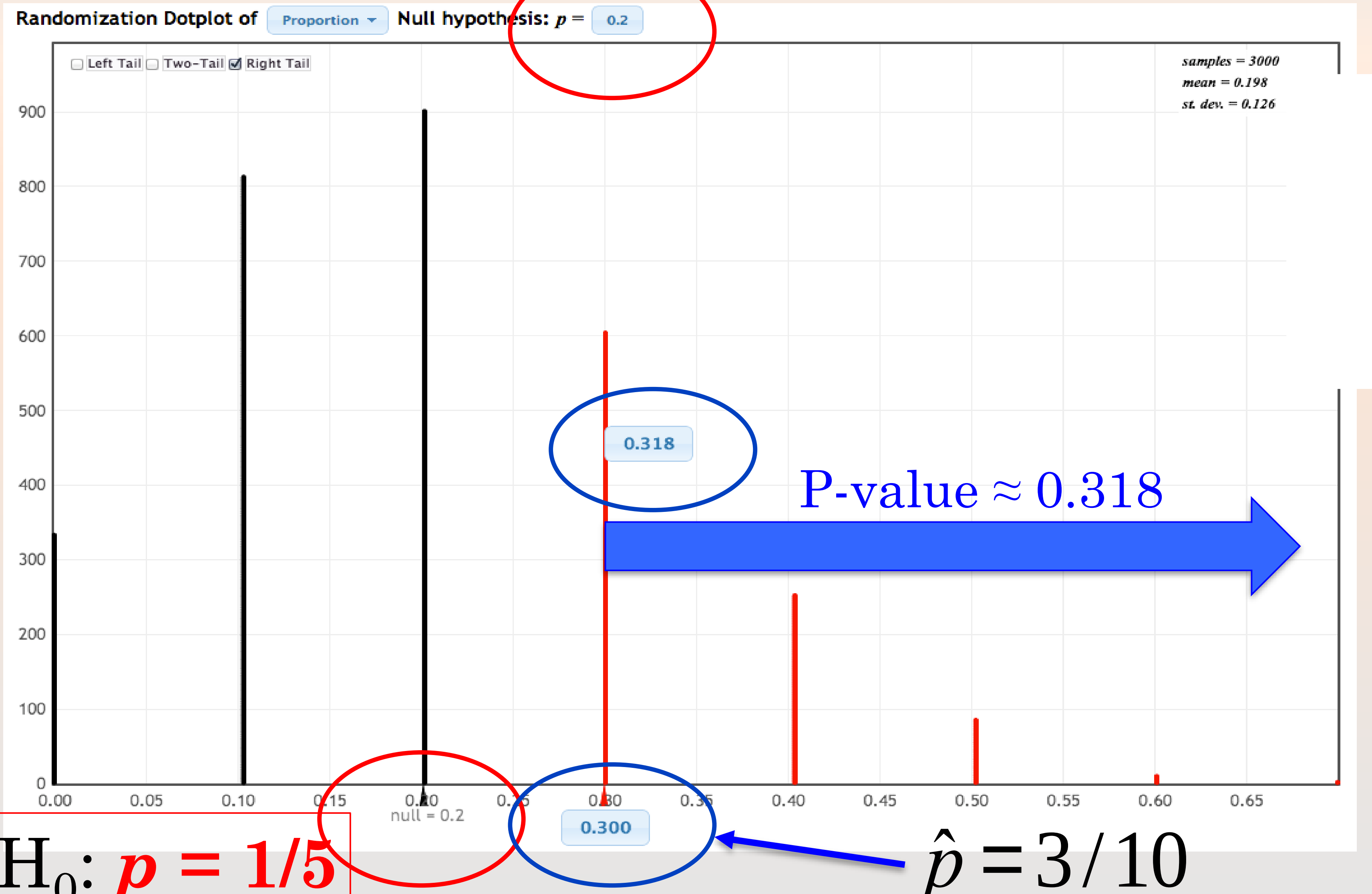
***How unusual is it to see a sample statistic as extreme as that observed, if $H_0$ is true?***

# P-VALUE

The ***p-value*** is the chance of obtaining a sample statistic as extreme (or more extreme) than the observed sample statistic, if the null hypothesis is true

- The p-value can be calculated as the proportion of statistics in a randomization distribution that are **as extreme (or more extreme) than the observed sample statistic**
- "**extreme**" is determined by the alternative hypothesis

# Randomization Distribution for ESP




$H_0$: $\boldsymbol{p = 1/5}$

# P-VALUE FOR ESP (EXAMPLE 1)

- The ***p-value*** is the chance of getting **at least 3 out of 10 guesses correct**, if $p = 0.2$.
  - P-value is about 0.318.
  - About 31% of the time we would get at least 3 out 10 guesses correct just by chance (no ESP). (interpretation)
  - Which conclusion does this p-value support?
    - A. Inconclusive, little evidence that supports ESP ($H_a$)
    - B. Borderline, weak evidence for ESP ($H_a$)
    - C. Strong statistically significant evidence for ESP ($H_a$)

# P-VALUE AND $H_0$

- If the **p-value is small**, then a statistic as extreme as that observed would be unlikely if the null hypothesis were true, providing evidence against $H_0$ and in **favor of the alternative**
- **Small p-value**
  - Results are statistically significant
  - Reject the null in favor of the alternative
- **Large p-value**
  - Results are not statistically significant
  - Do not reject the null in favor of the alternative

# P-VALUE (EXAMPLE 2)

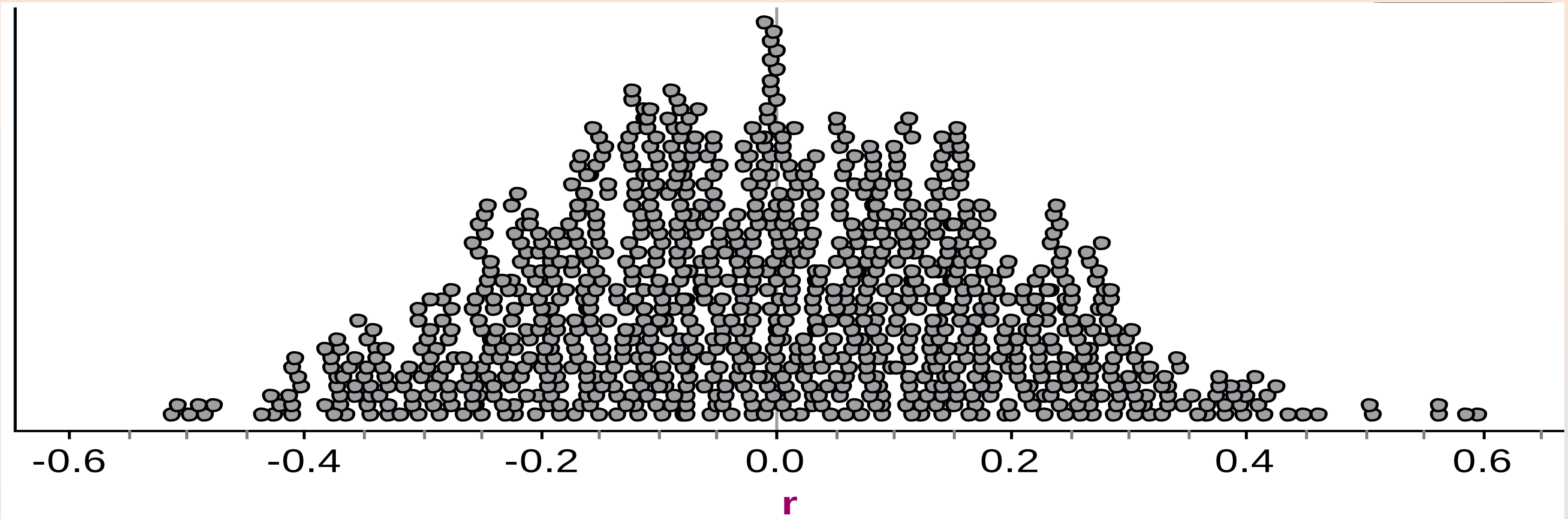
Using the randomization distribution below to test

$$H_0 : \rho = 0 \quad \text{vs} \quad H_a : \rho > 0$$

Match the sample correlation and p-values:

Sample Correlation: r = 0.1, r = 0.3, or r = 0.5

P-values: 0.005 , 0.15, or 0.35



# Sleep versus Caffeine (Example 3)

- Recall the sleep versus caffeine experiment
- $\mu_s$ and $\mu_c$ are the true mean number of words recalled after sleeping and after caffeine.
  - $H_0: \mu_s = \mu_c$
  - $H_a: \mu_s \neq \mu_c$

$$\boxed{\begin{array}{l} H_0: \mu_s - \mu_c = 0 \\ H_a: \mu_s - \mu_c \neq 0 \end{array}}$$

- How can we create a randomization distribution consistent with the null?
  - What statistic do we compute?
  - Sample difference: $\bar{x}_S - \bar{x}_C$
  - Where is the distribution centered?
  - Distribution centered at a difference of 0 (null)

# Sleep versus Caffeine Data

| Words | Group |
|---|---|
| 9 | sleep |
| 11 | sleep |
| 13 | sleep |
| 14 | sleep |
| 14 | sleep |
| 15 | sleep |
| 16 | sleep |
| 17 | sleep |
| 17 | sleep |
| 18 | sleep |
| 18 | sleep |
| 21 | sleep |

$\bar{x}_S = 15.25$

| Words | Group |
|---|---|
| 6 | caffeine |
| 7 | caffeine |
| 10 | caffeine |
| 10 | caffeine |
| 12 | caffeine |
| 12 | caffeine |
| 13 | caffeine |
| 14 | caffeine |
| 14 | caffeine |
| 15 | caffeine |
| 16 | caffeine |
| 18 | caffeine |

$\bar{x}_C = 12.25$

$$\bar{x}_S - \bar{x}_C = 3$$

*What kinds of results would you see, just by random chance, **if sleep or caffeine were equivalent for memory**?*

***Rerandomize sleep/caffeine**, but do not change the number of words recalled.*

# Sleep versus Caffeine – one **rerandomized** data set (under $H_0$)

| Words | Group |
|---|---|
| 9 | sleep |
| 11 | caffeine |
| 13 | caffeine |
| 14 | sleep |
| 14 | sleep |
| 15 | caffeine |
| 16 | sleep |
| 17 | caffeine |
| 17 | sleep |
| 18 | sleep |
| 18 | caffeine |
| 21 | sleep |

| Words | Group |
|---|---|
| 6 | caffeine |
| 7 | sleep |
| 10 | sleep |
| 10 | caffeine |
| 12 | caffeine |
| 12 | caffeine |
| 13 | caffeine |
| 14 | caffeine |
| 14 | sleep |
| 15 | sleep |
| 16 | sleep |
| 18 | caffeine |

$\bar{x}_S = 14.25$ $\bar{x}_C = 13.25$

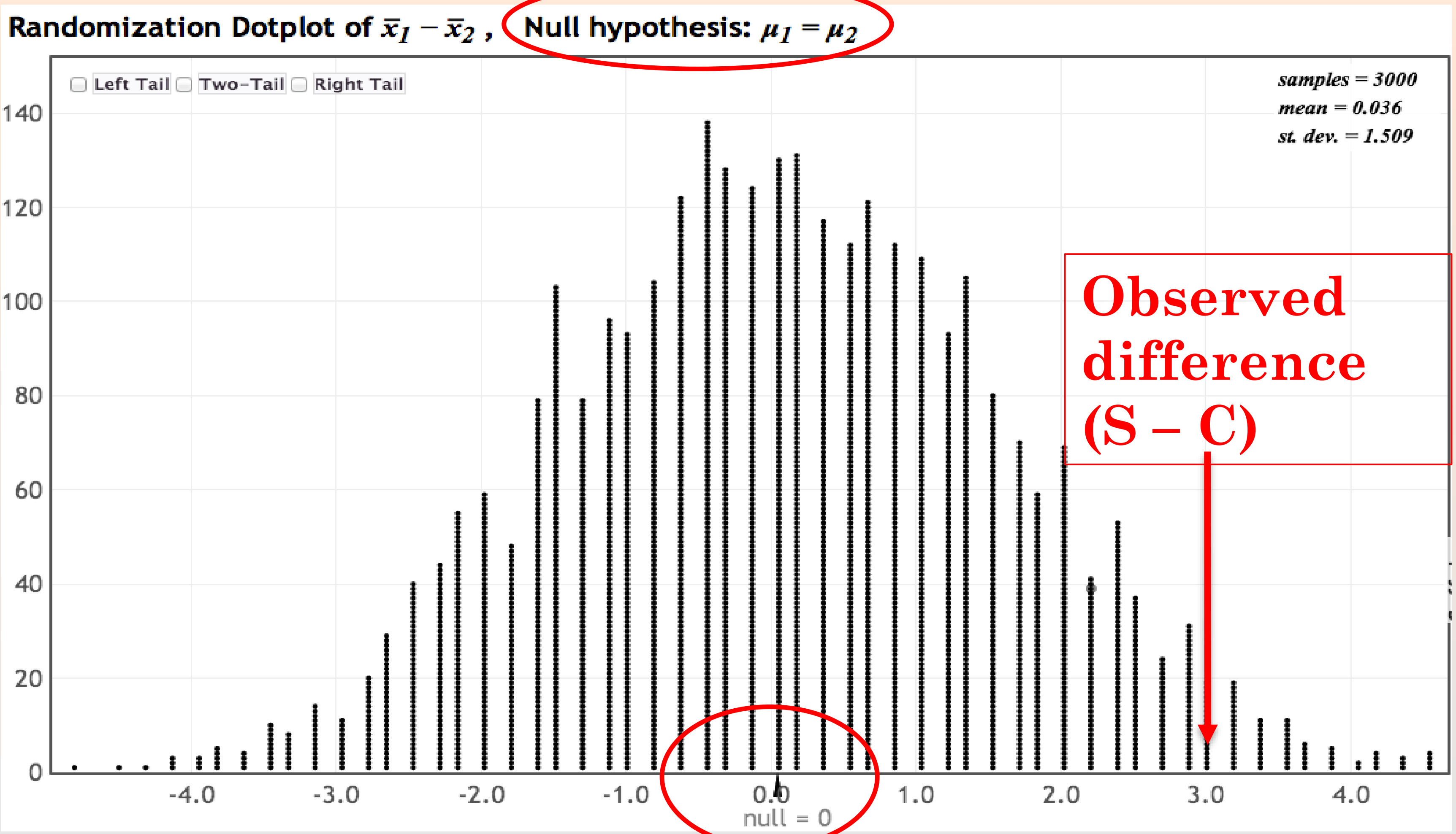
$\bar{x}_S - \bar{x}_C = 1$

*What kinds of results would you see, just by random chance, **if sleep or caffeine were equivalent for memory**?*

***Rerandomize sleep/caffeine**, but do not change the number of words recalled.*

# Sleep vs. Caffeine: Randomization Distribution

- Rerandomize many, many times.
- Compute difference in means for each rerandomized sample.

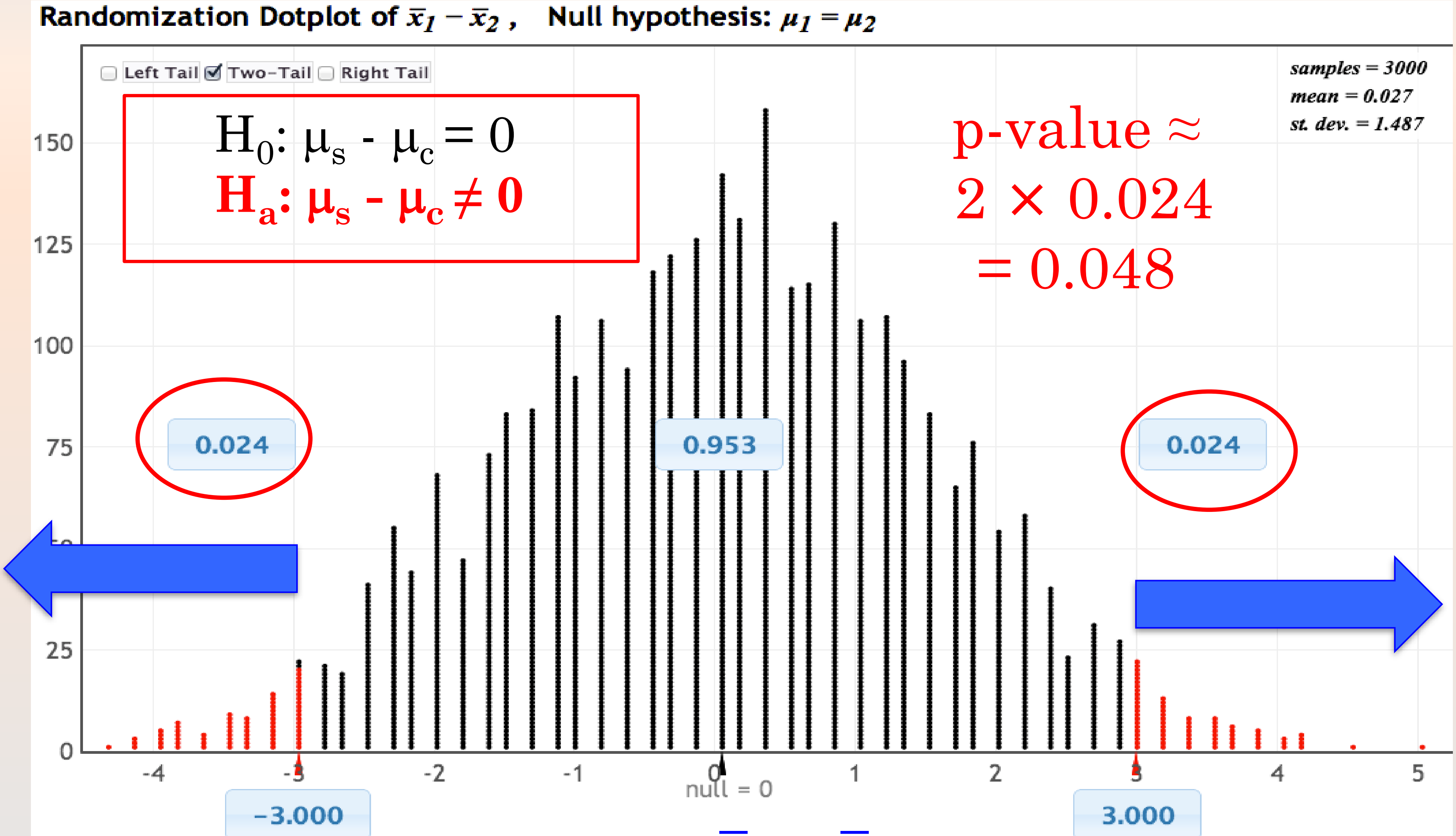


# Sleep versus Caffeine

$H_0$: $\mu_s$ - $\mu_c = 0$
$H_a$: $\mu_s$ - $\mu_c \neq 0$

- The observed difference is 3 words.
- The p-value is the proportion of samples that yield a **difference in means of 3 or more words** (under randomization model).
  - Two-sided alternative: no direction specified!

# Sleep versus Caffeine




$$\bar{X}_S - \bar{X}_C = 3$$

# Sleep versus Caffeine (Example 3)

$$H_0: \mu_s - \mu_c = 0$$
$$H_a: \mu_s - \mu_c \neq 0$$

- P-value is about 0.048
- About 4.8% of samples will yield a difference in means of 3 or more words if sleep and caffeine have the same influence on memory.
- Which hypothesis does this p-value support?
  A. Inconclusive, little evidence that suggests treatments differ
  B. Borderline, weak evidence that suggests treatments differ
  C. Strong statistically significant evidence that suggests treatments differ

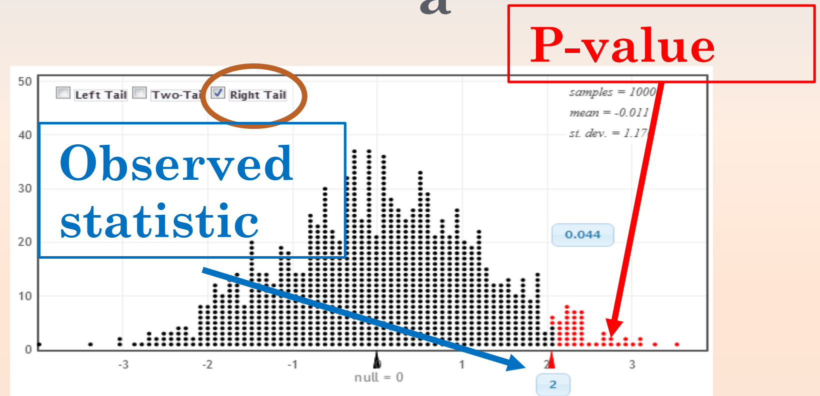
# Alternative Hypothesis

- The p-value is the proportion in the tail in the direction specified by $H_a$

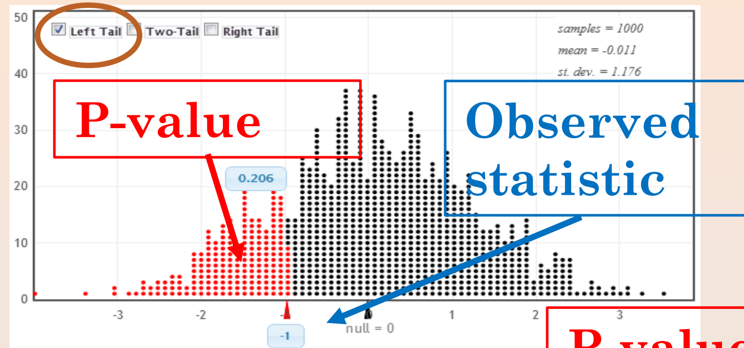
- For a two-sided alternative, the p-value is twice the proportion in the smallest tail

# Summary: p-value and $H_a$

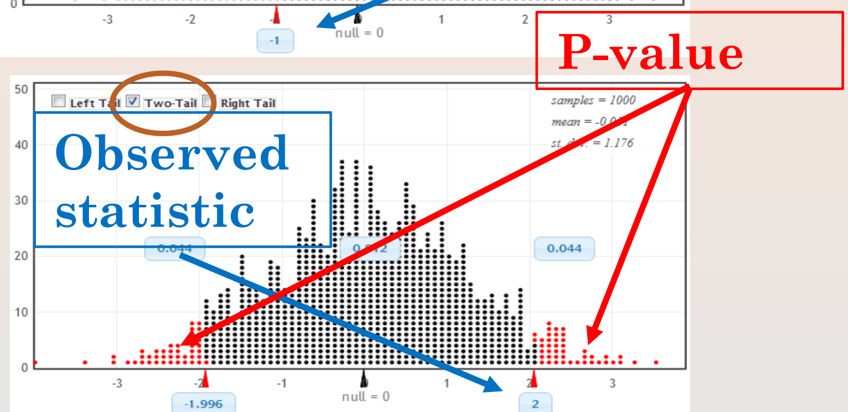
Upper-tail
(Right Tail)
$H_a$: parameter > null value




Lower-tail
(Left Tail)
$H_a$: parameter < null value




Two-tailed
$H_a$: parameter ≠ null value



# SUMMARY: RANDOMIZATION DISTRIBUTION FOR ONE PROPORTION

- Null: $H_0$: $p = p_0$ where $p_0$ is the null value of the population parameter $p$
- Creating a randomization distribution consistent with $H_0$:
  - Generate a sample of size n from a population with proportion $p_0$
  - Compute the sample proportion
  - Repeat lots of times

# SUMMARY: RANDOMIZATION DISTRIBUTION FOR COMPARING TWO GROUPS 1 AND 2

- Null: $H_0$: $\mu_1 - \mu_2 = 0$ OR $H_0$: $p_1 - p_2 = 0$
- Creating a randomization distribution consistent with $H_0$: **group membership arbitrary (no affect on response)**
  - Randomly permute (re-randomize) the group assignment for all cases
  - Compute the sample mean/proportion for each group and find the difference $\bar{x}_1 - \bar{x}_2$ OR $\hat{p}_1 - \hat{p}_2$ and repeat lots of times

Original data

| Case | response | Group |
|---|---|---|
| 1 | $x_1$ | 1 |
| 2 | $x_2$ | 1 |
| 3 | $x_3$ | 1 |
| 4 | $x_4$ | 2 |
| 5 | $x_5$ | 2 |
| 6 | $x_6$ | 2 |
| 7 | $x_7$ | 2 |

**Permute groups**

| Case | response | **Group** |
|---|---|---|
| 1 | $x_1$ | **2** |
| 2 | $x_2$ | **2** |
| 3 | $x_3$ | **1** |
| 4 | $x_4$ | **2** |
| 5 | $x_5$ | **2** |
| 6 | $x_6$ | **1** |
| 7 | $x_7$ | **1** |

# SUMMARY: RANDOMIZATION DISTRIBUTION FOR COMPARING TWO GROUPS 1 AND 2

- Comment:
  - Equivalently, we can permute (re-randomized) the response for all cases but leave the group assignments fixed.
  - Will get the same randomization distribution for the difference in means or proportions either way.

Original data

| Case | response | Group |
|---|---|---|
| 1 | $x_1$ | 1 |
| 2 | $x_2$ | 1 |
| 3 | $x_3$ | 1 |
| 4 | $x_4$ | 2 |
| 5 | $x_5$ | 2 |
| 6 | $x_6$ | 2 |
| 7 | $x_7$ | 2 |

Permute responses →

| Case | response | Group |
|---|---|---|
| 1 | $x_6$ | 1 |
| 2 | $x_7$ | 1 |
| 3 | $x_3$ | 1 |
| 4 | $x_5$ | 2 |
| 5 | $x_4$ | 2 |
| 6 | $x_1$ | 2 |
| 7 | $x_2$ | 2 |

# SUMMARY: RANDOMIZATION DISTRIBUTION FOR CORRELATION OR SLOPE

- Null: $H_0: \rho = 0$ OR $H_0: \beta = 0$
- Creating a randomization distribution consistent with $H_0$: **no association between x and y**
  - Randomly permute (re-randomize) one of the variables (either or x or y)
  - Compute the sample correlation/slope r or b.
  - Repeat lots of times

Original data

| Case | x variable | y variable |
|---|---|---|
| 1 | $x_1$ | $y_1$ |
| 2 | $x_2$ | $y_2$ |
| 3 | $x_3$ | $y_3$ |
| 4 | $x_4$ | $y_4$ |
| 5 | $x_5$ | $y_5$ |
| 6 | $x_6$ | $y_6$ |
| 7 | $x_7$ | $y_7$ |

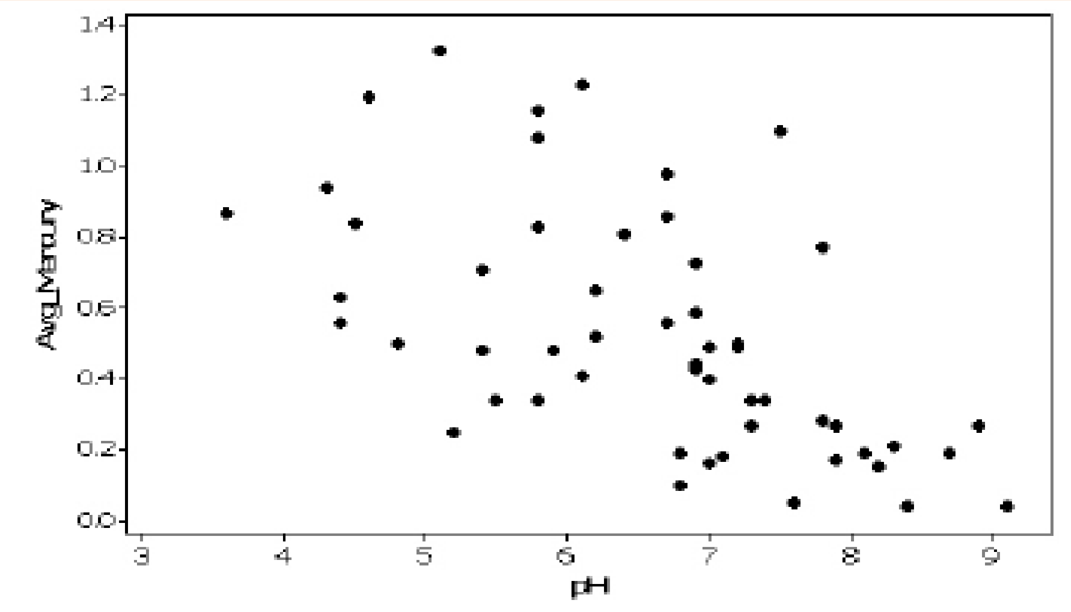
**Permute y variable** →

| Case | x variable | y variable |
|---|---|---|
| 1 | $x_1$ | $y_5$ |
| 2 | $x_2$ | $y_2$ |
| 3 | $x_3$ | $y_4$ |
| 4 | $x_4$ | $y_1$ |
| 5 | $x_5$ | $y_6$ |
| 6 | $x_6$ | $y_5$ |
| 7 | $x_7$ | $y_3$ |

# Mercury and pH in Lakes

- For Florida lakes, are lower pH levels (more acidity) associated with higher mercury levels?

$H_0: \beta = 0$ vs. $H_a: \beta < 0$

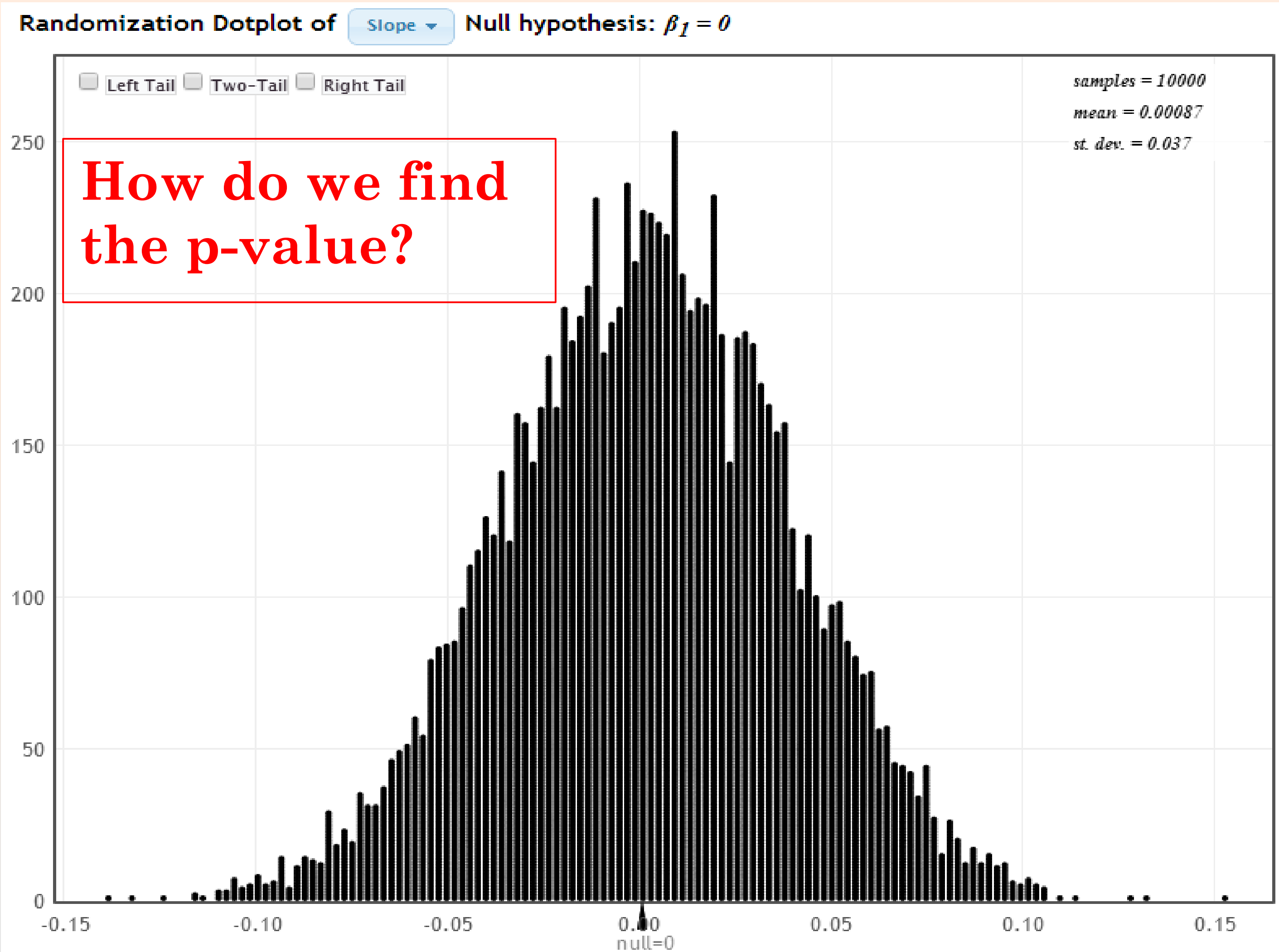



The regression line slope is b = -0.152.

Lange, Royals, and Connor, Transactions of the American Fisheries Society (1993)

# Mercury and pH in Lakes

$H_0: \beta = 0 \quad$ vs. $\quad H_a: \beta < 0$

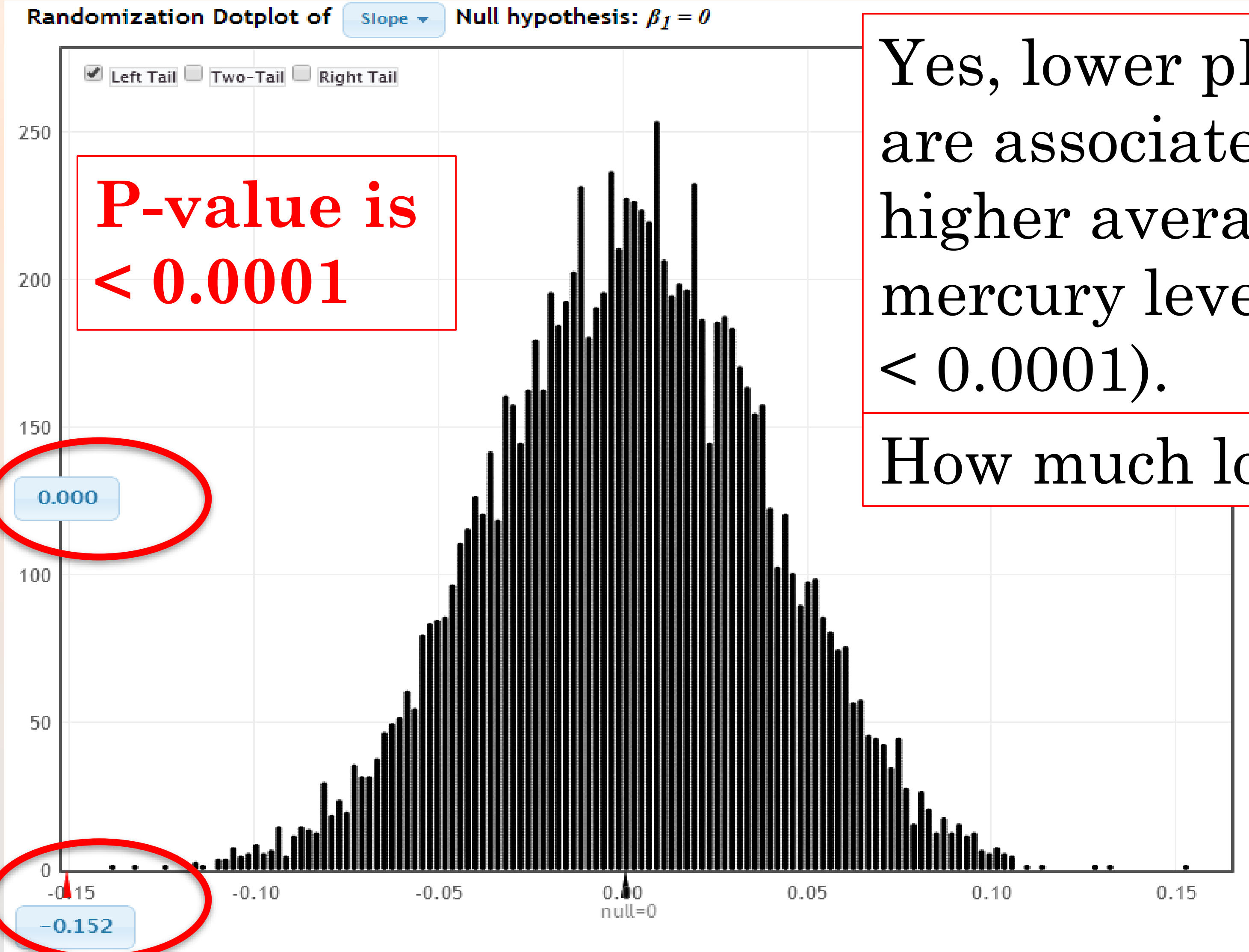



**How do we find the p-value?**

**Chance of getting a slope as small, or smaller than, the observed slope of b = -0.152.**

# Mercury and pH in Lakes

$H_0: \beta = 0$ vs. $H_a: \beta < 0$



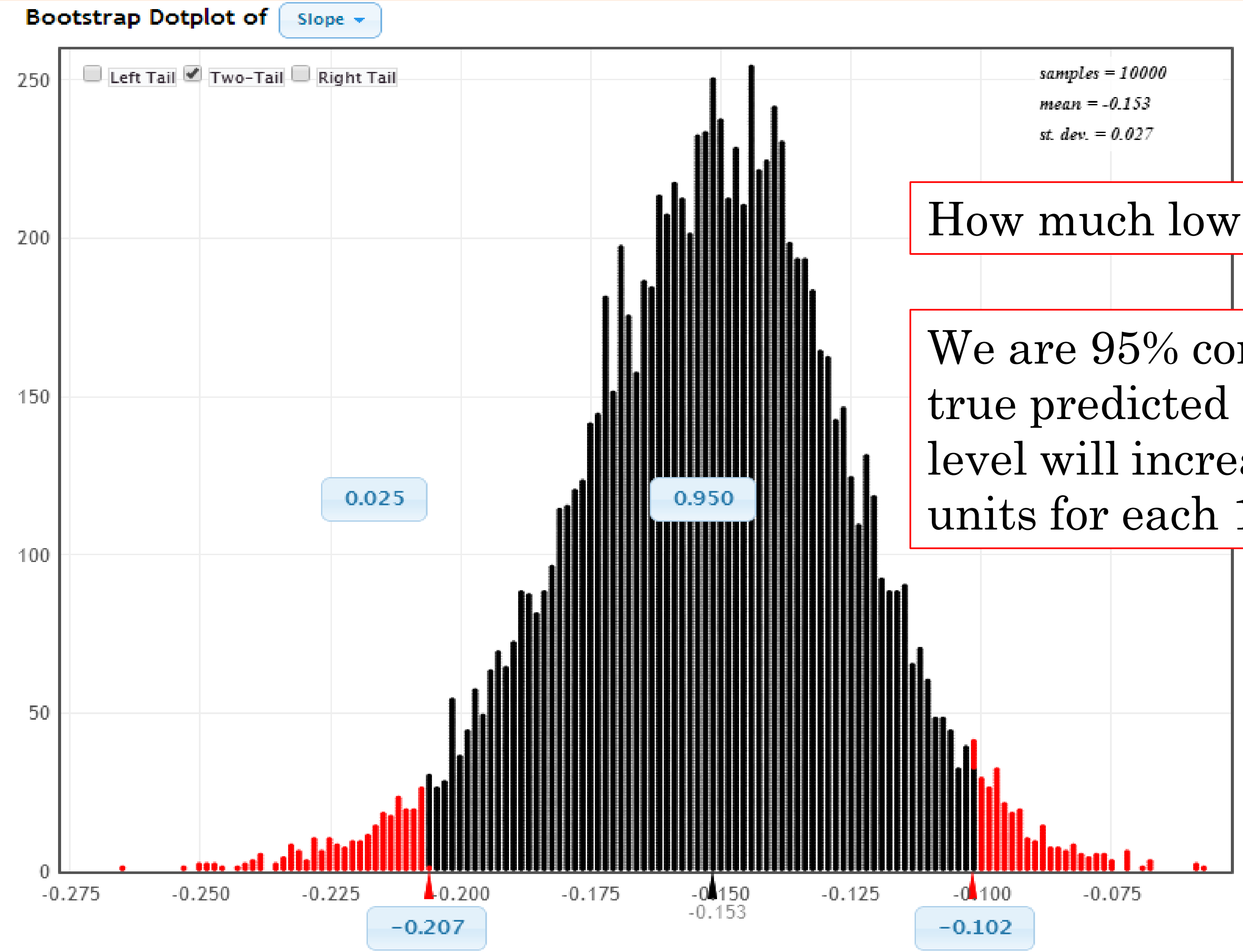

**P-value is < 0.0001**

Yes, lower pH levels are associated with higher average mercury levels (p-value < 0.0001).

How much lower?

# Mercury and pH in Lakes

Bootstrap distribution for the slope

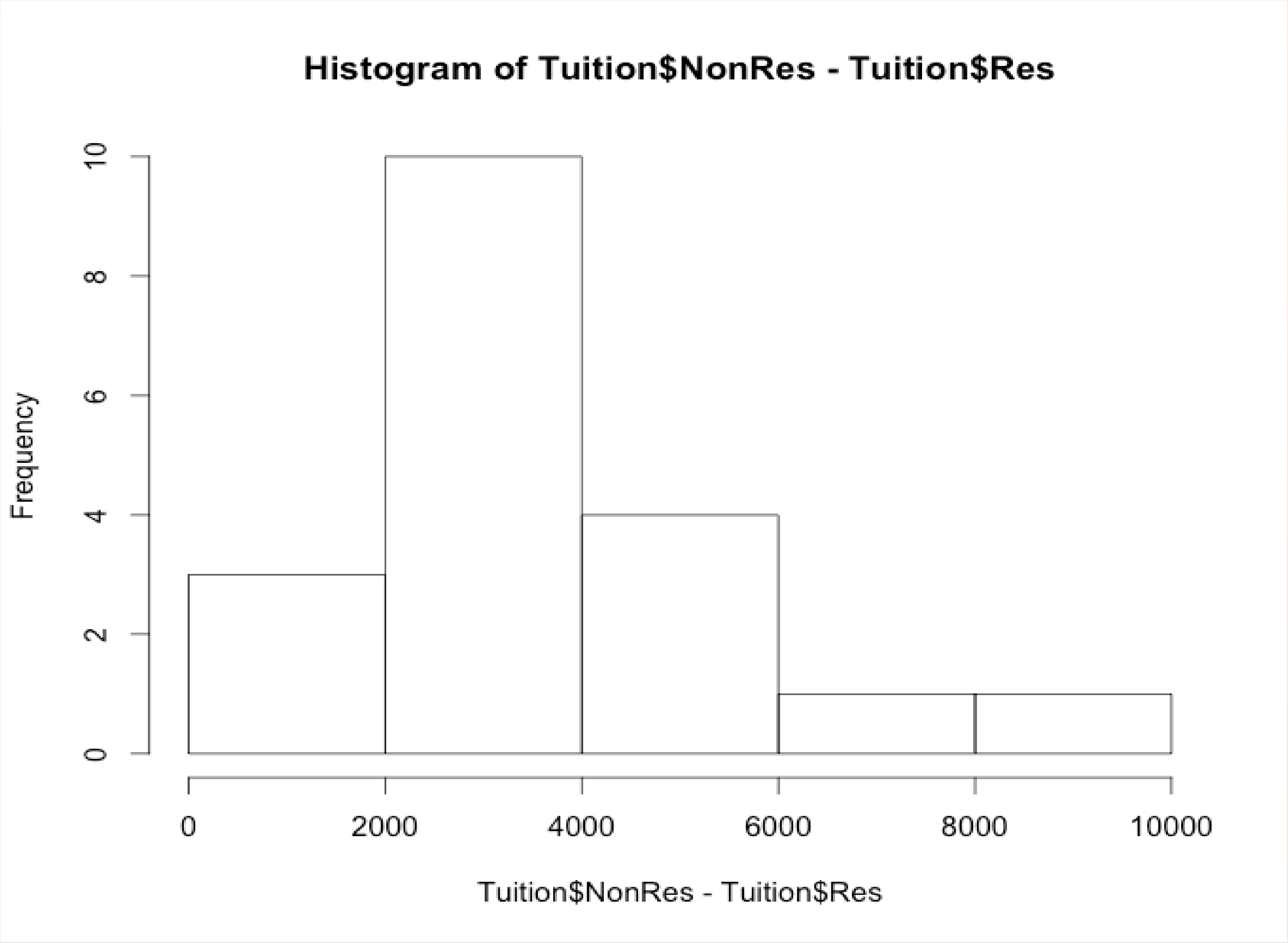



How much lower?

We are 95% confident that the true predicted average mercury level will increase by 0.1 to 0.2 units for each 1 unit drop in pH.

# TUITION: RESIDENT VS. NON-RESIDENT

- Tuition2006 data from the lab manual section 4.5
- We want to know if the average tuition charged to non-residents is higher than residents for all state colleges and universities
- Population: all state colleges and universities
- Parameters: $\mu$ = mean tuition (resident or non-resident) for all colleges and universities
- $H_0$: $\mu_{non\text{-}resident} - \mu_{resident} = 0$
- $H_a$: $\mu_{non\text{-}resident} - \mu_{resident} > 0$
- Data: **paired** tuition amounts (resident, non-resident) from a random sample of n=19 schools


Histogram of Tuition$NonRes - Tuition$Res
Frequency
0
2
4
6
8
10
0
2000
4000
6000
8000
10000
Tuition$NonRes - Tuition$Res

# TUITION: RESIDENT VS. NON-RESIDENT

- $H_0$: $\mu_{non\text{-}resident} - \mu_{resident} = 0$
- $H_a$: $\mu_{non\text{-}resident} - \mu_{resident} > 0$
- How can we create a randomization distribution for paired data?

Original Data (first 7 cases)

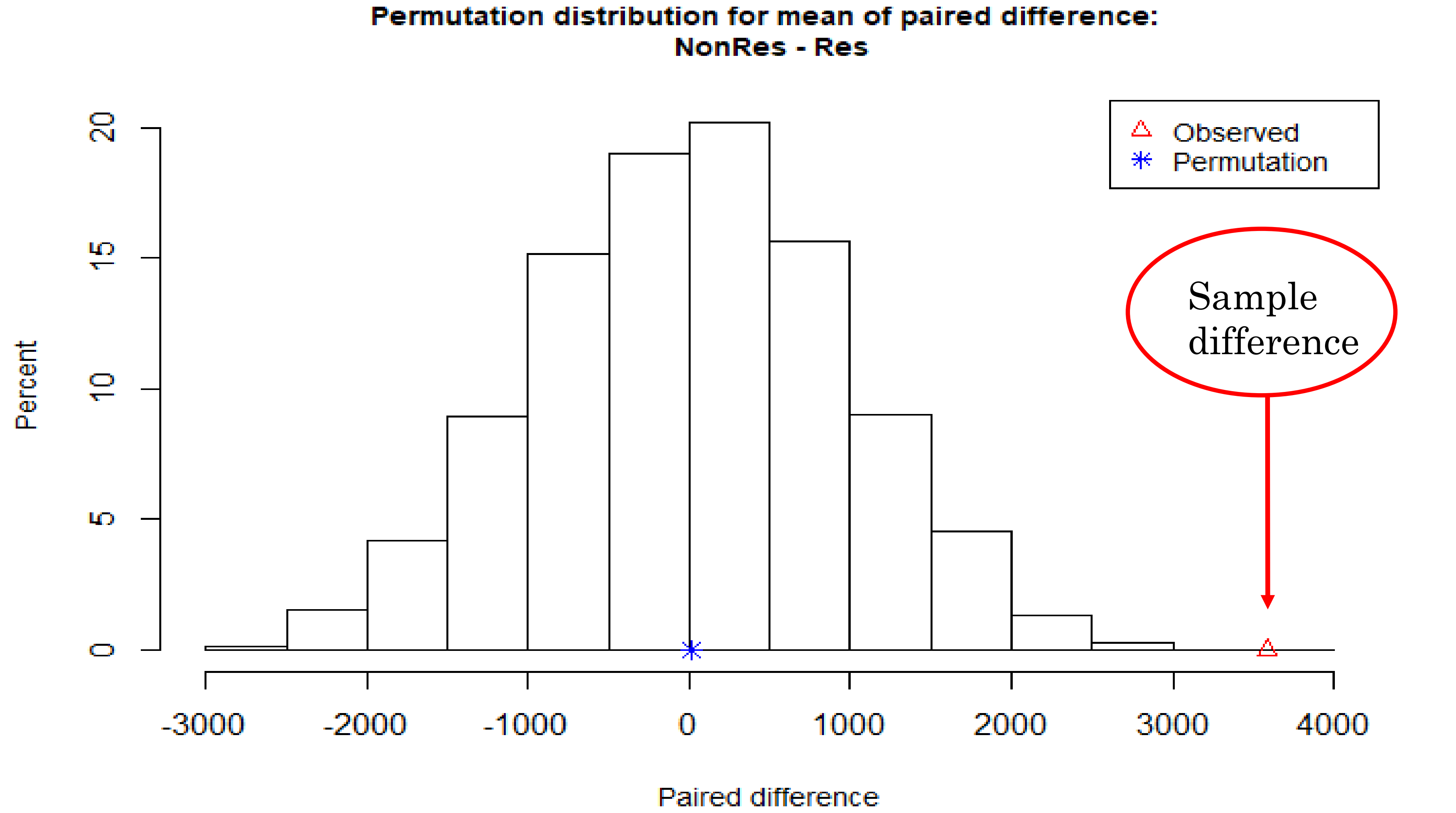
| Case | Non-resident tuition | Resident tuition |
|---|---|---|
| 1 | 8800 | 4200 |
| 2 | 3600 | 1900 |
| 3 | 8600 | 3400 |
| 4 | 7000 | 3200 |
| 5 | 12700 | 3400 |
| 6 | 5700 | 2600 |
| 7 | 5900 | 3300 |

**Randomly assign tuition amounts to resident or non-resident for each case**

| Case | Non-resident tuition | Resident tuition |
|---|---|---|
| 1 | **4200** | **8800** |
| 2 | **1900** | **3600** |
| 3 | **8600** | **3400** |
| 4 | **7000** | **3200** |
| 5 | **12700** | **3400** |
| 6 | **2600** | **5700** |
| 7 | **5900** | **3300** |

# Tuition: Resident vs. Non-Resident

- $H_0$: $\mu_{non\text{-}resident} - \mu_{resident} = 0$
- $H_a$: $\mu_{non\text{-}resident} - \mu_{resident} > 0$
- How can we create a randomization distribution for paired data?
  - For each case: Randomly re-assign tuition amounts to resident or non-resident
  - Compute the difference in tuition for non-residents and residents
  - Calculate the mean difference $\bar{x}_{difference}$
  - Repeat lots of times
- Use R to get this randomization distribution


Permutation distribution for mean of paired difference:
NonRes - Res
Observed
Permutation
Sample difference
Percent
0
5
10
15
20
-3000
-2000
-1000
0
1000
2000
3000
4000
Paired difference

# TUITION: RESIDENT VS. NON-RESIDENT

- $H_0$: $\mu_{non\text{-}resident} - \mu_{resident} = 0$
- $H_a$: $\mu_{non\text{-}resident} - \mu_{resident} > 0$

```
> permTestPaired(NonRes ~ Res, data= tuition, alt = "greater")

        ** Permutation test for mean of paired difference **

 Permutation test with alternative: greater
 Observed mean
  NonRes :  6405.263     Res :  2821.053
 Observed difference  NonRes - Res : 3584.211

 Mean of permutation distribution: 13.60926
 Standard error of permutation distribution: 948.5907
 P-value:  1e-04
```

- If there was no difference in mean tuition, we would see a mean difference (NR-R) of at least $3584 less than 0.01% of the time. We have very strong evidence that mean tuition for non-residents is higher than for residents.

# Formal Decisions

A formal hypothesis test has only two possible conclusions:

1. The p-value is small: reject the null hypothesis in favor of the alternative
2. The p-value is not small: do not reject the null hypothesis

***How small?***

# Significance Level

- The ***significance level***, $\alpha$, is the threshold below which the p-value is deemed small enough to reject the null hypothesis

$$\text{p-value} < \alpha \quad \Rightarrow \quad \text{Reject } H_0$$
$$\text{p-value} \geq \alpha \quad \Rightarrow \quad \text{Do not Reject } H_0$$

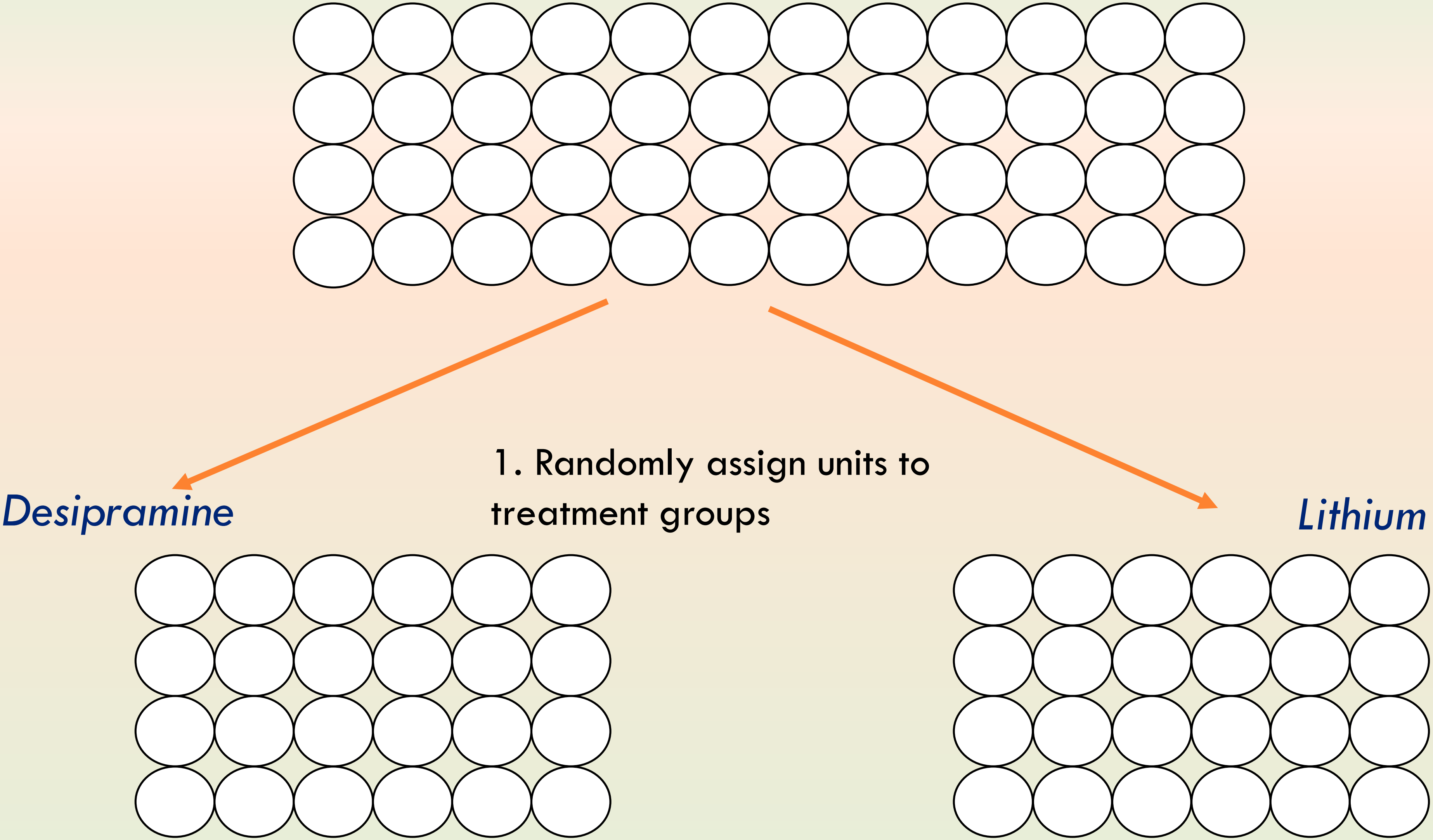
# Significance Level

- If the p-value is ***less than $\alpha$***, the results are ***statistically significant***, and we reject the null hypothesis in favor of the alternative
- If the p-value is ***not*** less than $\alpha$, the results are ***not*** statistically significant, and our test is inconclusive
- Often $\alpha = 0.05$ by default, unless otherwise specified

# Cocaine Addiction

- In a randomized experiment on treating cocaine addiction, 48 people were randomly assigned to take either Desipramine (a new drug), or Lithium (an existing drug), and then followed to see who relapsed
- Question of interest:

- We are testing to see if desipramine is better than lithium at treating cocaine addiction.

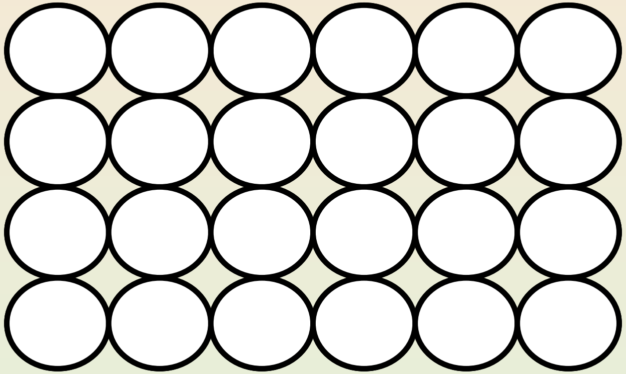

1. Randomly assign units to treatment groups
Desipramine
Lithium

2. Conduct experiment
3. Observe relapse counts in each group

R = Relapse
N = No Relapse

1. Randomly assign units to treatment groups

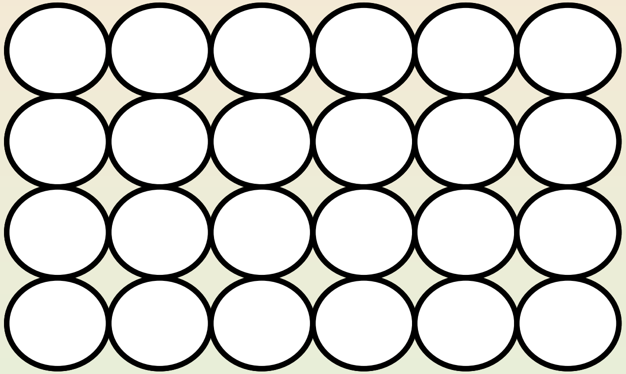
*Desipramine*



10 relapse, 14 no relapse

$$\hat{p}_D - \hat{p}_L$$
$$= \frac{10}{24} - \frac{18}{24}$$
$$= -.333$$

*Lithium*



18 relapse, 6 no relapse

# SUMMARY

- The randomization distribution shows what types of statistics would be observed, just by random chance, if the null hypothesis were true
- A p-value is the chance of getting a statistic as extreme as that observed, if $H_0$ is true
- A p-value can be calculated as the proportion of statistics in the randomization distribution as extreme as the observed sample statistic
- The smaller the p-value, the greater the evidence against $H_0$