

Hypothesis tests for more than two means

	5	3	2		7			8
6		1	5					2
2			9	1	3		5	
7	1	4	6	9	2			
	2						6	
			4	5	1	2	9	7
	6		3	2	5			9
1					6	3		4
8			1		9	6	7	

Overview

Hypothesis tests for more than two means

Hypothesis tests for correlation

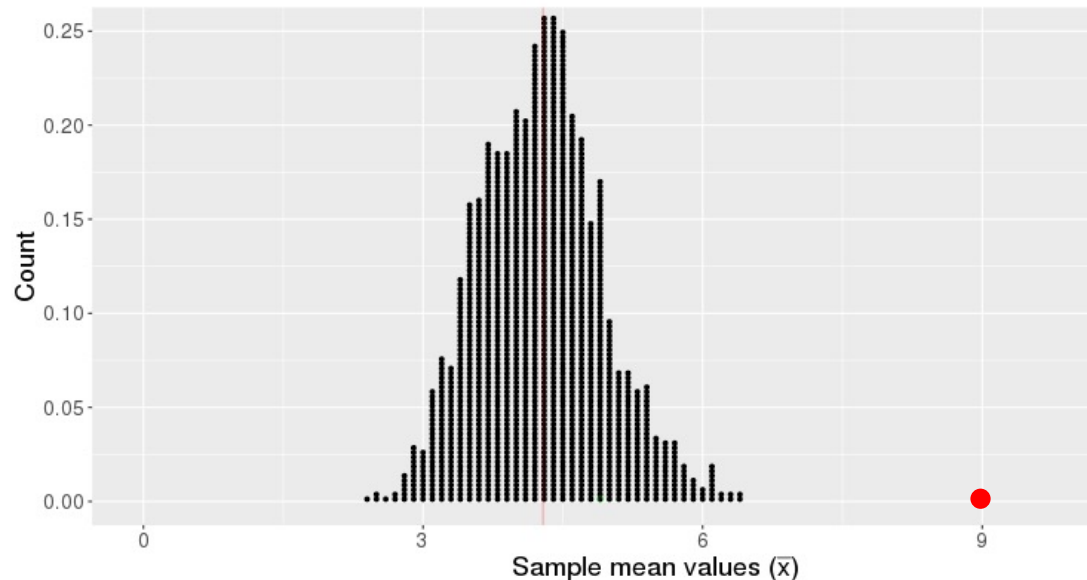
Hypothesis tests for more than two means

The logic of hypothesis tests...

We start with a claim about a population parameter

- E.g., $\mu = 4$

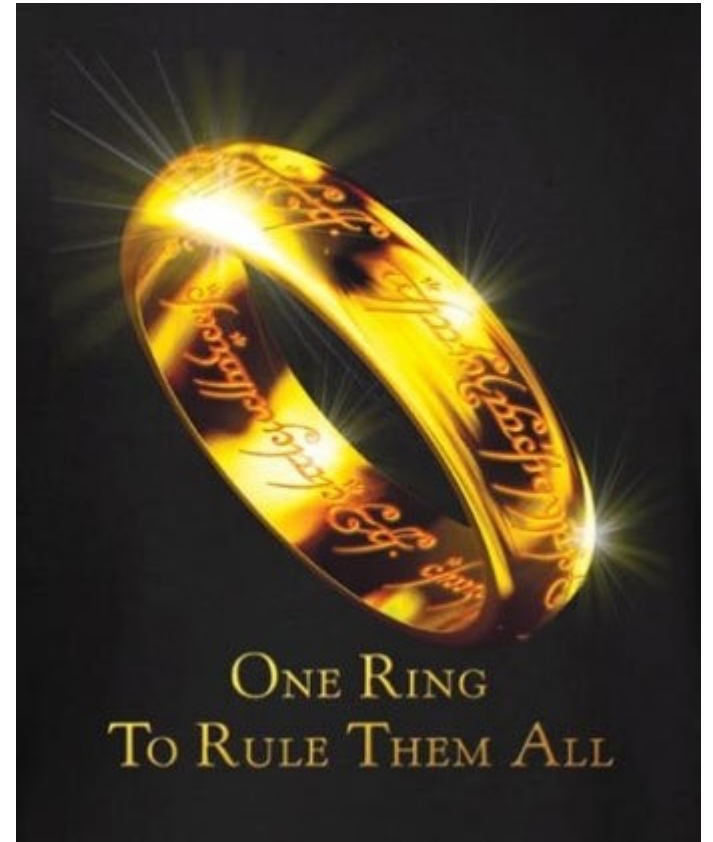
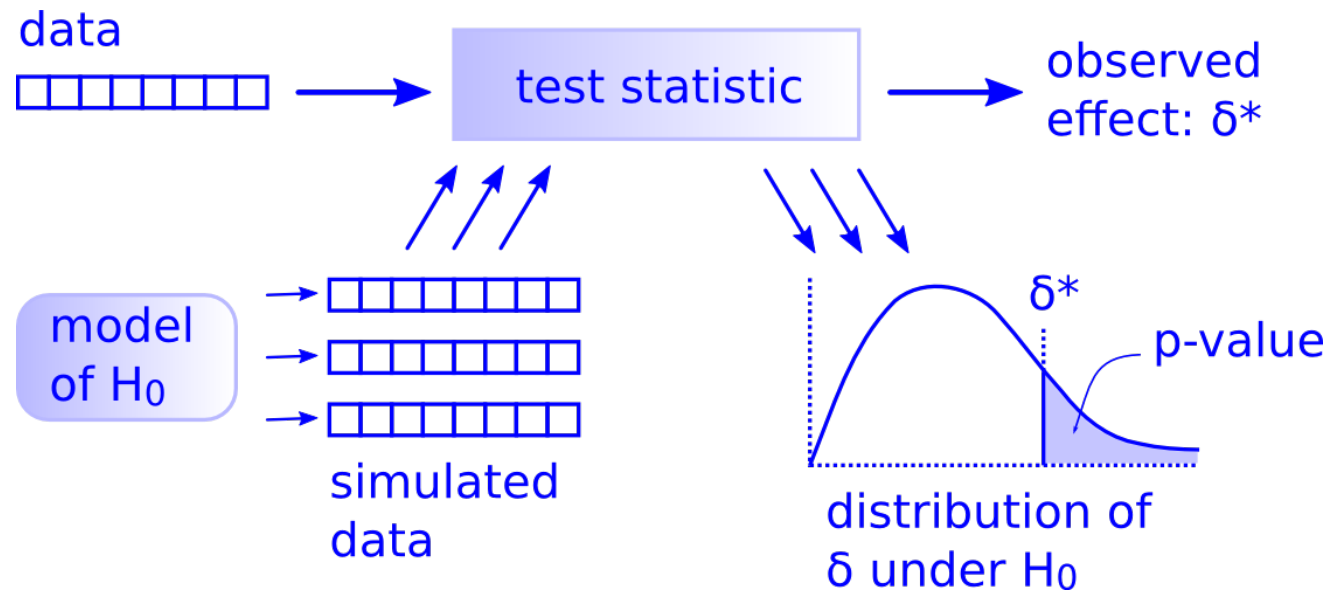
This claim implies we should get a certain distribution of statistics



If our observed statistic is highly unlikely, we reject the claim

The logic of hypothesis tests...

There is only one [hypothesis test](#)!

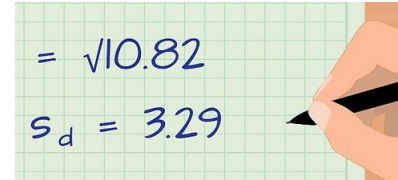


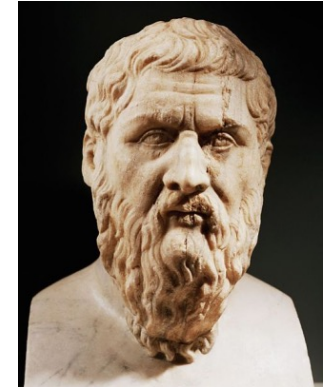
Just follow the 5 hypothesis tests steps!

Five steps of hypothesis testing

1. State H_0 and H_A

- Assume Gorgias (H_0) was right


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



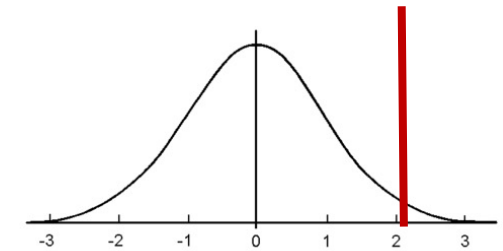
2. Calculate the actual observed statistic

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



Comparing more than two means

A group of Hope College students wanted to see if there was an association between a student's major and the time it takes to complete a small Sudoku-like puzzle

	5	3	2		7			8
6		1	5					2
2			9	1	3		5	
7	1	4	6	9	2			
	2						6	
			4	5	1	2	9	7
	6		3	2	5			9
1					6	3		4
8			1		9	6	7	

Comparing more than two means

A group of Hope College students wanted to see if there was an association between a student's major and the time it takes to complete a small Sudoku-like puzzle

They grouped majors into four categories

- Applied science (as)
- Natural science (ns)
- Social science (ss)
- Arts/humanities (ah)

What is the first step of hypothesis testing?

Sudoku by field

1. State the null and alternative hypotheses!

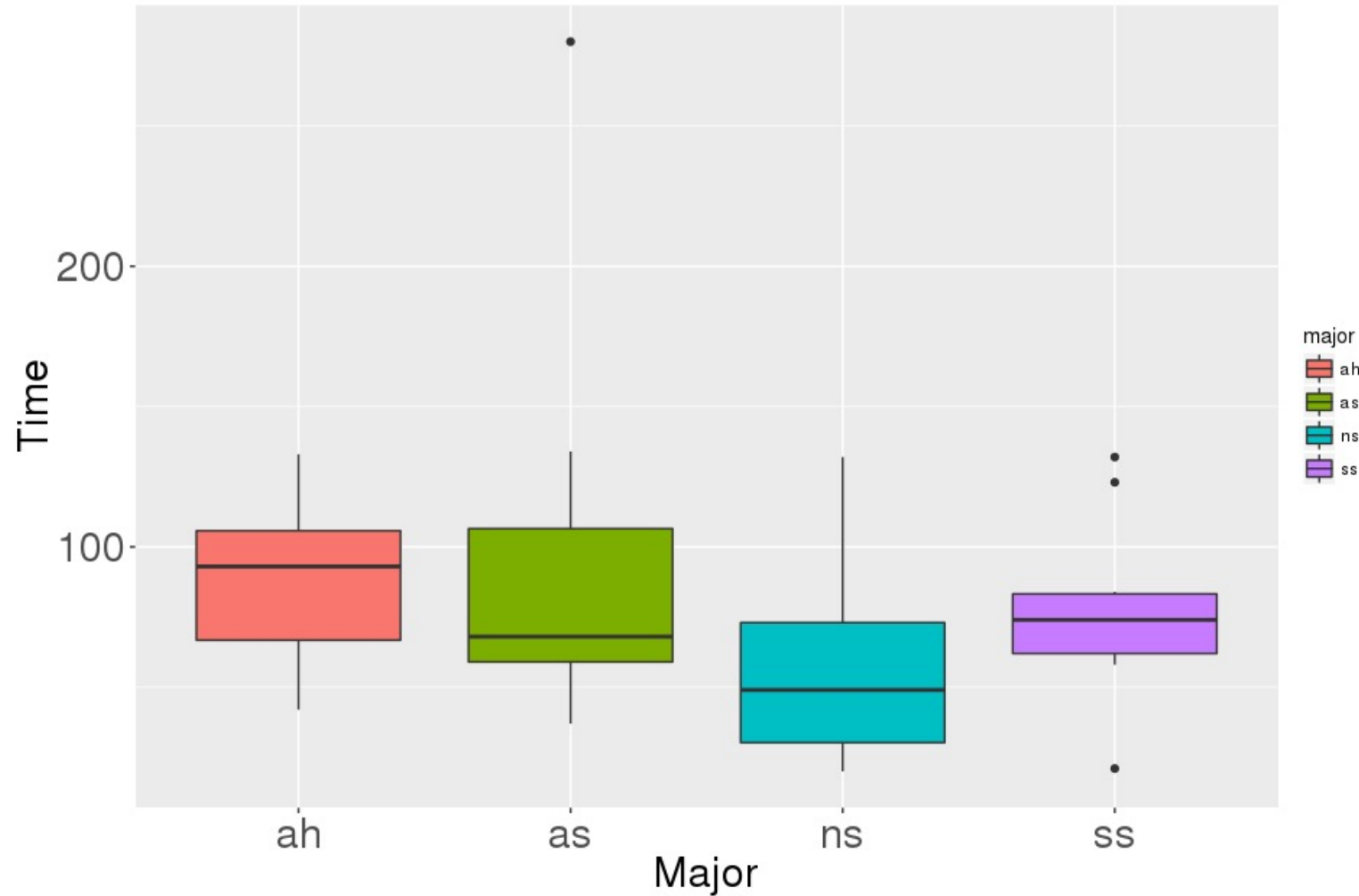
$$\mathbf{H}_0: \mu_{as} = \mu_{ns} = \mu_{ss} = \mu_{ah}$$

$$\mathbf{H}_A: \mu_i \neq \mu_j \text{ for one pair of fields of study}$$

What should we do next?

Let's plot the data first...

Step 2a: Plot of completion time by major



What should we do next?

Sudoku by field

1. State the null and alternative hypotheses!

$$\mathbf{H}_0: \mu_{as} = \mu_{ns} = \mu_{ss} = \mu_{ah}$$

$$\mathbf{H}_A: \mu_i \neq \mu_j \text{ for one pair of fields of study}$$

Thoughts on the statistic of interest?

Comparing multiple means

There are many possible statistics we could use. A few choices are:

1. Group range statistic:

$$\max \bar{x} - \min \bar{x}$$

2. Mean absolute difference (MAD):

$$(|\bar{x}_{as} - \bar{x}_{ns}| + |\bar{x}_{as} - \bar{x}_{ss}| + |\bar{x}_{as} - \bar{x}_{ah}| + |\bar{x}_{ns} - \bar{x}_{ss}| + |\bar{x}_{ns} - \bar{x}_{ah}| + |\bar{x}_{ss} - \bar{x}_{ah}|)/6$$

3. F statistic:

$$F = \frac{\text{between-group variability}}{\text{within-group variability}}$$

Using the MAD statistic

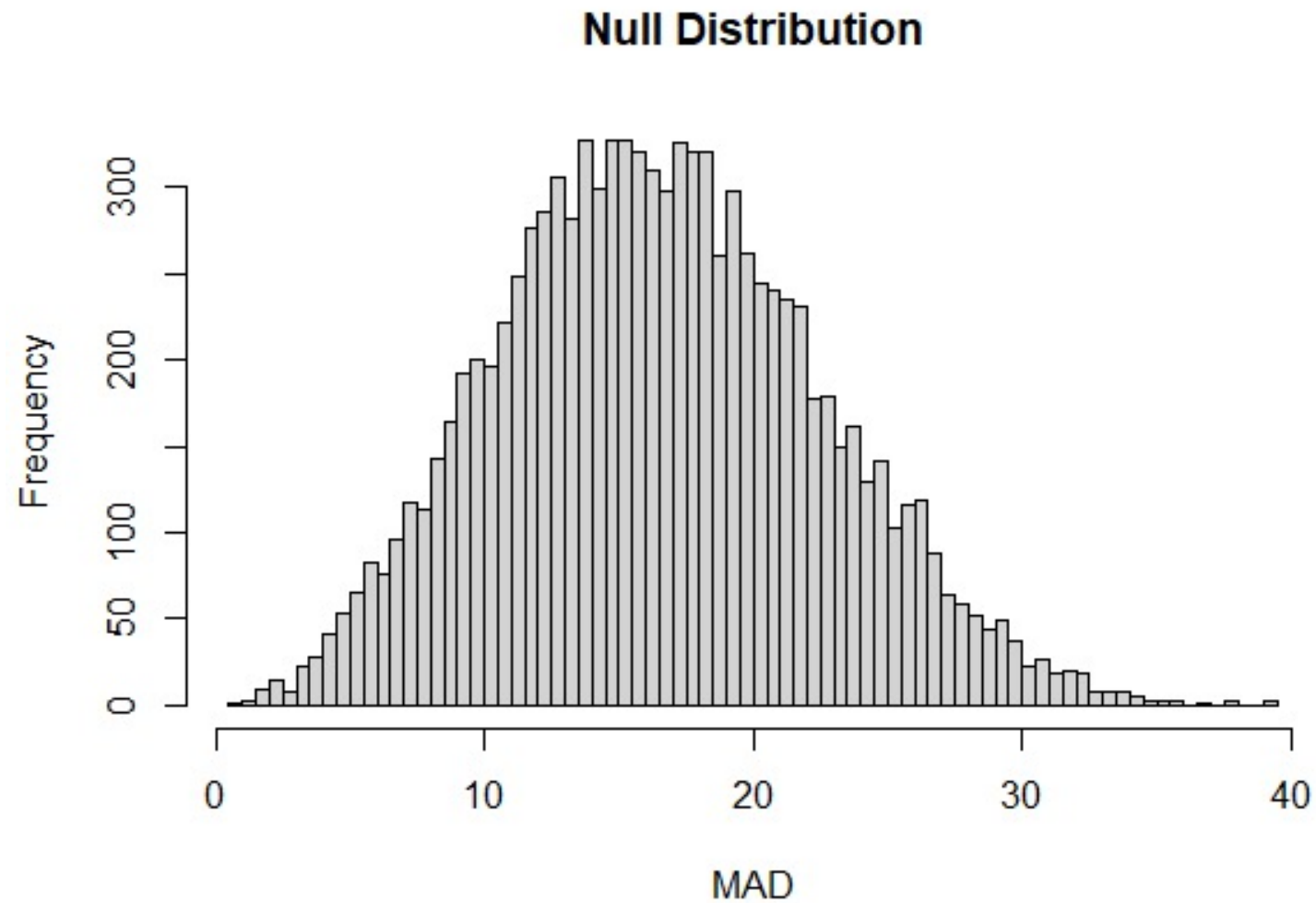
Mean absolute difference (MAD):

$$(|\bar{x}_{as} - \bar{x}_{ns}| + |\bar{x}_{as} - \bar{x}_{ss}| + |\bar{x}_{as} - \bar{x}_{ah}| + |\bar{x}_{ns} - \bar{x}_{ss}| + |\bar{x}_{ns} - \bar{x}_{ah}| + |\bar{x}_{ss} - \bar{x}_{ah}|)/6$$

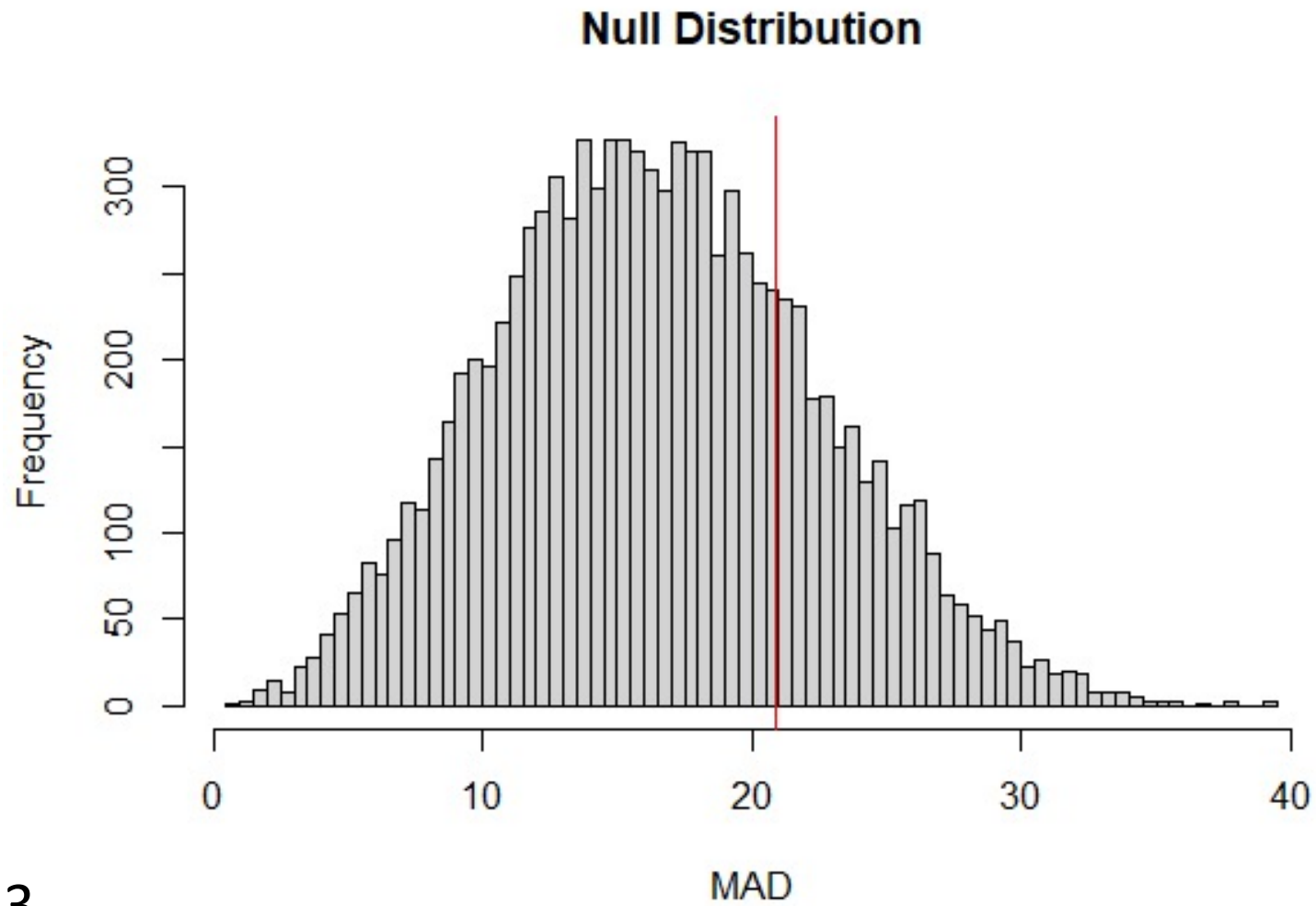
Observed statistic value = 20.88

How can we create the null distribution?

Null distribution



P-value



Conclusions?



Hypothesis tests for more than two means in R

Step 1: null and alternative hypotheses...

$$H_0: \mu_{as} = \mu_{ns} = \mu_{ss} = \mu_{ah}$$

$$H_A: \mu_i \neq \mu_j \text{ for one pair of fields of study}$$

	5	3	2		7			8
6		1	5					2
2			9	1	3		5	
7	1	4	6	9	2			
	2						6	
			4	5	1	2	9	7
	6		3	2	5			9
1					6	3		4
8			1		9	6	7	

Let's try this analysis in R...

```
# get the data
```

```
library(SDS100)
```

```
download_data("MajorPuzzle.txt")
```

```
sudoku_data <- read.table("MajorPuzzle.txt", header = TRUE)
```

```
# Extract vectors from the data frame (how do we do this?)
```

```
completion_time <- sudoku_data$time
```

```
major <- sudoku_data$major
```

Visualize the data

How can we visualize the data?

```
# We can create side-by-side boxplots using  
boxplot(completion_time ~ major,  
         xlab = "Major", ylab = "Time (s)")
```

Calculating the statistic of interest

We can get the MAD statistic using the `get_MAD_stat()` function

`get_MAD_stat(data_vector, grouping_vector)`

- `data_vector`: a vector of quantitative data
- `grouping_vector`: a vector indicating which group the quantitative data is in

Can you get the MAD statistic for the sudoku data?

```
obs_stat <- get_MAD_stat(completion_time, major)
```

Creating the null distribution

Q: How could we create one point in a null distribution?

- A: Shuffle the grouping_vector (major vector) and calculate the MAD statistic

Q: How can we do this in R?

```
shuffled_majors <- shuffle(major)
```

```
get_MAD_stat(completion_time, shuffled_majors)
```

Creating the null distribution

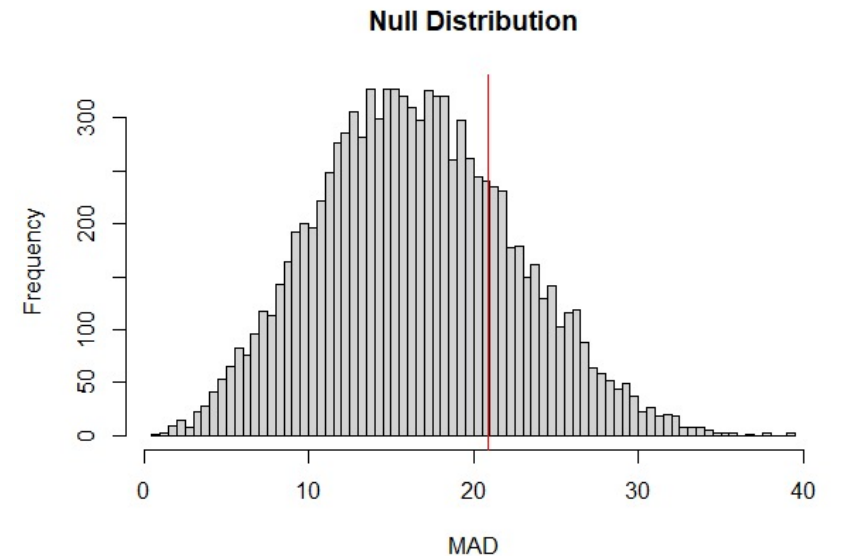
Q: How can we create a full null distribution?

```
null_dist <- do_it(10000) * {  
  shuffled_majors <- shuffle(major)  
  get_MAD_stat(completion_time, shuffled_majors)  
}
```

visualize the null distribution

```
hist(null_dist, breaks = 100)
```

```
abline(v = obs_stat, col = "red")
```



Creating the null distribution

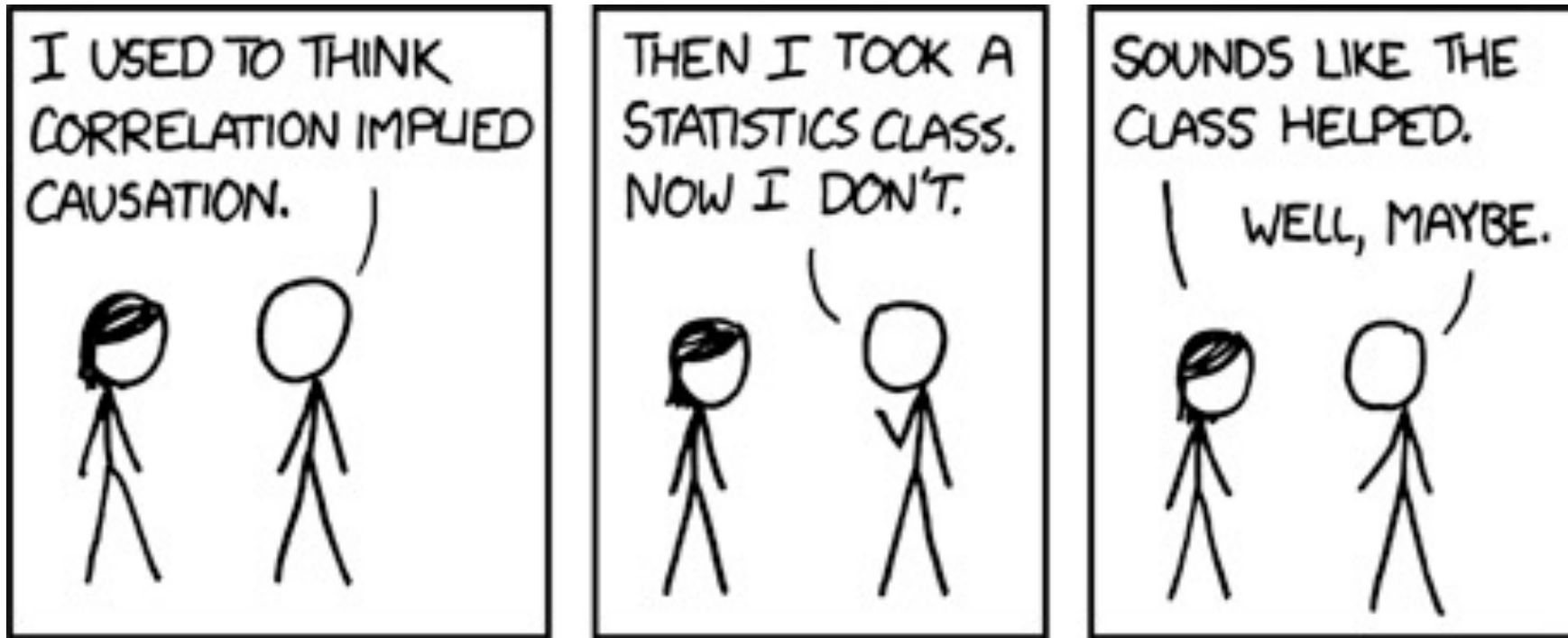
Q: What do we do next and how do we do it?

- A: We get the p-value

`pnull(obs_stat, null_dist, lower.tail = FALSE)`

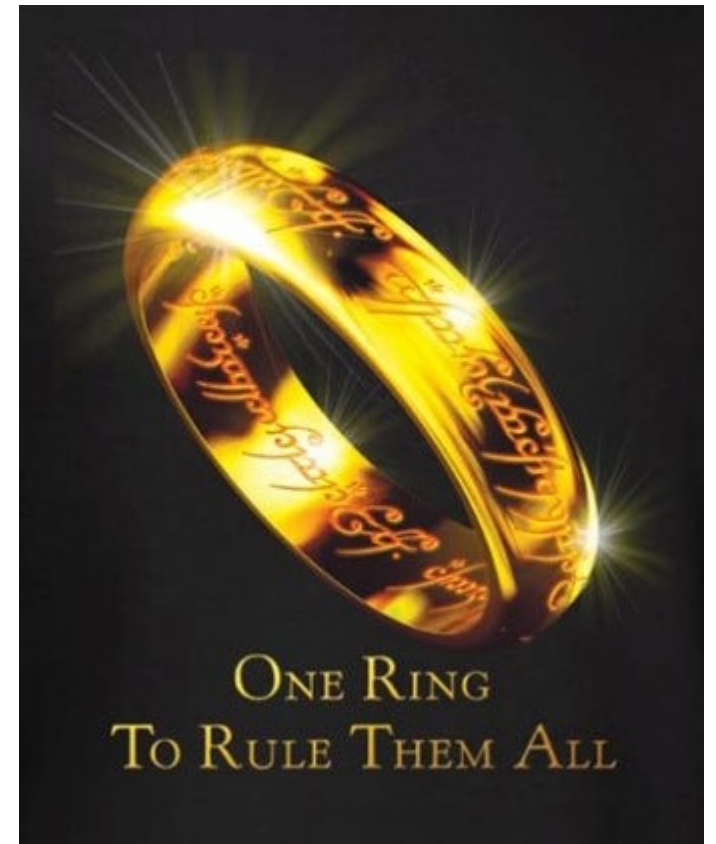
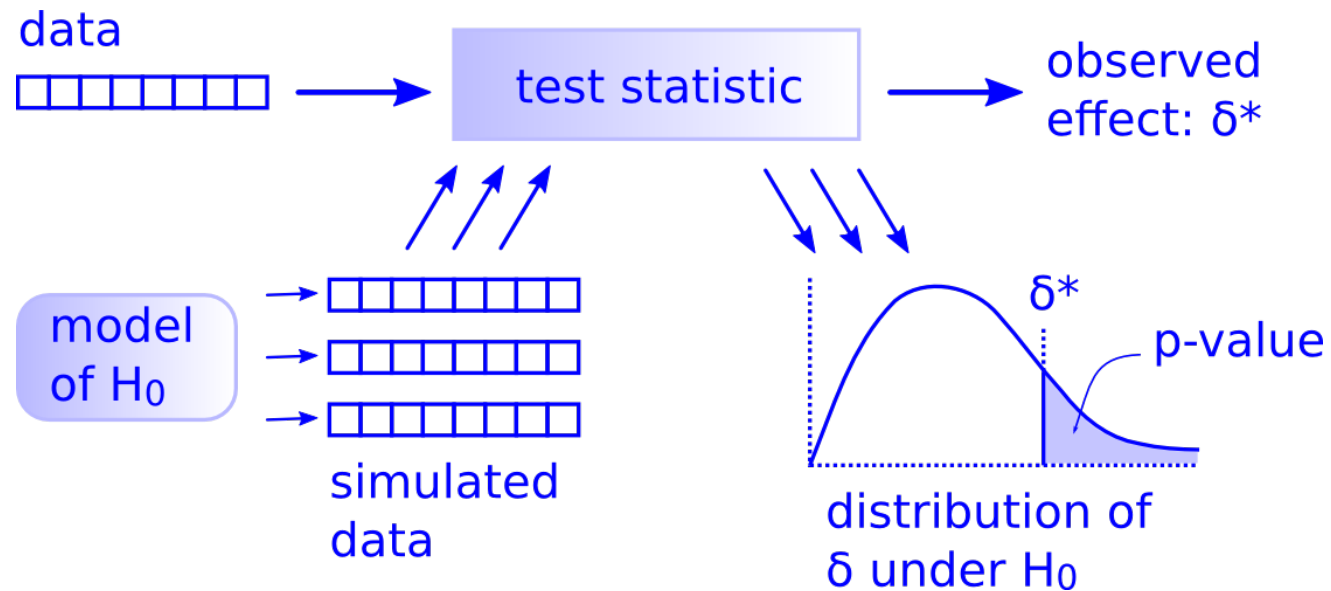


Hypothesis tests for correlation



The logic of hypothesis tests

There is only one [hypothesis test](#)!



Just follow the 5 hypothesis tests steps!

Hypothesis tests for correlation

Is there a positive correlation between the number of carbohydrates in a cereal and the number calories?



What is the population parameter and the statistic of interest?

Significance tests for correlation

Let's look at data from 30 randomly selected cereals

	Calories	Carbohydrates
AppleJacks	117	27
Boo Berry	118	27
Cap'n Crunch	144	31
Cinnamon Toast Crunch	169	32

What is the first step we should do for running a hypothesis test?

Hypothesis testing for correlation

1. Write down the null and alternative in symbols and words
2. Load the data and compute the observed statistic:
 - > `download_data("cereal.Rda")`
 - > `load("cereal.Rda")`
3. Let's extract the calories and carbohydrates from the data frame
 - > `calories <- cereal$Calories`
 - > `carbs <- cereal$Carbs`

Try this at home!

Step 2: What is the observed statistic?

- Also say whether you think you will be able to reject the null hypothesis based on a plot of your data

Step 3: Create the null distribution

- To start with: how we can create one point in the null distribution?
 - Hint: think about shuffling the data
 - If you are able to create a null distribution, upload a plot of it to Canvas

Step 4: What is the p-value that you get?

Step 5: What decision would you make?

We will pick up from here next class...

