

Hypothesis tests for two means

Overview

Review/continuation of hypothesis testing for a single proportion

One tailed vs. two-tailed hypothesis tests

If there is time: Hypothesis tests for two means

Announcement

Homework 6 has been posted!

It is due on Gradescope on **Sunday March 1st at 11pm**

- **Be sure to mark each question on Gradescope!**

Announcement: Midterm exam

Exam is during regular class time on Thursday
March 5th

- Exam is on paper

If you have accommodations, please schedule
the exam with SAS

A practice exam (last year's exam) has been
posted



Midterm exam “cheat sheet”

You are allowed an exam “cheat sheet”

One page, double sided, that contains only code and equations

- No code comments allowed

Cheat sheet must be on a regular 8.5 x 11 piece of paper

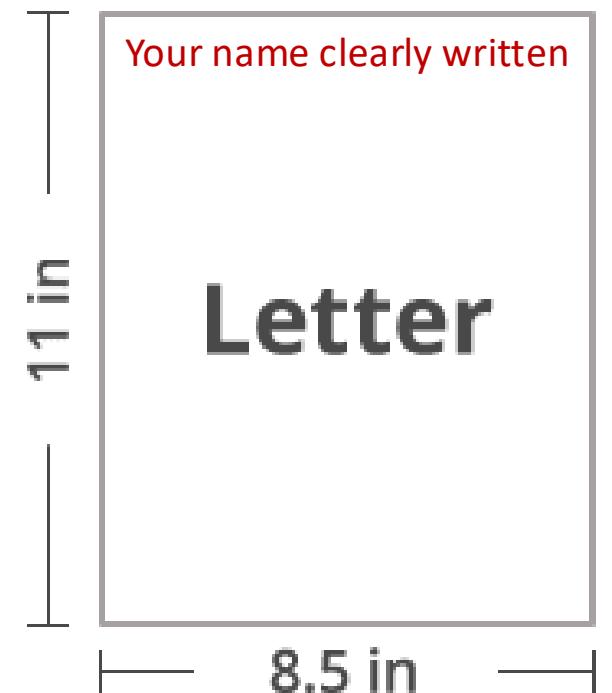
- Your name on the upper left of both sides of the paper

Recommend making a typed list of all functions discussed in class and on the homework

- This will be useful beyond the exam

You must turn in your cheat sheet with the exam

- Failure to do so will result in a 20 point deduction



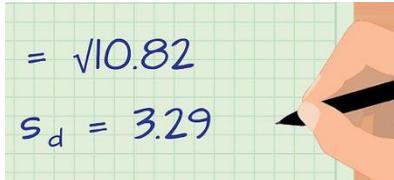
Hypothesis tests for a single proportion continued

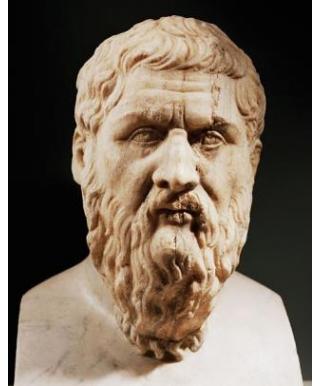
Five steps of hypothesis testing

1. State H_0 and H_A

- Assume Gorgias (H_0) was right

2. Calculate the actual observed statistic


$$= \sqrt{10.82}$$
$$s_d = 3.29$$

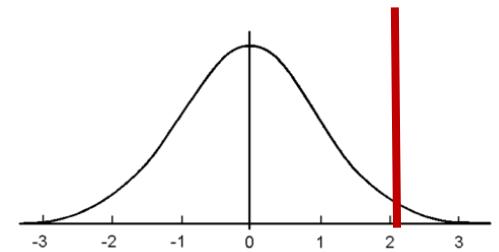


3. Create a **null distribution** of statistics that are consistent with H_0

- i.e., a distribution of statistics that we would expect if Gorgias is right

4. Get the probability we would get a statistic more than the observed statistic from the null distribution

- p-value



5. Make a judgement

- Assess whether the results are statistically significant



Are lie detectors more than 60% accurate?

A study by Hollien, Harnsberger, Martin and Hollien (2010) tried to assess the accuracy of lie detection software

A sample of 48 participants were gathered and attached to a lie detection device. They were asked to read deceptive (lying) material out loud

The lie detector correctly reported that 31 out of the 48 participants were lying

Does this provide evidence that lie detectors are more than 60% accurate?

Questions about the study

1. What are the cases here?
2. What is the variable of interest and is it categorical or quantitative?
3. What is the observed statistic - and what symbols should we use to denote it?
4. What is the population parameter we are trying to estimate - and what symbol should we use to denote it?

Sketch the dataset!

Variable

Cases

5 steps to null-hypothesis significance testing (NHST)

5 steps of hypothesis testing:

1. State null and alternative hypotheses
2. Calculate statistic of interest
3. Create a null distribution
4. Calculate a p-value
5. Assess if there is convincing evidence to reject the null hypothesis

Let's go through these 5 steps now!

Step 1: State the null and alternative hypotheses

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

Alternative Hypothesis (H_a): Claim for which we seek significant evidence

These claims are always made in terms of population parameters!

Lie detector study

Q: What is the null hypothesis? (please state it using words)

Q: How would you write it in terms of the population parameter?

Q: What is the alternative hypothesis?

Step 2: Calculate statistic of interest

For the lie detector study, what was the observed statistic?

Step 3: Create a null distribution

Q: Please describe what the null distribution is here

Q: How can we create a null distribution?

Step 3: Create a null distribution

Please answer the following questions for the lie detector study

- 1. How many coins should we flip?**

- 2. What should the probability of heads be on each flip?**

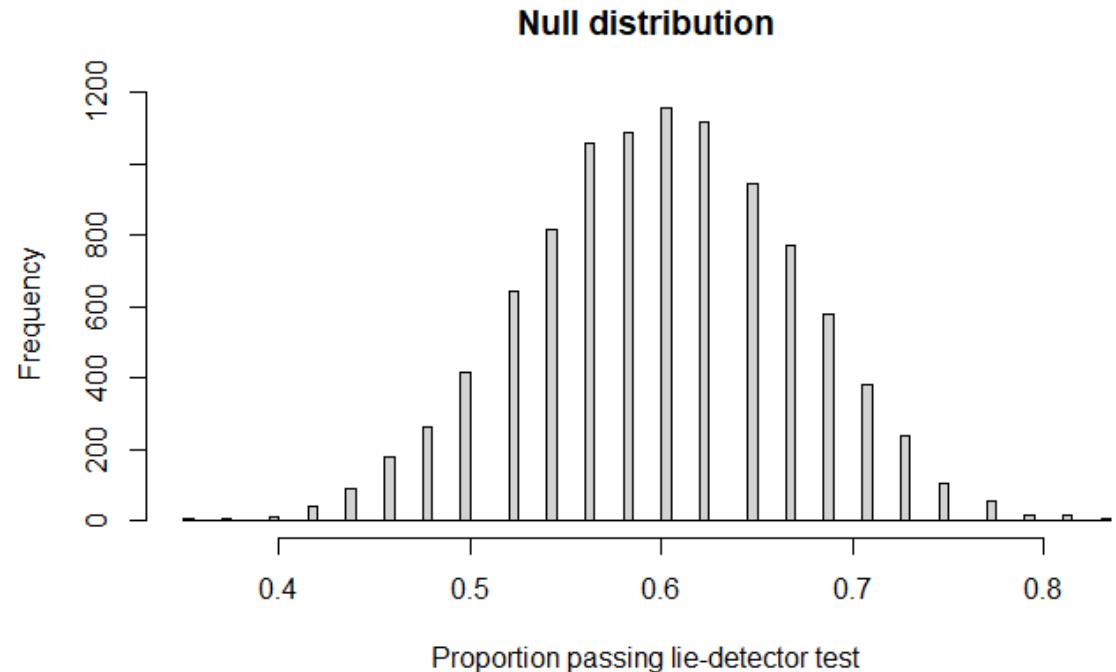
- 3. How many simulations should we run?**

Step 3: Create a null distribution

A null distribution (\hat{p} 's) based on:

- ***10,000 simulations***
- Each simulation consists of flipping 48 coins
- With the probability of getting a head on each flip of 0.60

```
null_dist <- do_it(10000) * {  
  rflip(48, prob = .6)  
}
```



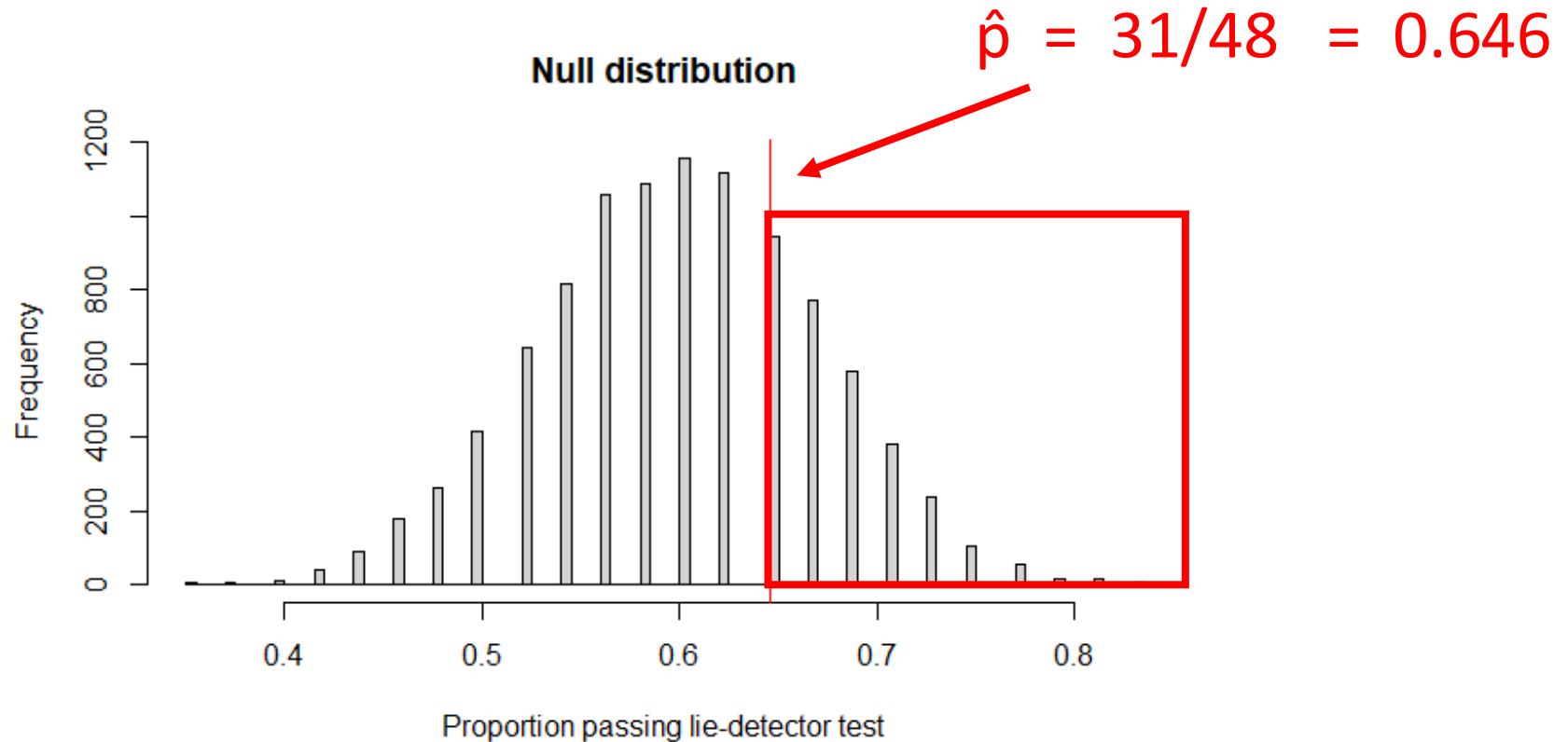
Step 4: Calculate a p-value

The **p-value** is the probability, when the null hypothesis is true, of obtaining a statistic as extreme or more extreme than the observed statistic

$$P(\text{STAT} \geq \text{observed statistic} \mid H_0 = \text{True})$$

The smaller the p-value, the stronger the statistic evidence is against the null hypothesis and in favor of the alternative

Step 4: Calculate a p-value



What is the p-value here?

Step 5a: Assess if results are statistically significant

When our observed sample statistic is unlikely to come from the null distribution, we say the sample results are **statistically significant**

- i.e., we have a small p-value
- Often if the p-value is less than 0.05 we say the results are "statistically significant"
 - (although perhaps we should use a lower threshold of 0.01 or 0.005 and/or not use this terminology at all)

'Statistically significant' results mean we have convincing evidence against H_0 in favor of H_A

Step 5b: Make a decision

Are the results from the lie detector study statistically significant?



Let's try the lie detector example in R...

One-sided vs. two-sided tests

One-sided vs. two-sided

In the examples we have seen, we were just interested if the parameter was greater (or less) than a hypothesized value

$$H_0: \pi = 0.60 \quad H_A: \pi > 0.60$$

In other cases, we might not have a directional alternative hypothesis

Testing whether a lie detector is not 60% accurate

Suppose we wanted to test what whether the lie detector was correct more **or less** than 60% of the time

- i.e., we are testing whether the lie detector is **not** 60% accurate

Step 1: Write down the null and alternative hypotheses

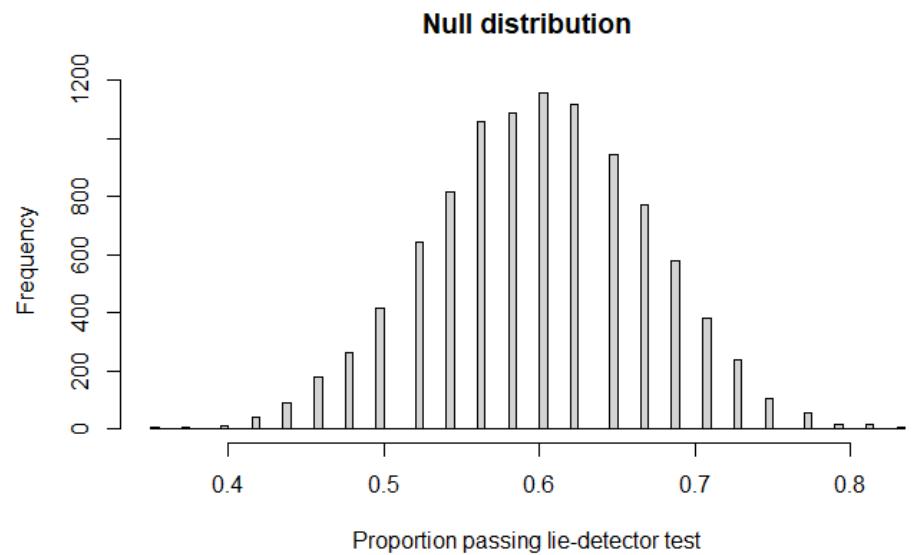
$$H_0: \pi = 0.60$$

$$H_A: \pi \neq 0.60$$

Testing whether a lie detector is not 60% accurate

Step 2: Would the statement of hypotheses affect the observed statistic value?

Step 3: Would this change in statement of hypotheses affect the null distribution?

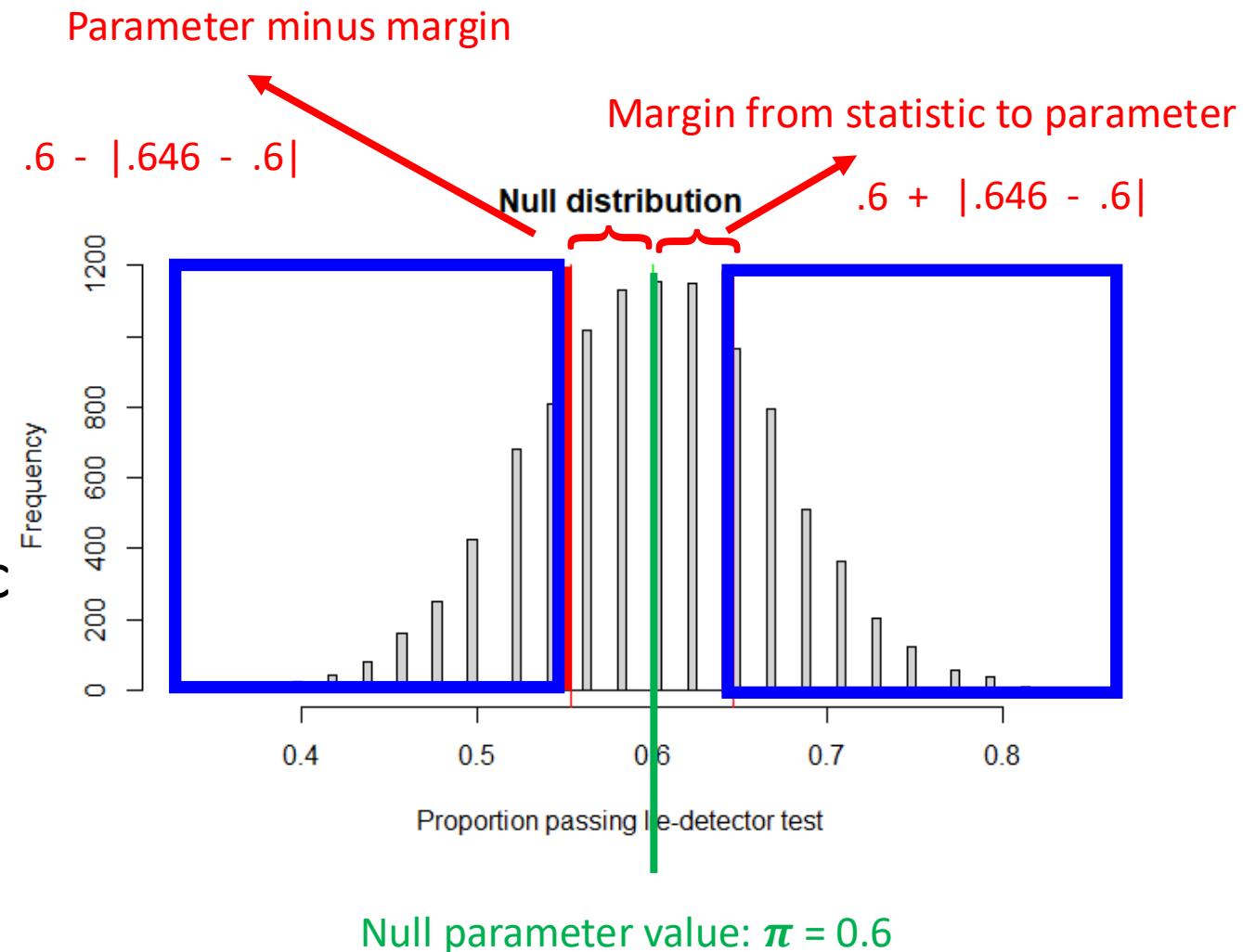


Testing whether a lie detector is not 60% accurate

Step 4: Would the statement of hypotheses affect p-value?

We need to look for values **more extreme** than the observed statistic

Thus, the p-value for a two-tailed test is about twice as large



Statement of alternative hypothesis is important

We need to state what you expect before analyzing the data

Our expectation (hypothesis statement) can change the p-value!

Estimating a p-value from a null distribution

For a one tailed alternative: Find the proportion of statistics in the null distribution that equal or exceed the original statistic in the direction (tail) indicated by the alternative hypothesis

For a two-tailed alternative: Find the proportion of statistics in the null distribution beyond the deviation of the observed statistic from the parameter value in both tails

- Alternatively, find the proportion of statistics in the null distribution beyond the original statistic in one of the tails, and then double the proportion to account for the other tail

How to estimate two sided p-values in R?

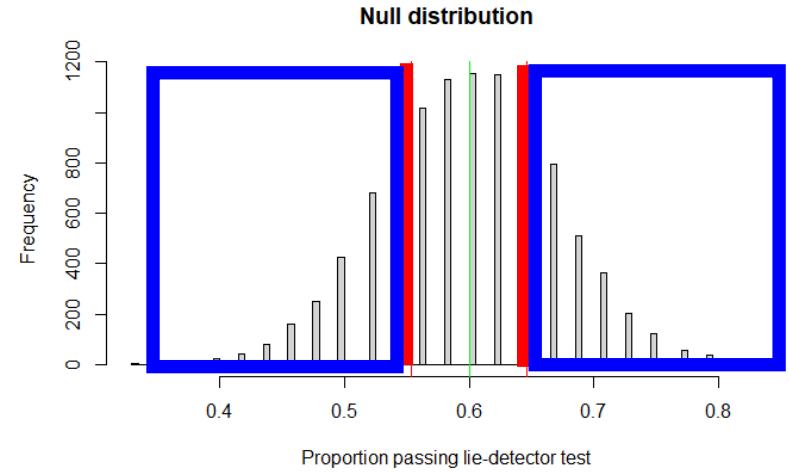
```
null_distribution <- do_it(10000) * {  
  rflip(48, prob = .6)/48  
}
```

```
stat_param_margin <- abs(obs_stat - .6)
```

```
pval_right_tail <- pnull(.6 + stat_param_margin, null_dist, lower.tail = FALSE)
```

```
pval_left_tail <- pnull(.6 - stat_param_margin , null_dist, lower.tail = TRUE)
```

```
p_value <- p_right_tail + p_left_tail
```



Let's try it in R...

Hypothesis tests for comparing two means

Testing whether a pill is effective

How would we design a study?

What would the cases and variables be?

What are the null and alternative hypotheses?

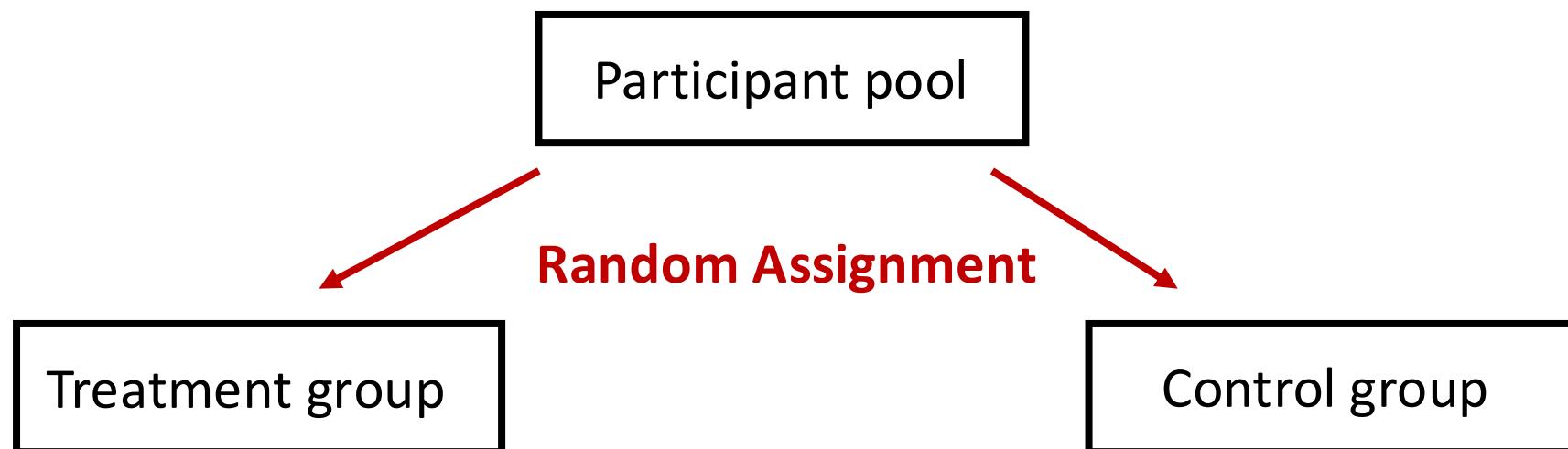
- Assume we are looking for differences in means between the groups

What would the statistic of interest be?

Experimental design: randomized controlled trial

Take a group of participant and ***randomly assign***:

- Half to a *treatment group* where they get the pill
- Half in a *control group* where they get a fake pill (placebo)
- See if there is more improvement in the treatment group compared to the control group



Example: Does calcium reduce blood pressure?

A randomized controlled trial by Lyle et al (1987) investigated whether calcium lowered blood pressure

- A treatment group of 10 men received a calcium supplement for 12 weeks
- A control group of 11 men received a placebo during the same period

The blood pressure of these men was taken before and after the 12 weeks of the study

1. What are the null and alternative hypotheses?

Hypothesis tests for differences in two group means

1. State the null and alternative hypothesis

$$H_0: \mu_{\text{Treatment}} = \mu_{\text{Control}} \quad \text{or} \quad \mu_{\text{Treatment}} - \mu_{\text{Control}} = 0$$

$$H_A: \mu_{\text{Treatment}} > \mu_{\text{Control}} \quad \text{or} \quad \mu_{\text{Treatment}} - \mu_{\text{Control}} > 0$$

2. Write down the statistic of interest using appropriate symbols

$$\bar{x}_{\text{Effect}} = \bar{x}_{\text{Treatment}} - \bar{x}_{\text{Control}}$$

Does calcium reduce blood pressure?

Treatment data (n = 10):

Begin	107	110	123	129	112	111	107	112	136	102
End	100	114	105	112	115	116	106	102	125	104
Decrease	7	-4	18	17	-3	-5	1	10	11	-2

Control data (n = 11):

Begin	123	109	112	102	98	114	119	112	110	117	130
End	124	97	113	105	95	119	114	114	121	118	133
Decrease	-1	12	-1	-3	3	-5	5	2	-11	-1	-3

2. What is the observed statistic of interest?

- $\bar{x}_{\text{Effect}} = 5 - -.2727 = 5.273$

3. What is step 3?

3. Create the null distribution!

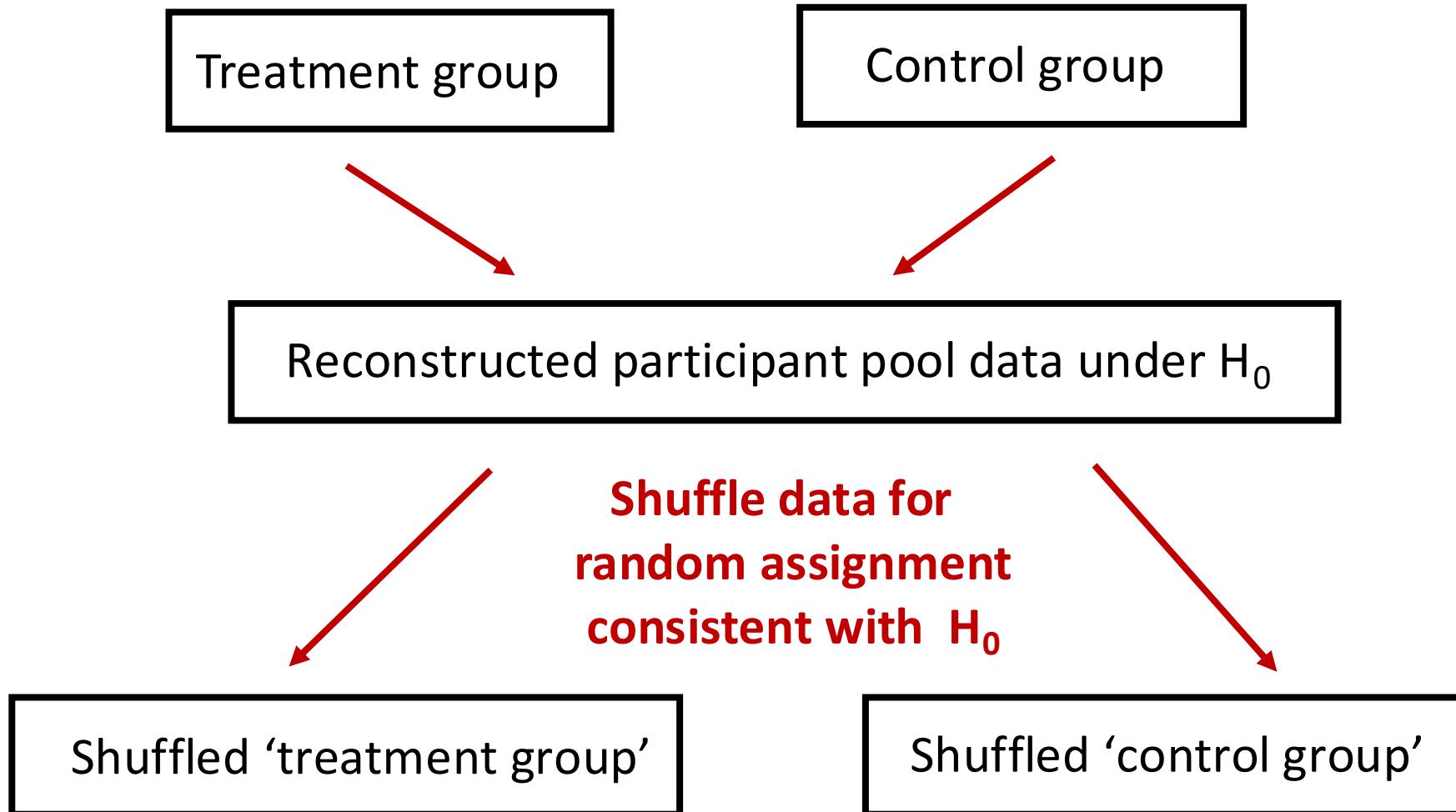
How could we create the null distribution?

Need to generate data consistent with H_0 : $\mu_{\text{Treatment}} - \mu_{\text{Control}} = 0$

- i.e., we need fake \bar{x}_{Effect} that are consistent with H_0

Any ideas how we could do this?

3. Create the null distribution!



One null distribution statistic: $\bar{x}_{\text{Shuff_Treatment}} - \bar{x}_{\text{Shuff_control}}$

3. Create a null distribution

1. Combine data from both groups
2. Shuffle data
3. Randomly select 10 points to be the ‘shuffled’ treatment group
4. Take the remaining points to the ‘shuffled’ control group
5. Compute the statistic of interest on these ‘shuffled’ groups
6. Repeat 10,000 times to get a null distribution

3. Creating a null distribution in R

```
# the data from the calcium study
treat <- c(7, -4, 18, 17, -3, -5, 1, 10, 11, -2)
control <- c(-1, 12, -1, -3, 3, -5, 5, 2, -11, -1, -3)

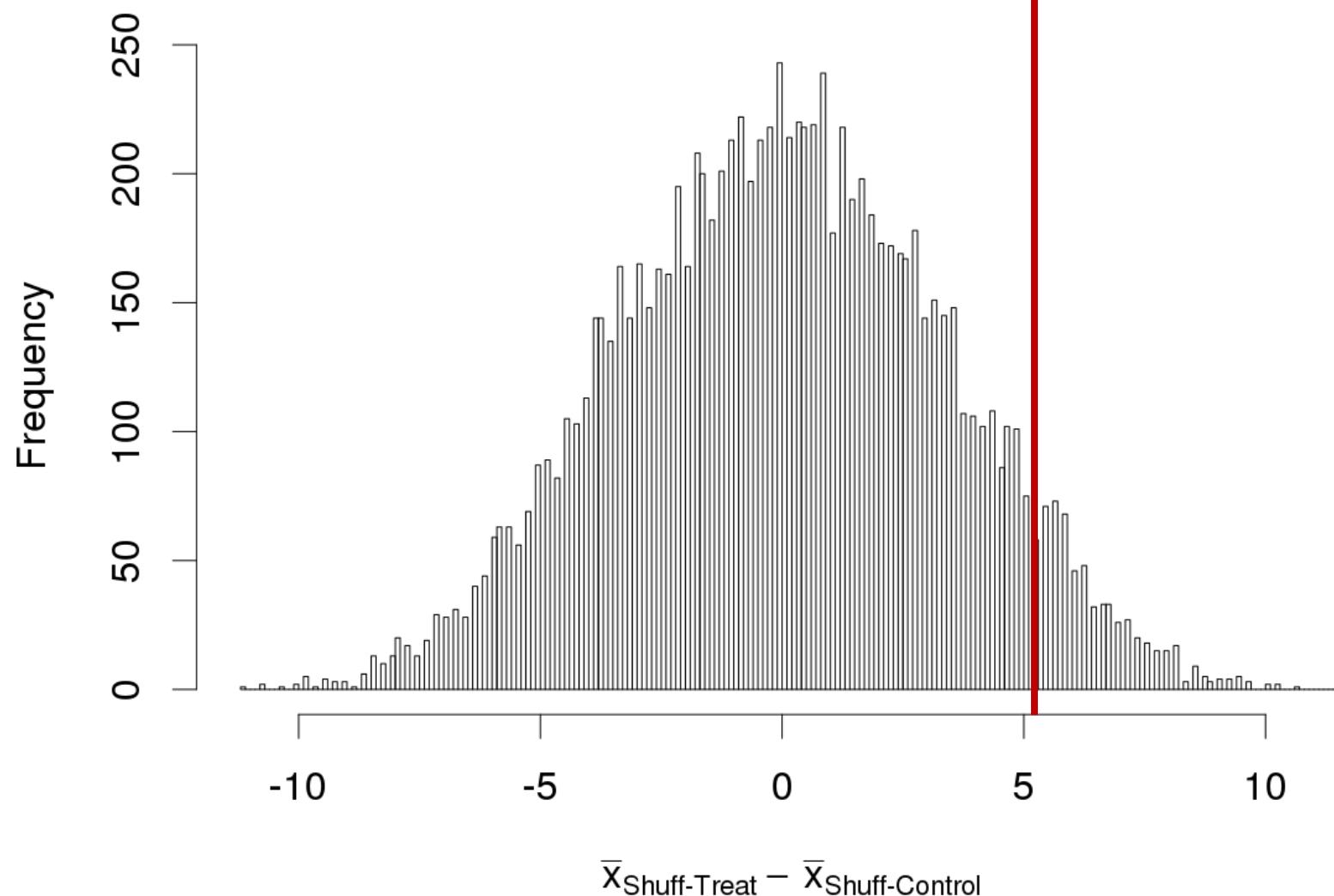
# observed statistic
obs_stat <- mean(treat) - mean(control)

# Combine data from both groups
combined_data <- c(treat, control)
```

3. Creating a null distribution in R

```
null_distribution <- do_it(10000) * {  
  
  # shuffle data  
  shuff_data <- shuffle(combined_data)  
  
  # create fake treatment and control groups  
  shuff_treat <- shuff_data[1:10]  
  shuff_control <- shuff_data[11:21]  
  
  # save the statistic of interest  
  mean(shuff_treat) - mean(shuff_control)  
}
```

Null distribution



`hist(null_distribution, breaks = 200)`

Next step?

4. Calculate the p-value

```
# Calculate the p-value
```

```
p_value <- pnull(obs_stat, null_distribution, lower.tail = FALSE)
```

p-value = .064

Next step?

5. Are the results statistically significant?



What should we do?