Hypothesis tests for more than two means and for correlation

Overview

Class logistics and taking stock of where we are

Review/continuation hypothesis tests for more than two means

Hypothesis tests for correlation

Plan for the rest of the semester

Lectures will be recorded and posted on Canvas twice a week

- Prior to our usual class meeting time of 9am EST

I will try to hold online (Zoom) office hours 3 times a week

- 1 private office hours, 2 open office hours

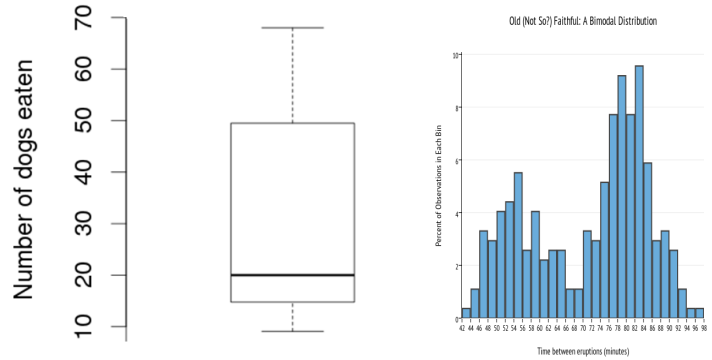
Assignments will be the same as before

- Weekly homework assigned on Tuesdays and due at 11:30pm on Sundays
- Final exam

We are we and where are we going?

Where we are: what we have covered

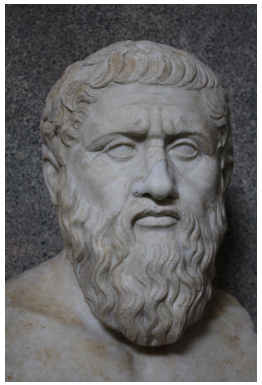
Descriptive statistics

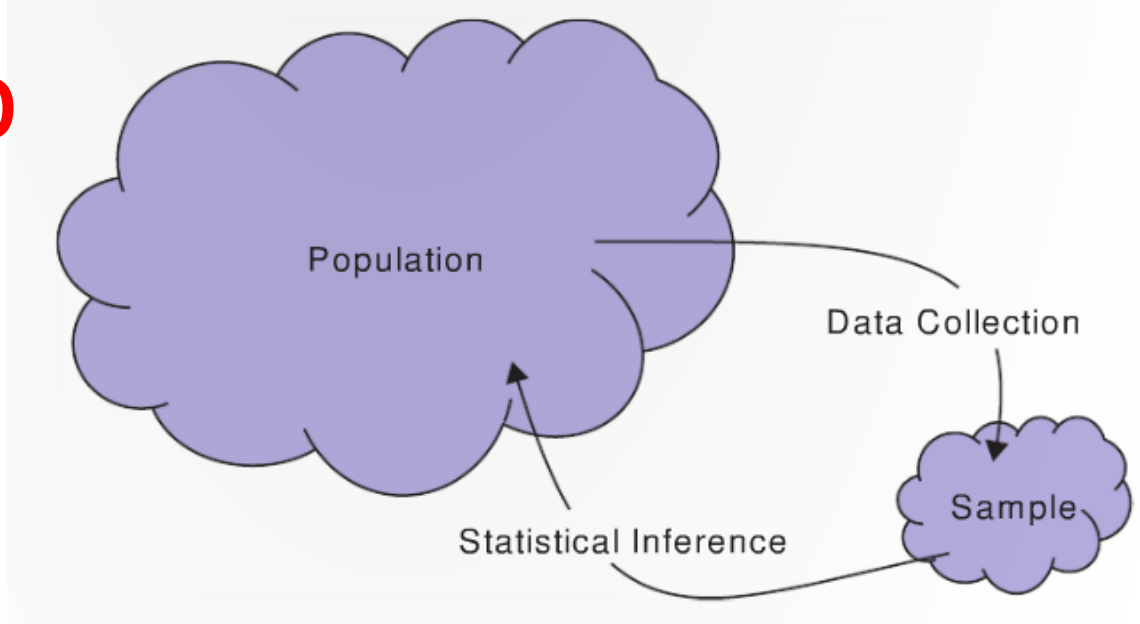


Statistical inference

- We have used computational methods for inference

π, μ, σ, ρ



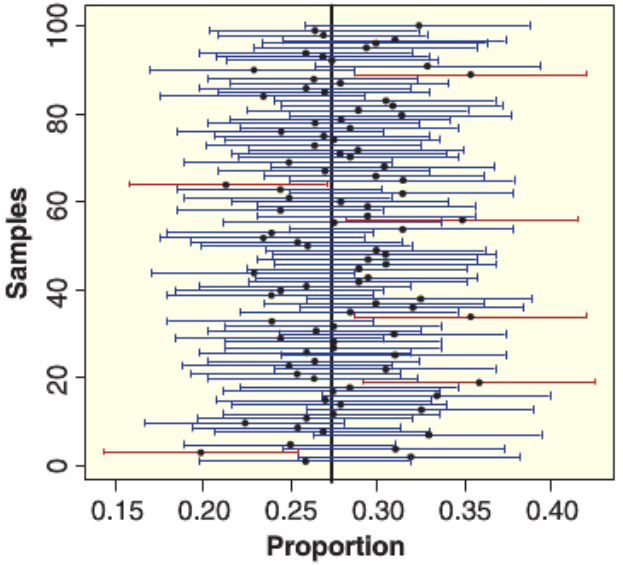


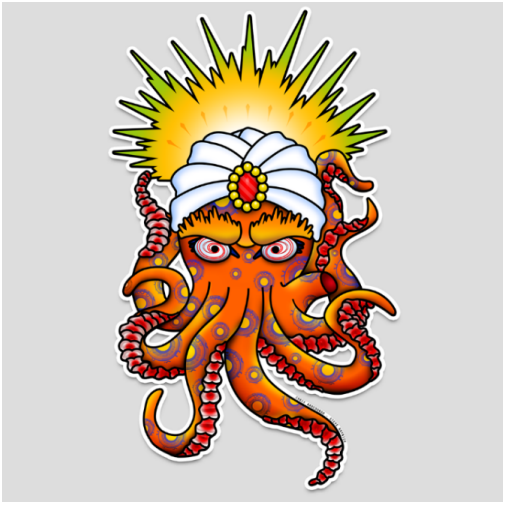
$\hat{p}, \bar{x}, s, r$

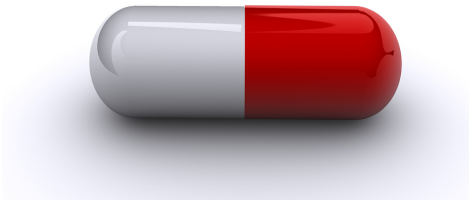
Where we are: what we have covered

Statistical inference using computational methods

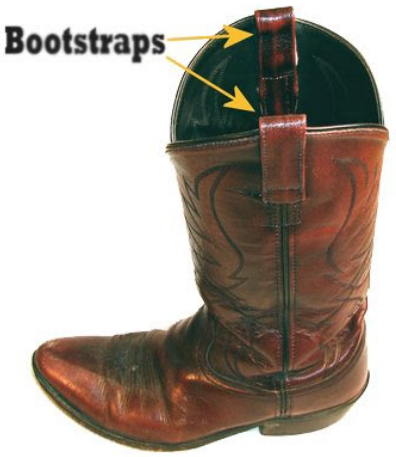
- Confidence intervals using the bootstrap
- Hypothesis tests using permutation/randomization tests
  - Single proportion, two means, more than two means







	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	



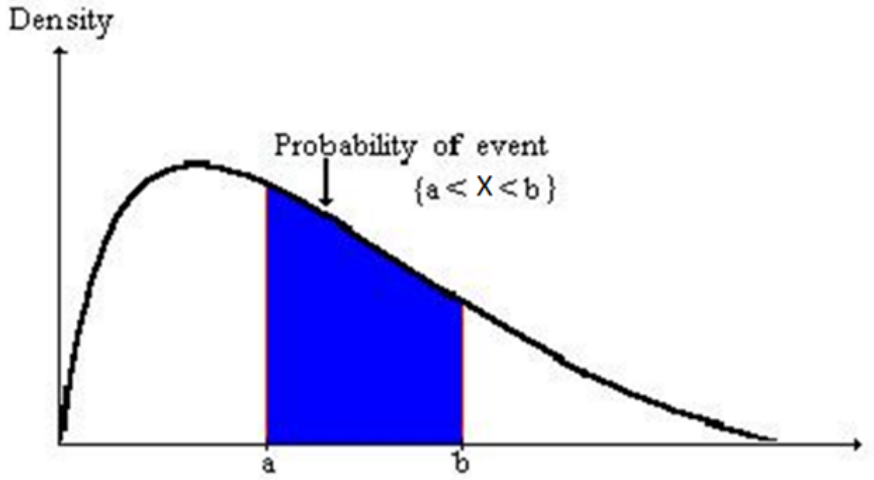
Where we are: where are we going…

Continuation of statistical inference using computational methods

- More than two means continued, correlation, theories of hypothesis tests

Statistical inference based on math/theory

- t-tests, ANOVA, regression, etc.



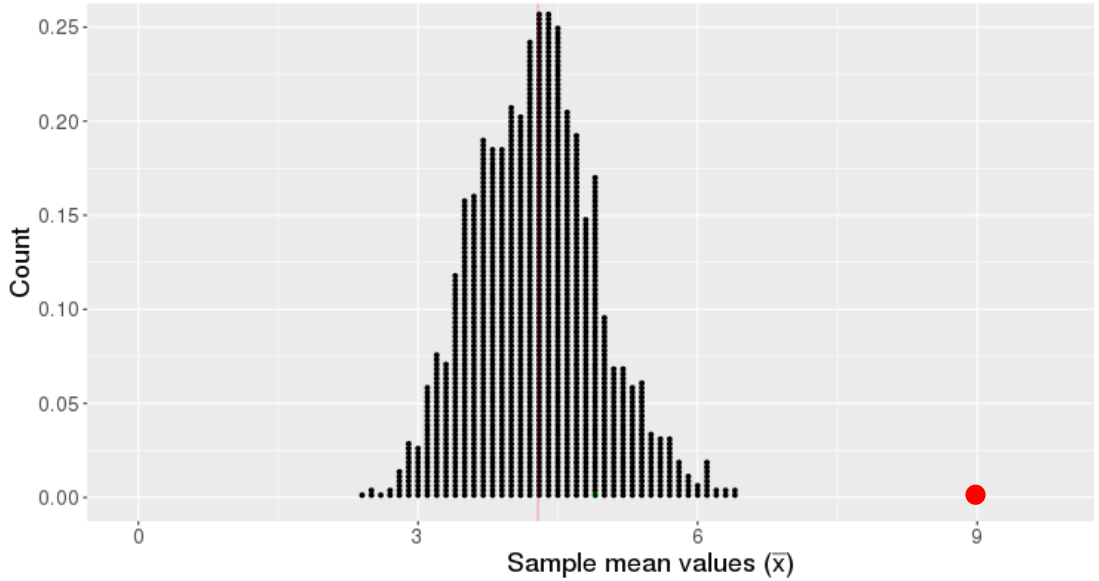
Hypothesis test for more than two means continued…

The logic of hypothesis tests

We start with a claim about a population parameter

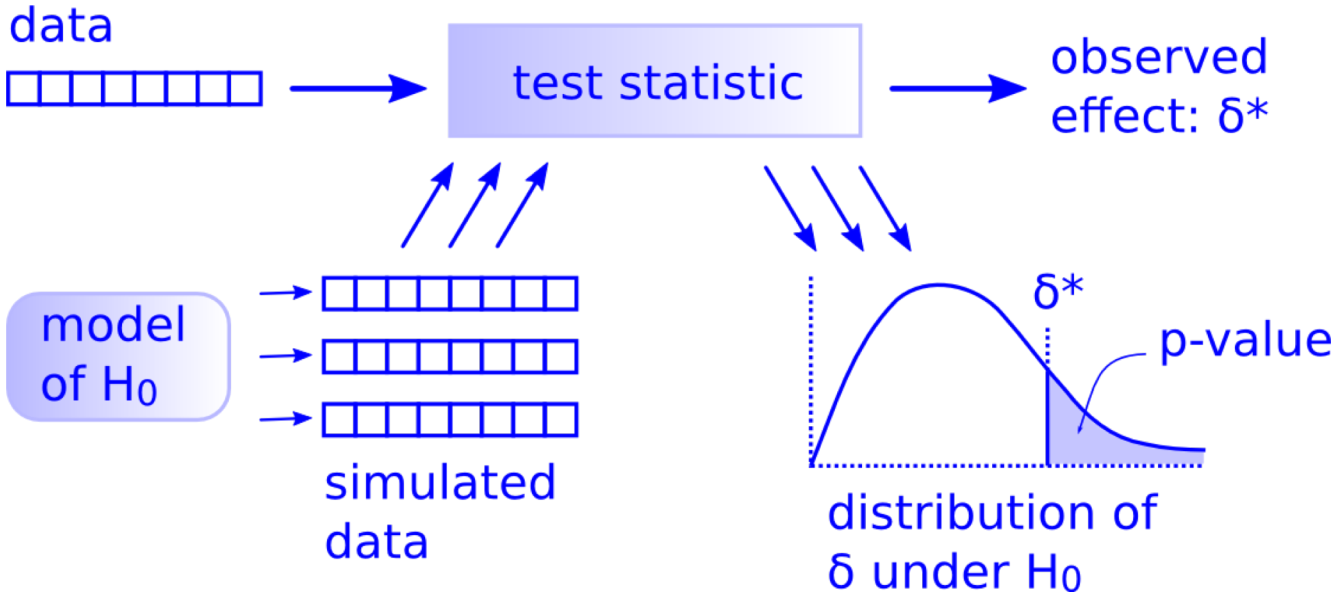
- E.g., μ = 4.2
- This claim is the null hypothesis

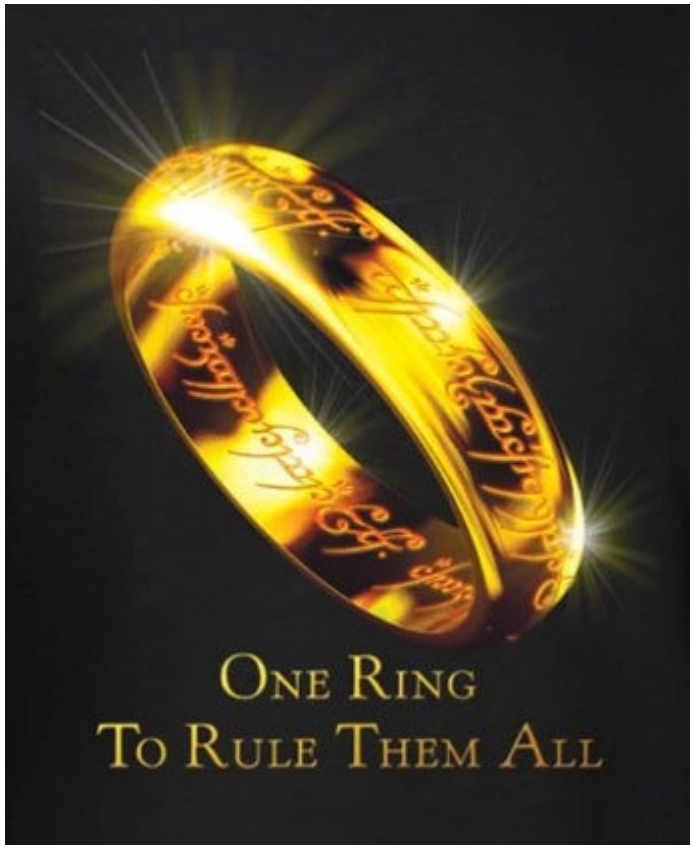
This claim implies we should get a certain distribution of statistics



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

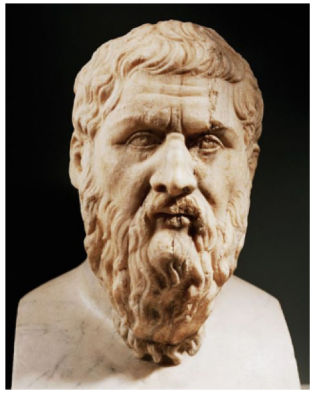
There is only one hypothesis test!





Just follow the 5 hypothesis tests steps!

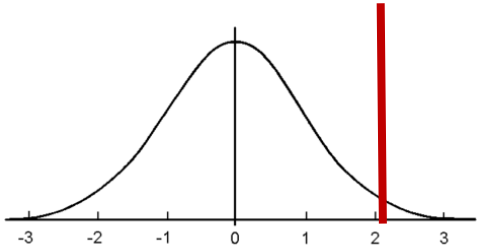
Five steps of hypothesis testing



1. State H_0 and H_A
   - Assume Gorgias (H_0) was right

2. Calculate the actual observed statistic

3. Create a distribution of what statistics would look like if Gorgias is right
   - Create the **null distribution** (that is consistent with H_0)



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

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?

	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	

Sudoku by field

1. State the null and alternative hypotheses!

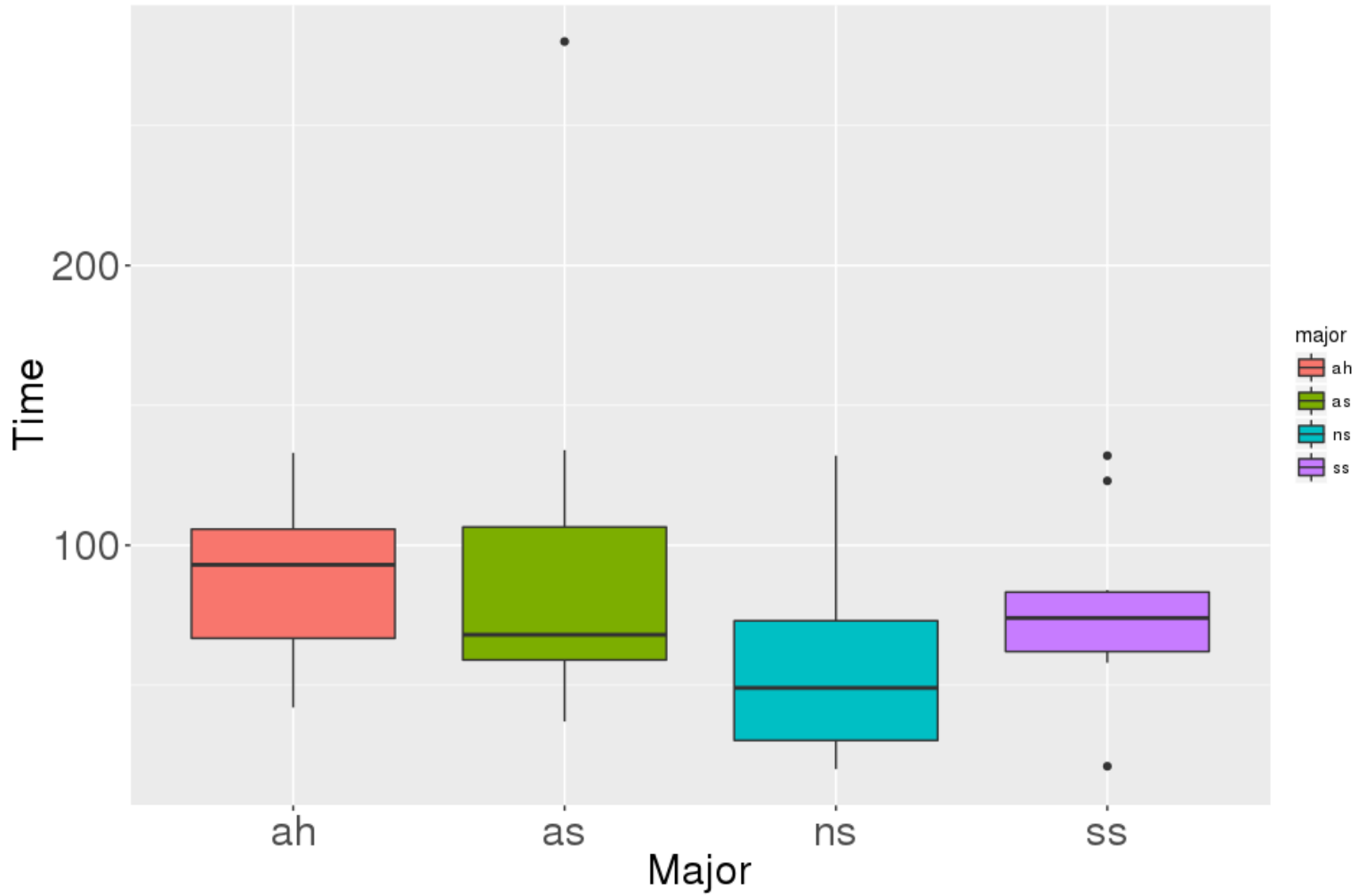
H_0: $\mu_{as} = \mu_{ns} = \mu_{ss} = \mu_{ah}$

H_A: $\mu_i \neq \mu_j$ 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!

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

$\mathbf{H_A}$: $\mu_i \neq \mu_j$ 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

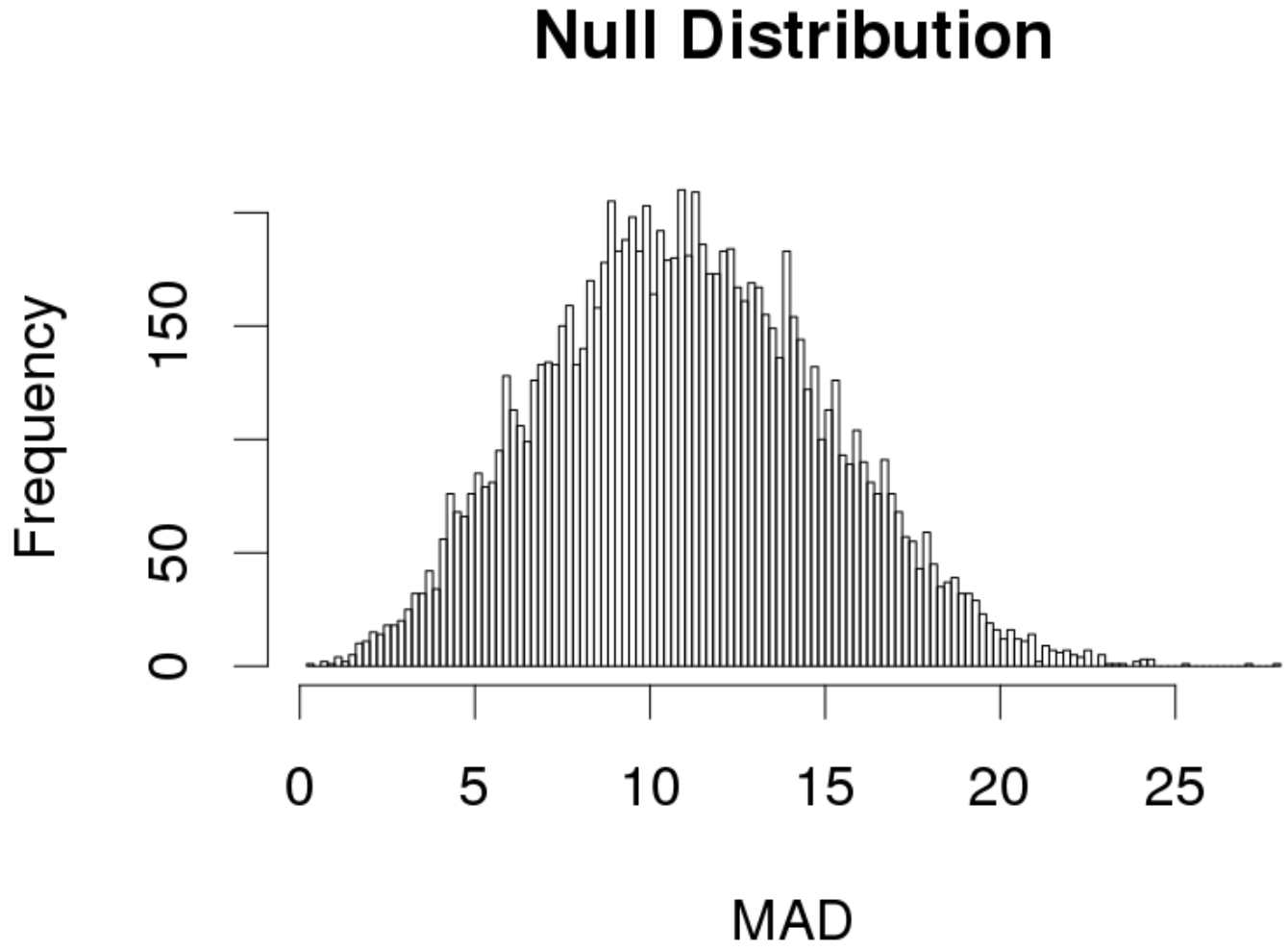
Mean absolute difference (MAD):

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

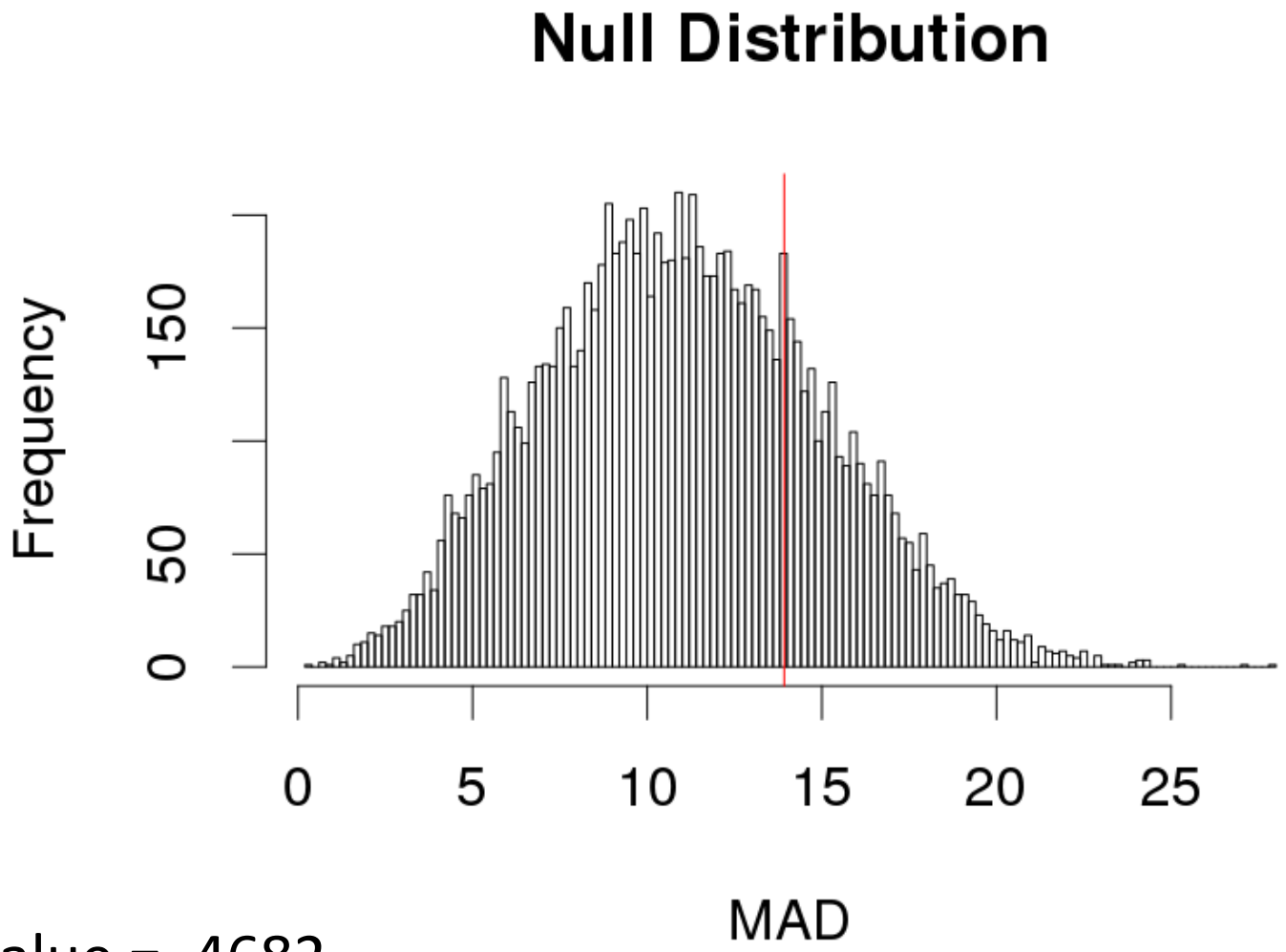
Observed statistic value = 13.92

How can we create the null distribution?

Null distribution



P-value



p-value = .4682

Conclusions?

Hypothesis tests for more than 2 means in R

Let’s try this analysis in R…

Either run:

```
reinstall_class_package()
```

Or use workspace 3

- Link is on Canvas

```
# get the data
library(ClassTools)
download_class_data("MajorPuzzle.txt")
sudoku_data <- read.table("MajorPuzzle.txt", header = TRUE)
```

Let’s try this analysis in R…

```
# Extract vectors from the data frame  (how do we do this?)
completion_time <- sudoku_data$time
major <- sudoku_data$major
```

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)
```

Can you visualize the data?

```
boxplot(completion_time ~ shuffled_majors)
```

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

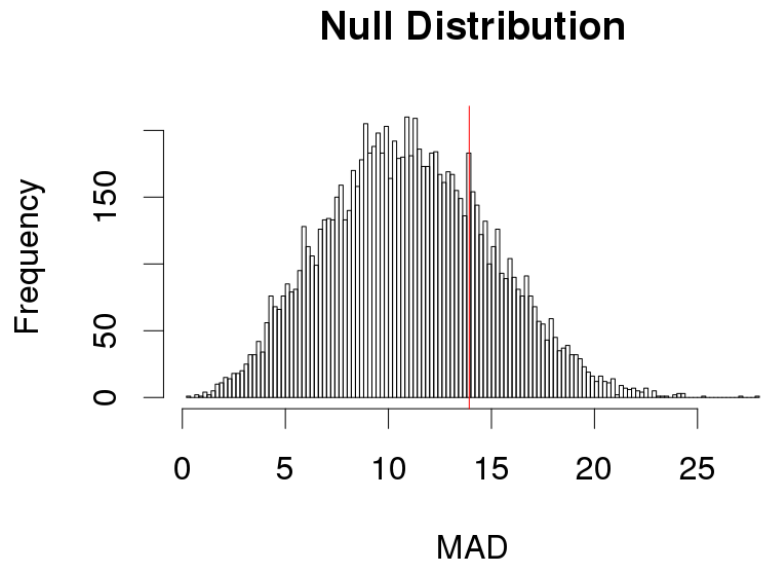
Q: How can we do this in R?

```
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")
```



Let’s try this analysis in R…

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

- A: We get the p-value

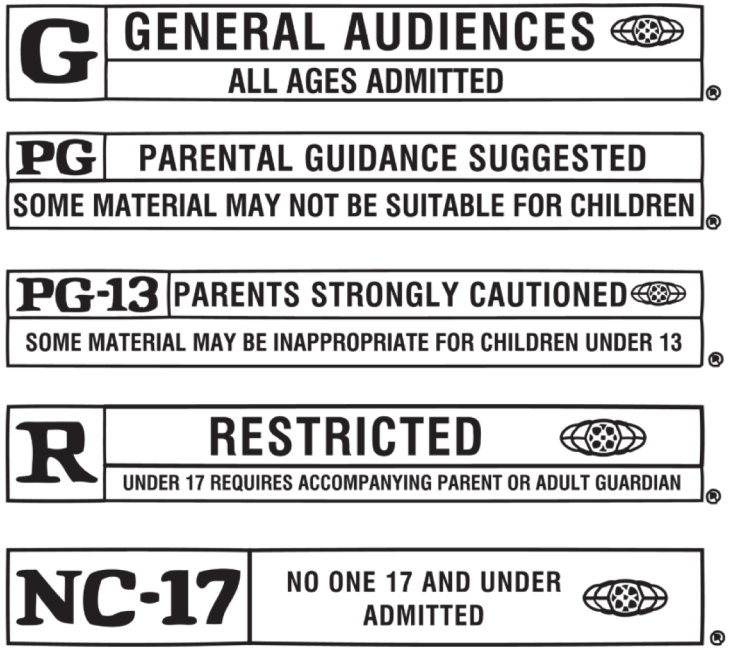
```
pnull(obs_stat, null_dist, lower.tail = FALSE)
```

Questions?

If so, come to online office hours!

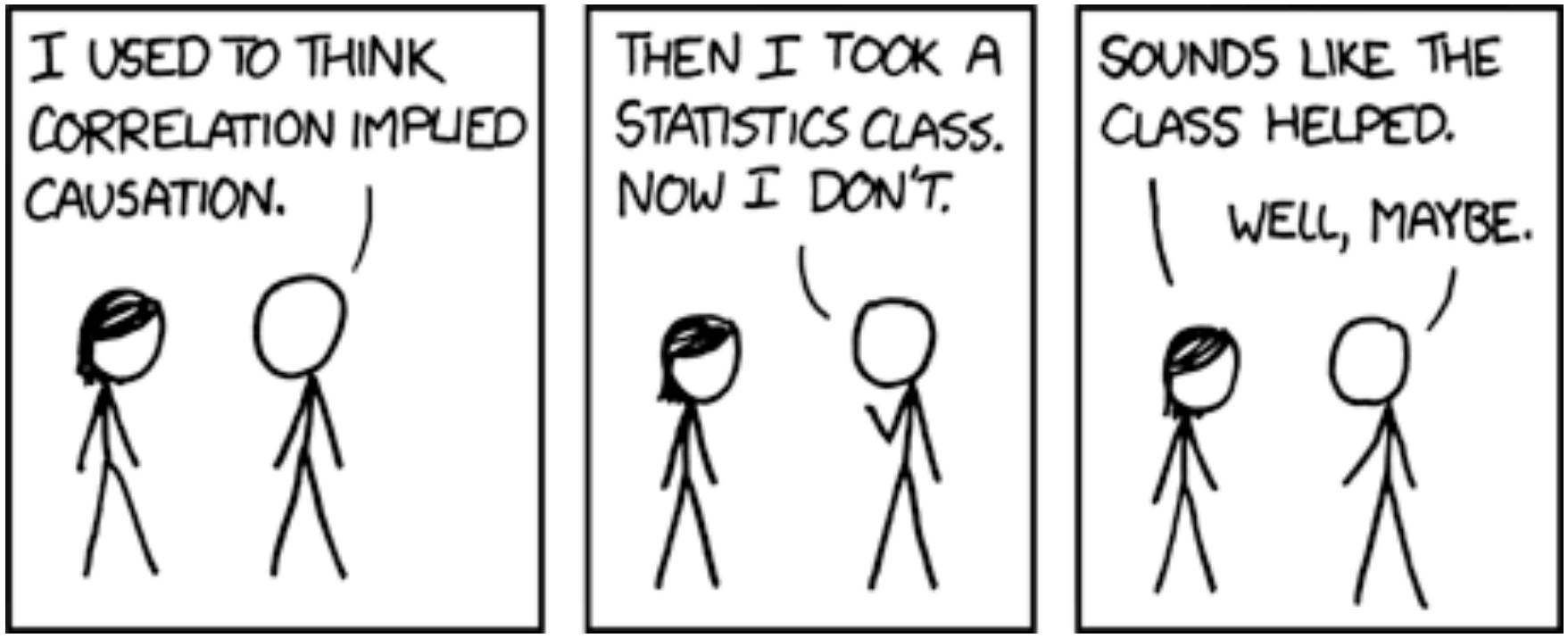
Homework 7

Rotten Tomatoes is a website that provides movie ratings and reviews



Question: Do critics' rate movies the same on average regardless of their MPAA ratings?

Hypothesis tests for correlation



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

Suppose we had some 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_class_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: Calculate the observed statistic in R and plot the 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

Step 4: Calculate the p-value

Step 5: Make a decision!

We will pick up from here next class…