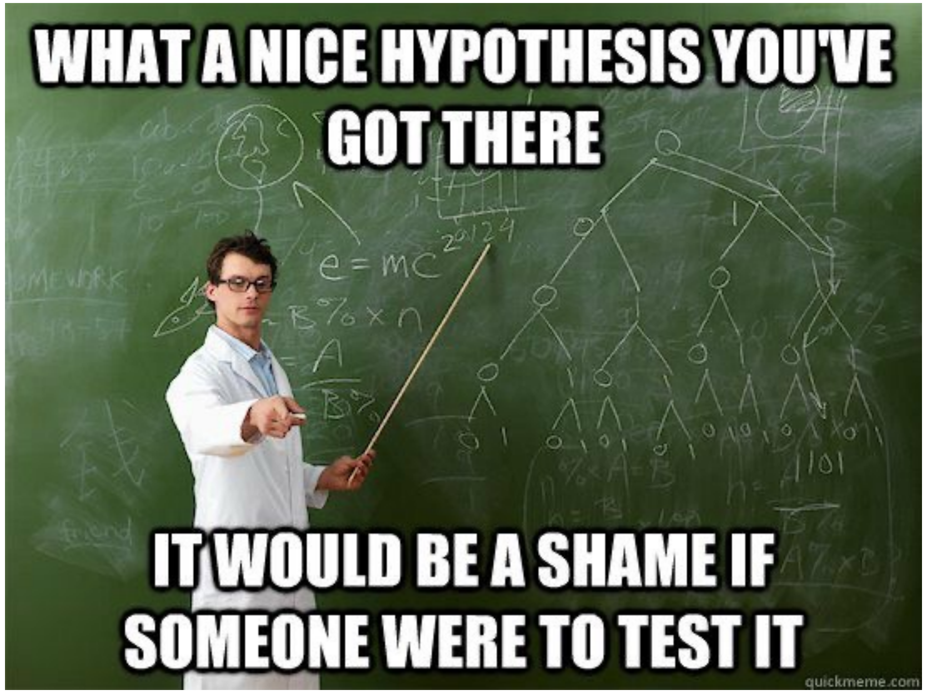
YData: Introduction to Data Science

Class 22: Confidence intervals

Overview

Hypothesis tests assessing causality

Confidence intervals



Project timeline

~~Tuesday, April 11th~~

- ~~Projects are due on Gradescope at 11pm on~~
- Also, email a pdf of your project to your peer reviewers
    - A list of whose paper you will review has been posted to Canvas

Wednesday, April 19th

- Jupyter notebook files with your reviews need to be sent to the authors and a pdf need to be submitted to Gradesscope
- A template for doing your review is available on Canvas

Sunday, April 30th

- Project is due on Gradescope
    - Add peer reviews to an Appendix of your project

Project peer review

A template for your project peer review has been posted

- import YData
- YData.download.download_class_file('reviewer_template.ipynb, 'homework')

Please review the projects by 11pm on Wednesday April 19th and:

- 1. Post a pdf of each of your reviews to Gradescope
- 2. Send a filled out Jupyter Notebook with your review to the project author
  - If you run into any logistic issues post to Ed and then ask our course manage Zihe (zihe.zheng@yale.edu)

In your final project, please add the three reviews in the Appendix section, and discuss how you addressed the reviewers' comments.

Also, homework 8 is due on Sunday April 16th

- Thanks to Rose, it is not too long

Review of Statistical Inference

Review: Statistical Inference

Statistical Inference: Making conclusions about a population based on data in a random sample

A parameter is number associated with the population

- e.g., population proportion π
- e.g., the proportion of voters who voted for Biden

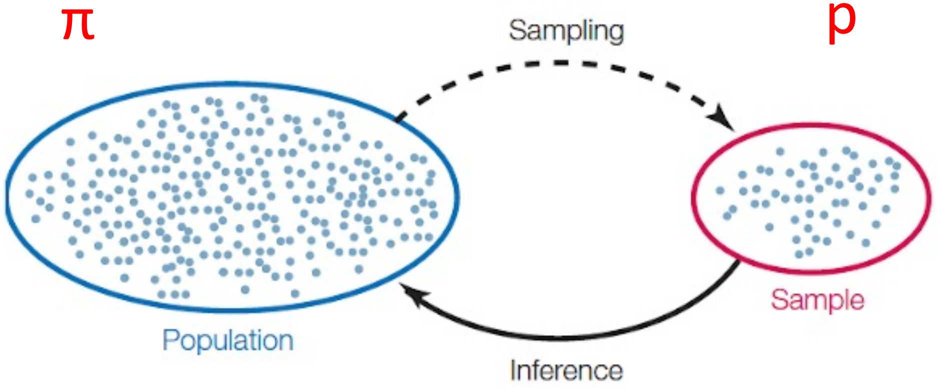
A **statistic** is number calculated from the sample

- e.g., sample proportion p̂
- e.g., the proportion of Biden's vote out of 1,000 people in our sample

A statistic can be used as an estimate of a parameter

parameter **statistic**

π p̂



	Sample Statistic	Population Parameter
Mean	$\bar{x}$	μ
Proportion	p̂	π

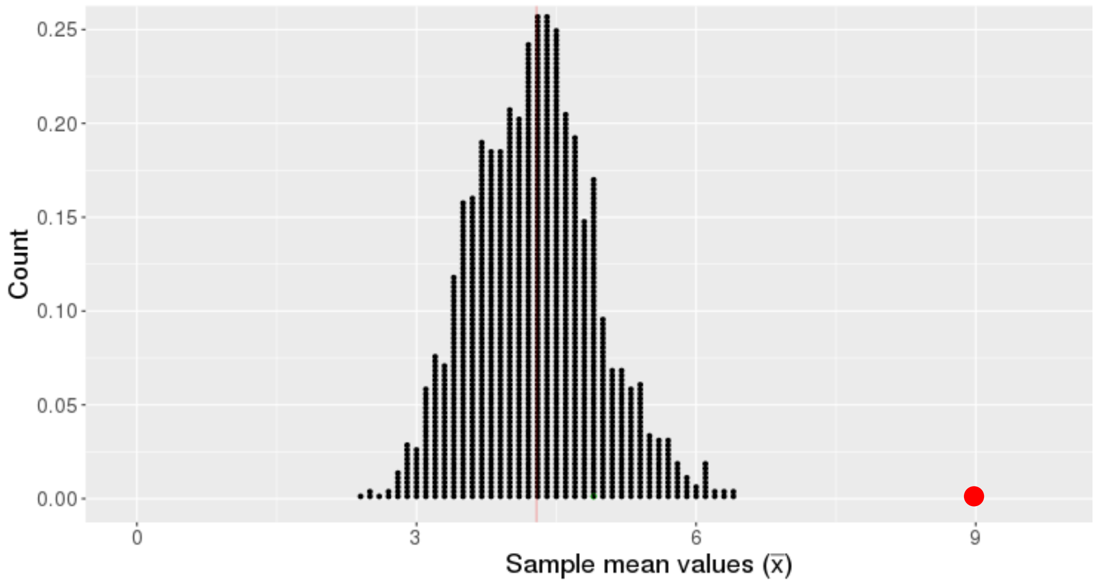
Hypothesis tests

Basic hypothesis test logic

We start with a claim about a population parameter

- E.g., μ = 4

This claim implies we should get a certain distribution of statistics

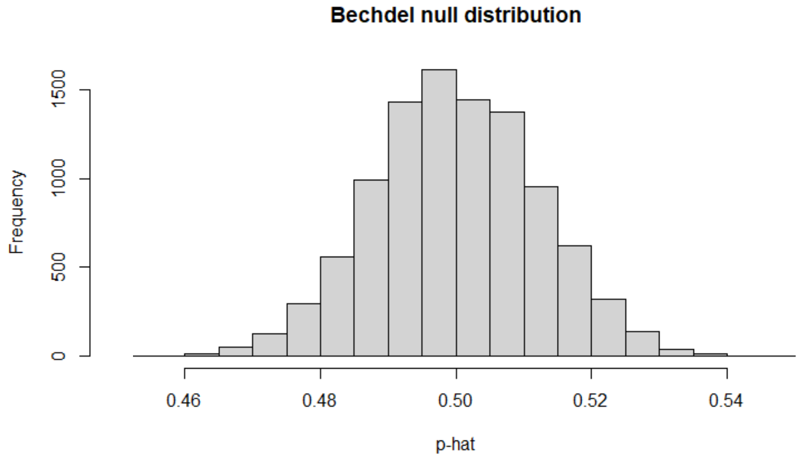


If our observed statistic is highly unlikely, we reject the claim

Null and Alternative hypotheses

Null hypothesis

- A hypothesis where ”nothing interesting” happened
  - E.g., our experiment failed
- We can simulate data under the assumptions of this model to get a "null distribution" of statistics



Alternative hypothesis

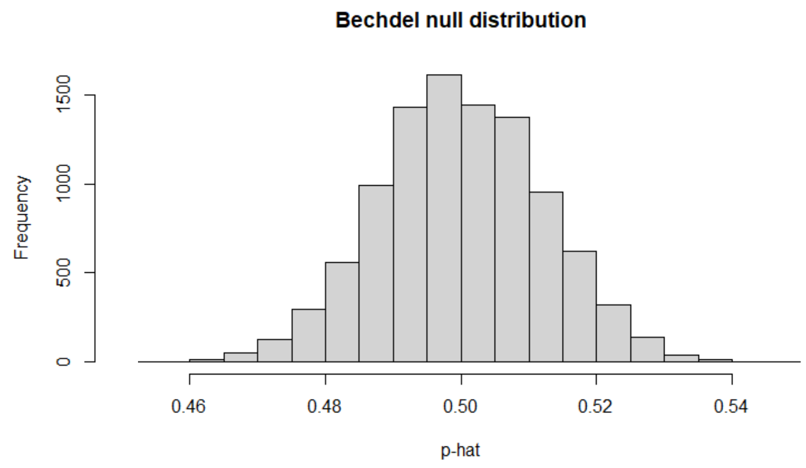
- The hypothesis we believe in (would like to see true)

A test statistic is the statistic we choose to simulate in order decide between the two hypotheses

Testing the null hypothesis

To resolve choice between null and alternative hypotheses:

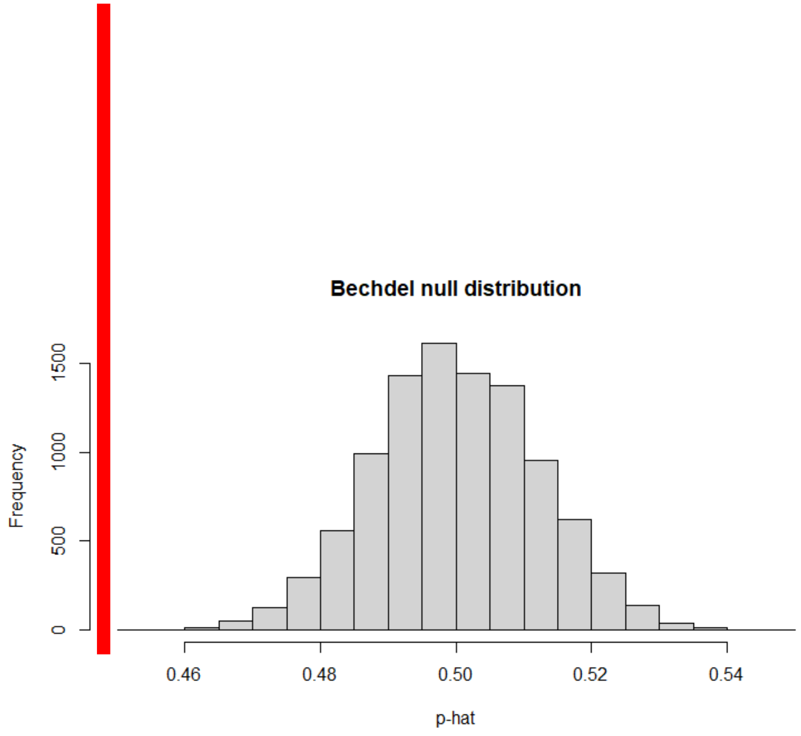
- We compare the observed test statistic to the statistic values in the null distribution
- If the observed statistic is not consistent with the null distribution, then we can reject the null hypothesis
  - And we accept the alternative hypothesis



The p-value

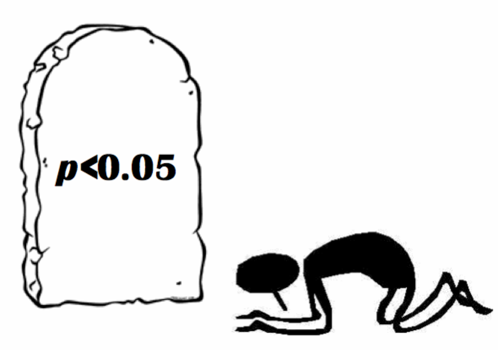
The p-value is the probability, that we get a statistic as or more extreme than the observed statistic from the null distribution

- P(Null_Stat ≥ obs_stat | H_0)



If the P-value is small, this is evidence against the null hypothesis and the results are often called "statistically significant"

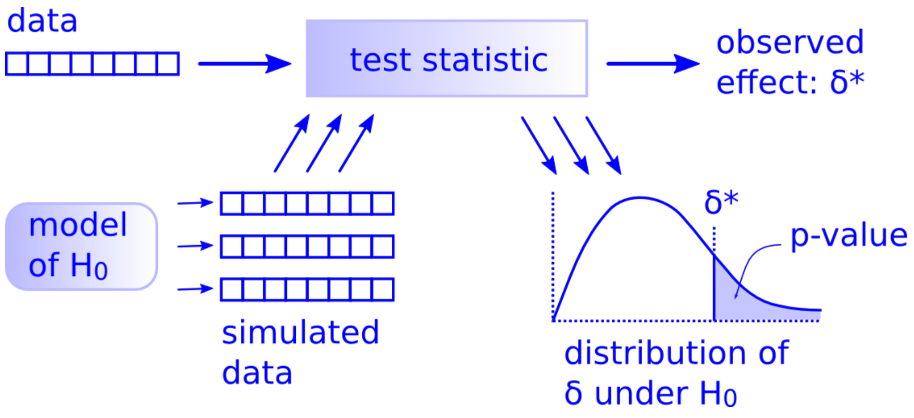
- Convention, p-value < 0.05

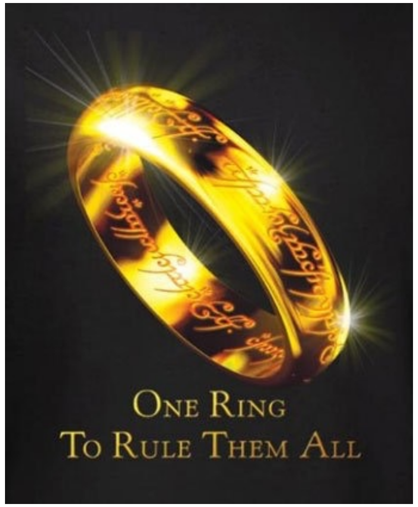


Steps needed to run a hypothesis test

To run a hypothesis test, we can use 5 steps:

1. State the null and alternative hypothesis
2. Calculate the observed statistic of interest
3. Create the null distribution
4. Calculate the p-value
5. Make a decision





Bechdel (hypothesis) test

1. State the null hypothesis and the alternative hypothesis

- 50% of the movies pass the Bechdel test: H_0: $\pi = 0.5$
- Less than 50% of movies pass the: H_A: $\pi < 0.5$

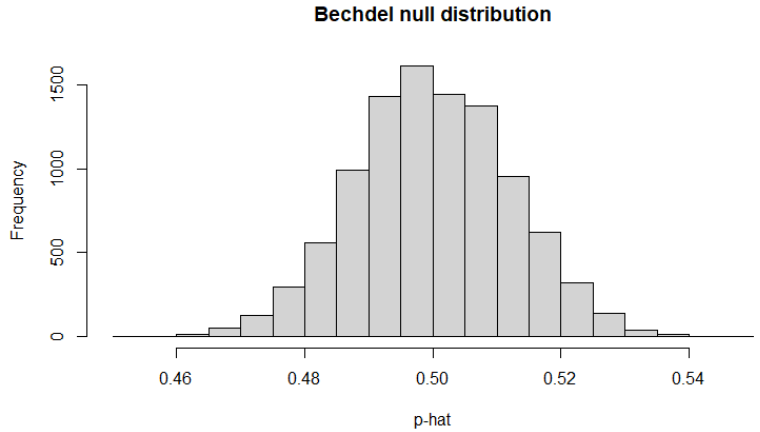
2. Calculate the observed statistic

- 803 out of 1794 movies passed the Bechdel test

$\hat{p} = .448$

3. Create a null distribution that is consistent with the null hypothesis

- i.e., the statistics we expect if 50% of the movies passed the Bechdel test



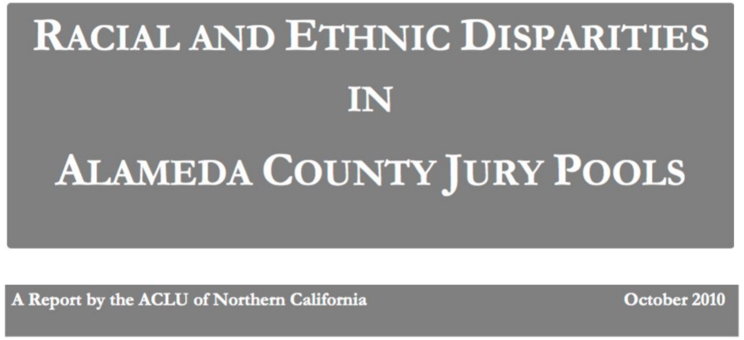
4. Examine how likely the observed statistic is to come from the null distribution

- What is the probability that only 803 of 1794 movies would pass the Bechdel test ($\hat{p}$ = .448) if the null hypothesis was true?
- i.e., what is the p-value?

5. Make a judgement

- A small p-value this means that $\pi = .5$ is unlikely, and so it is likely $\pi < .5$
- i.e., we say our results are 'statistically significant'

Jury selection in Alameda county



RACIAL AND ETHNIC DISPARITIES IN ALAMEDA COUNTY JURY POOLS

A Report by the ACLU of Northern California — October 2010

1. State the null hypothesis and the alternative hypothesis

- Jury panels match population demographics: H_0: $\pi_A = .15$, $\pi_L = 0.12$, etc.
- At least one ethnicity is not correctly represented: H_A: π_i differs from H_0

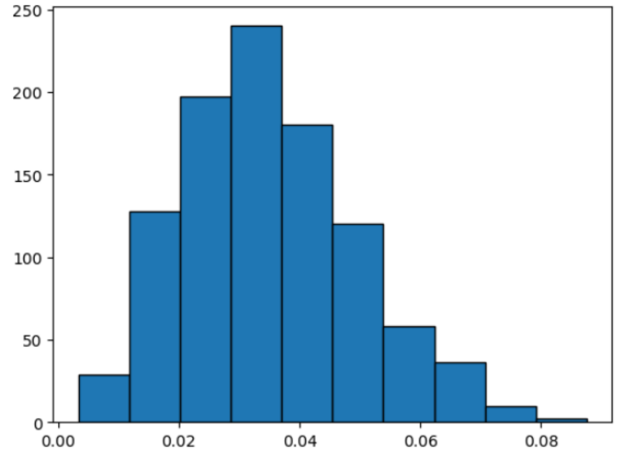
2. Calculate the observed statistic

$$TVD = \sum_{i=1}^{k} |\pi_i - \hat{p}_i|$$

TVD = .28

3. Create a null distribution that is consistent with the null hypothesis

- The TVD statistics we expect if the null hypothesis was true
- i.e., the TVD statistics we would expect if the sample demographics matched the population demographics



4. Examine how likely the observed statistic is to come from the null distribution

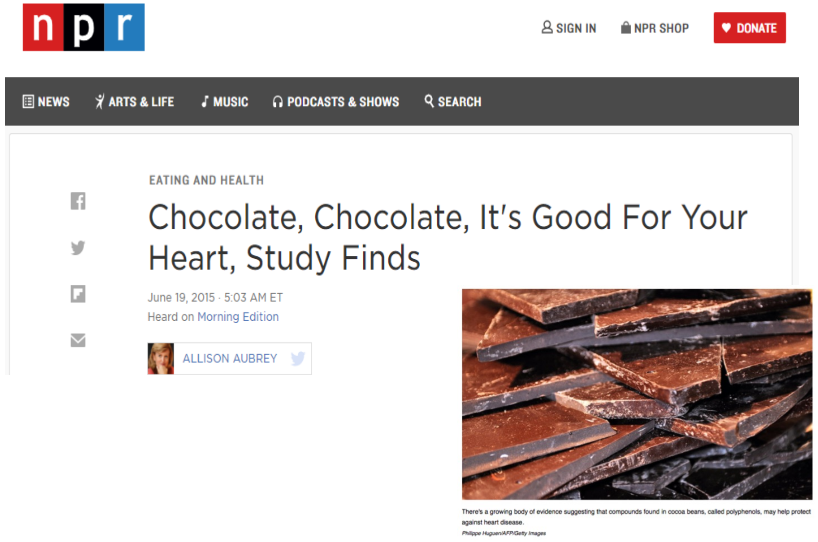
- What is the probability that we would get a TVD statistic larger than 0.28 if the null hypothesis was true?
- i.e., what is the p-value?

5. Make a judgement

- A small p-value this means that at least one demographic on juries does not match their representations in the population
- i.e., we say our results are ‘statistically significant’

Assessing causal relationships

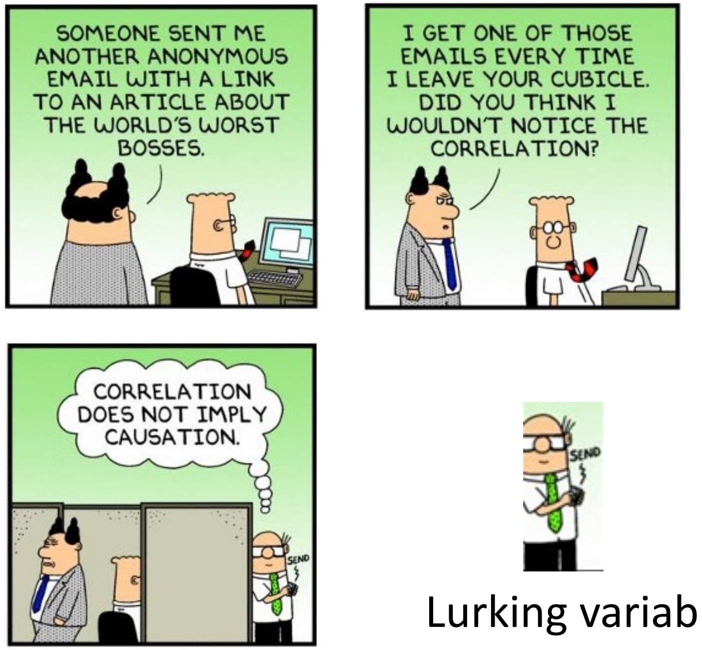
Review: Causality



Recall from class 2:

- **An association** is the presence of a reliable relationship between the treatments an outcome
- **A causal relationship** is when changing the value of a treatment variable influences the value outcome variable

- A **confounding variable** (also known as a **lurking variable**) is a third variable that is associated with both the treatment (explanatory) variable and the outcome (response) variable
  - A confounding variable can offer a plausible explanation for an association between the other two variables of interest



Lurking variable

Randomized Controlled Experiment

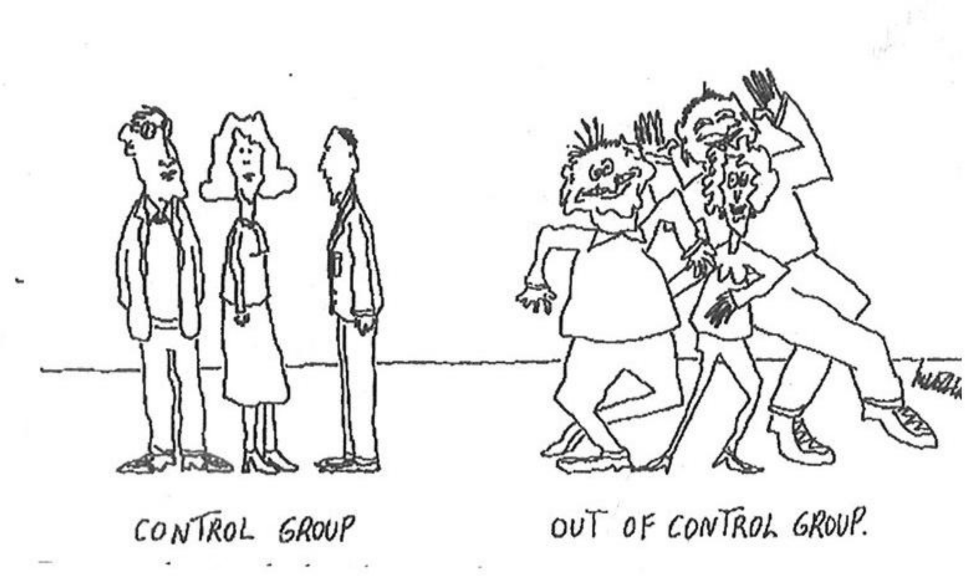
Sample A: control group

Sample B: treatment group

If members of the treatment and control groups are selected at random; this allows causal conclusions!

In particular, any difference in outcomes between the two groups could be due to:

- Chance
- The treatment



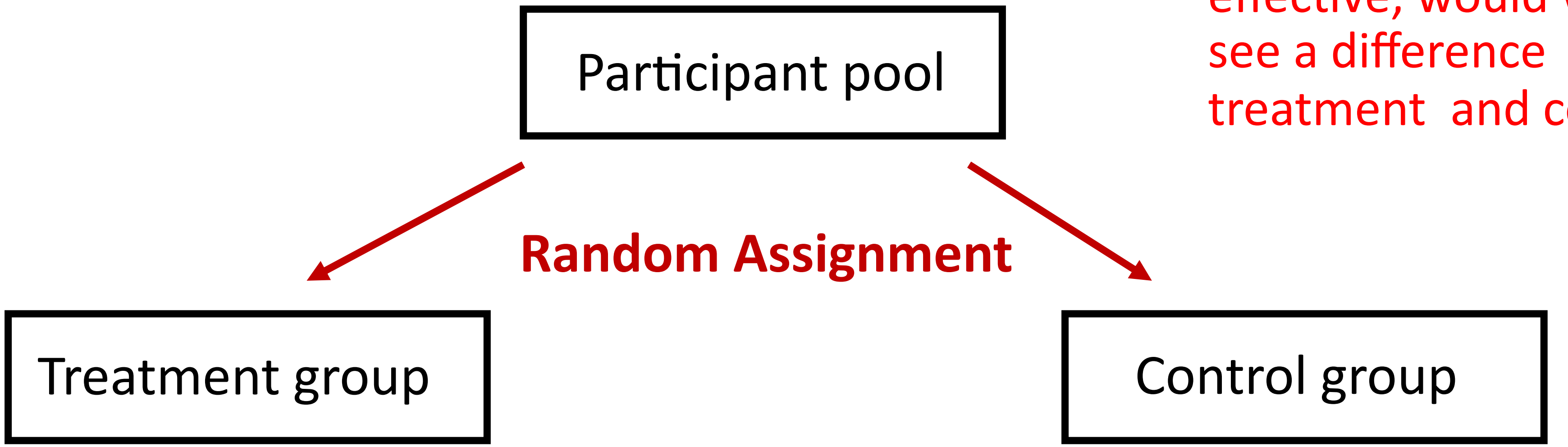
Randomly assigning participants to treatment and control groups allows us to separate what expected by chance and consequently what is due to the treatment

Randomized Controlled Experiment

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

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

Q: If the treatment was not effective, would we expect to see a difference between the treatment and control groups?



Case study

RCT to study Botulinum Toxin A (BTA) as a treatment to relieve chronic back pain

- 15 patients in the treatment group (received BTA)
- 16 in the control group (normal saline)

Trials were run double-blind: neither doctors nor patients knew which group they were in.

Results

- 2 patients in the control group had relief from pain (outcome=1)
- 9 patients in the treatment group had relief.

Can this difference be just due to chance?

Neurology®

May 22, 2001; 56 (10) ARTICLES

Botulinum toxin A and chronic low back pain

A randomized, double-blind study

Leslie Foster, Larry Clapp, Marleigh Erickson, Bahman Jabbari

First published May 22, 2001, DOI: https://doi.org/10.1212/WNL.56.10.1290

Step 1: The hypotheses

Null:

- BTA does not lead to an increase in pain relief
  - i.e., if many people were to get BTA and saline, the proportion of people who experienced pain relief would be the same in both groups.
  - H_0: $\pi_{treat} = \pi_{control}$

Alternative:

- BTA leads to an increase in pain relief
  - i.e., if many people were to get BTA and saline, the proportion of people who experienced pain relief would be higher for those who received BTA
  - H_A: $\pi_{treat} > \pi_{control}$

Neurology

May 22, 2001; 56 (10) ARTICLES

Botulinum toxin A and chronic low back pain

A randomized, double-blind study

Leslie Foster, Larry Clapp, Marleigh Erickson, Bahman Jabbari

First published May 22, 2001, DOI: https://doi.org/10.1212/WNL.56.10.1290

Step 2: The observed statistic

To calculate an observed statistic we need data:

Let's have our observed statistic mirror our hypotheses

- H_0: $\pi_{treat} - \pi_{control} = 0$

Observed statistic is: $\hat{p}_{treat} - \hat{p}_{control}$

$= 9/15 - 2/16$

$= 0.475$

	Group	Result
19	Treatment	1.0
7	Control	0.0
6	Control	0.0
26	Treatment	0.0
17	Treatment	1.0
9	Control	0.0
13	Control	0.0
3	Control	0.0
1	Control	1.0
30	Treatment	0.0
28	Treatment	0.0

3. Create the null distribution!

Treatment group

Control group

Reconstructed participant pool data under H_0

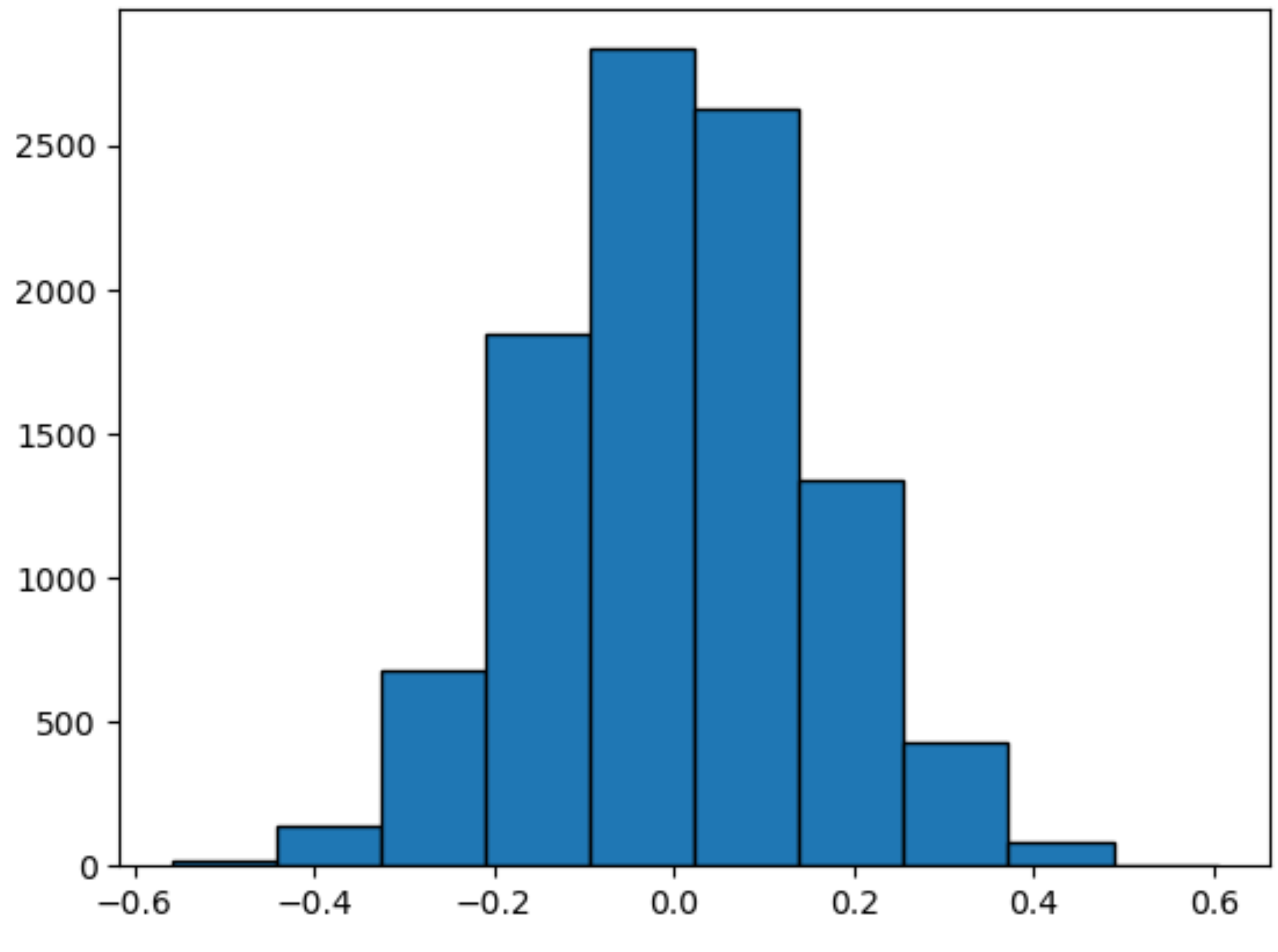
Shuffle data for random assignment consistent with H_0

Shuffled 'treatment group'

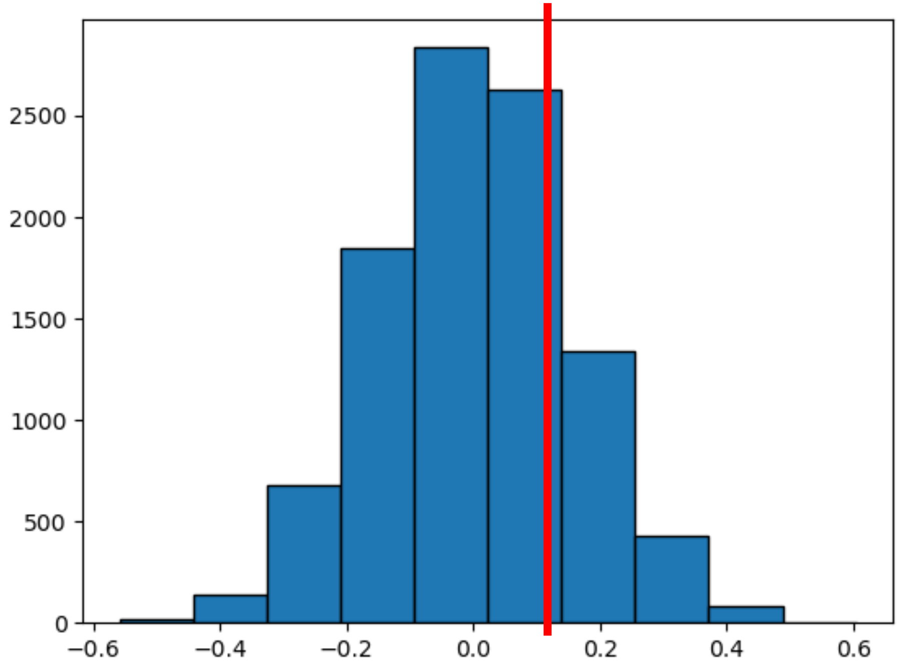
Shuffled 'control group'

One null distribution statistic: $\hat{p}_{Shuff\_Treatment} - \hat{p}_{Shuff\_control}$

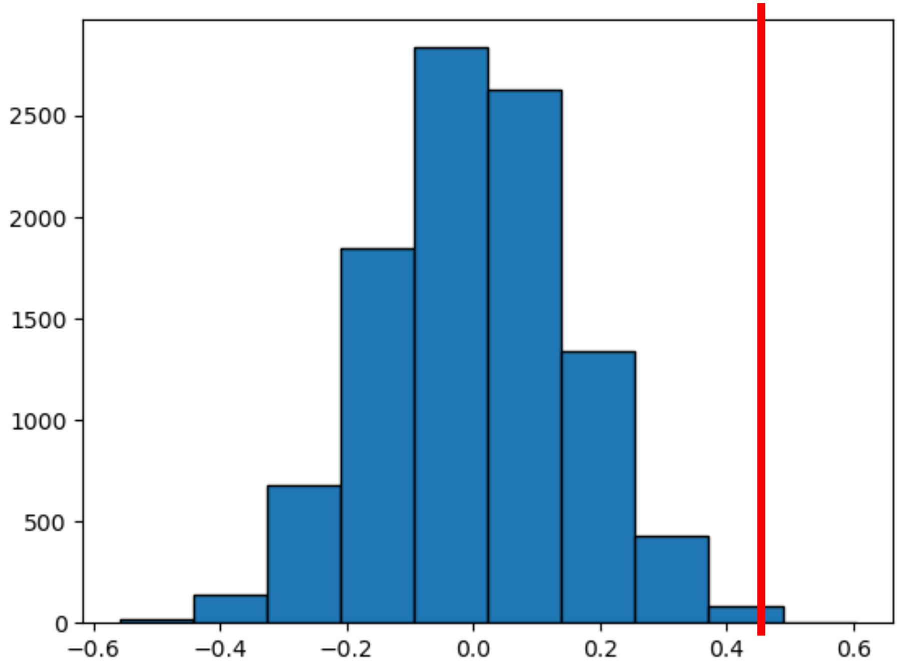
Step 3: Create a null distribution



Step 4: Calculate the p-value

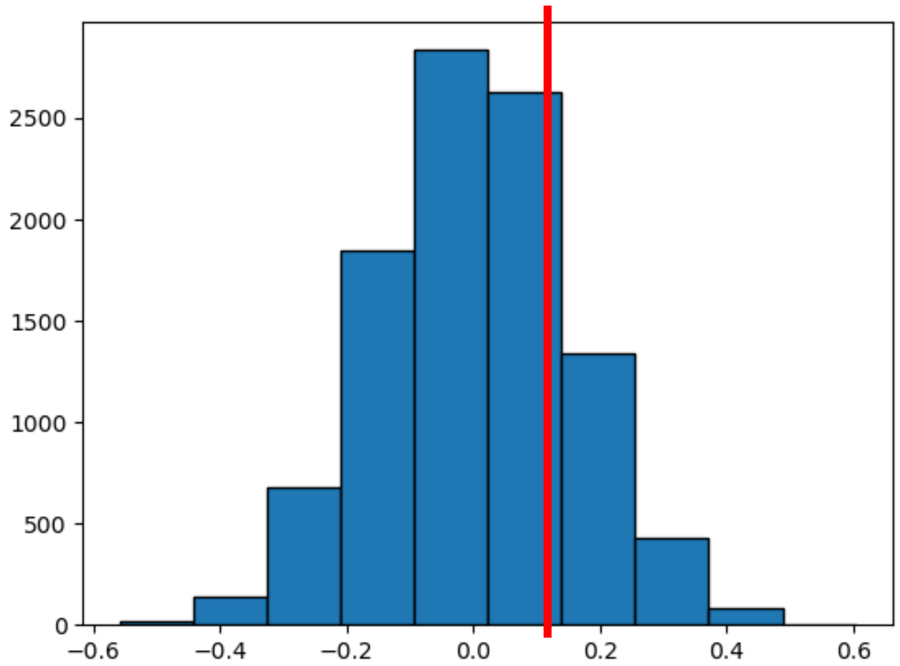


If $\hat{p}_{treat}$ - $\hat{p}_{control}$ = 0.1 what would the p-value be?

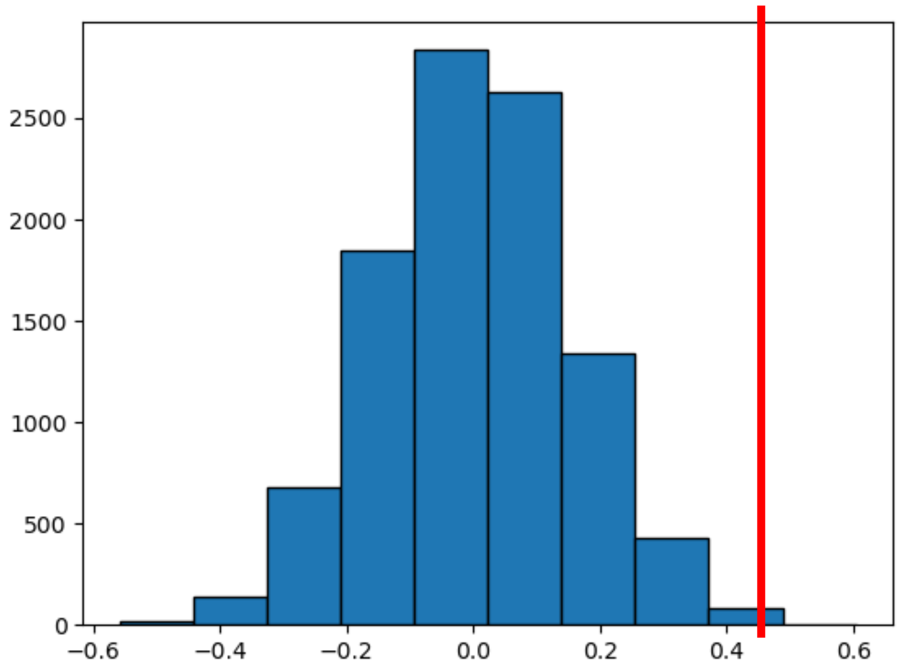


If $\hat{p}_{treat}$ - $\hat{p}_{control}$ = 0.5 what would the p-value be?

Step 5: Draw a conclusion



If the p-value was 0.19 what would we conclude?



If the p-value was 0.0007 what would we conclude?

Let's explore this in Jupyter!

Confidence intervals

Interval estimate based on a margin of error

Null hypothesis tests tell us if a particular parameter value is implausible

- E.g., in the Bechdel data we rejected $\pi = .5$

An **interval estimate** give a range of plausible values for a population parameter

Example: 42% of American approve of Biden's job performance, plus or minus 3%

How do we interpret this?

Says that the population parameter π lies somewhere between 39% to 45%

- i.e., if they sampled all voters the true population proportion would be likely be in this range

Confidence Intervals

A **confidence interval** is an interval <u>computed by a method </u>that will contain the *parameter* a specified percent of times

- i.e., if the estimation were repeated many times, the interval will have the parameter x% of the time

The **confidence level** is the percent of all intervals that contain the parameter

Think ring toss…

Parameter exists in the ideal world

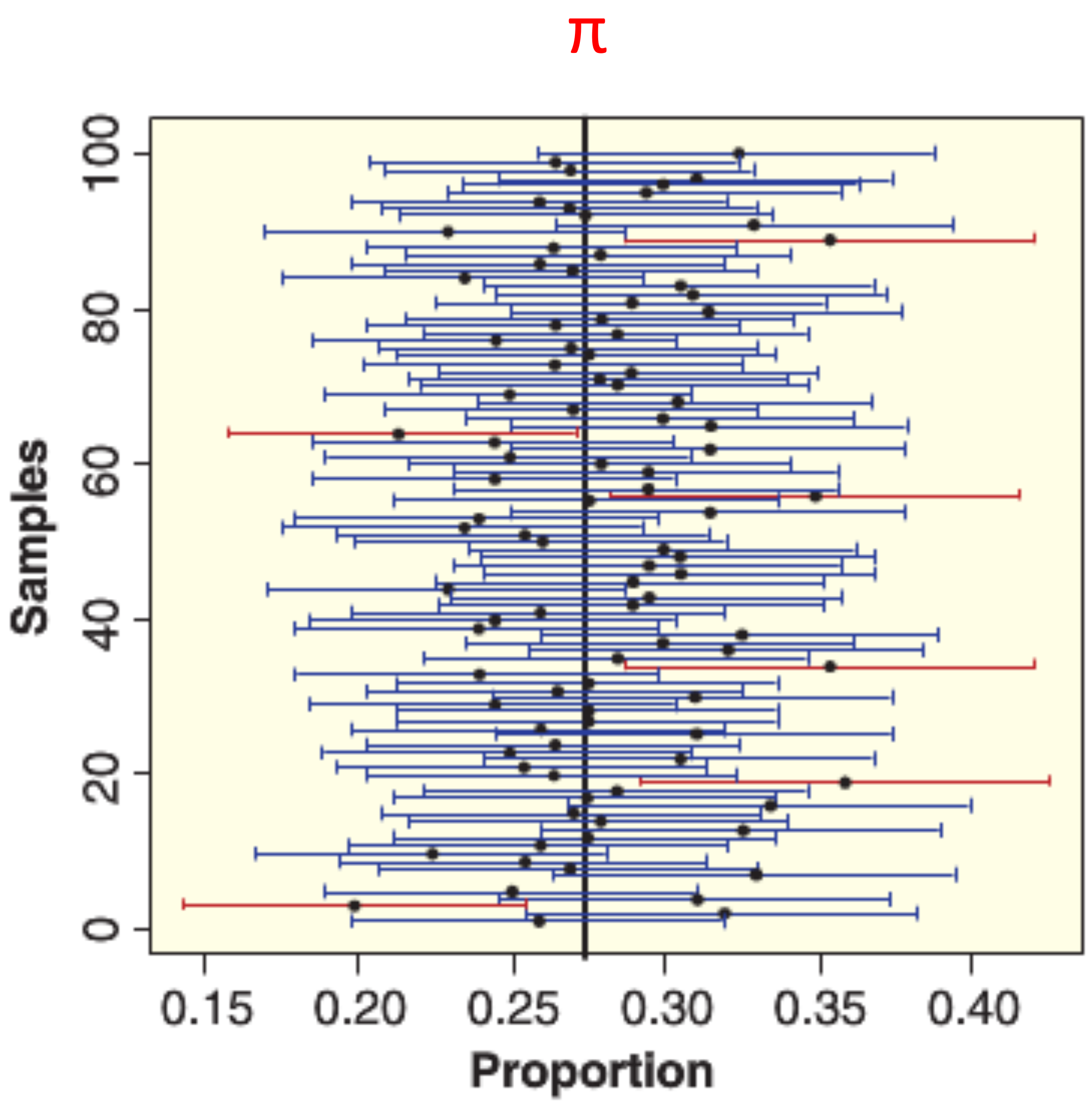
We toss intervals at it

95% of those intervals capture the parameter

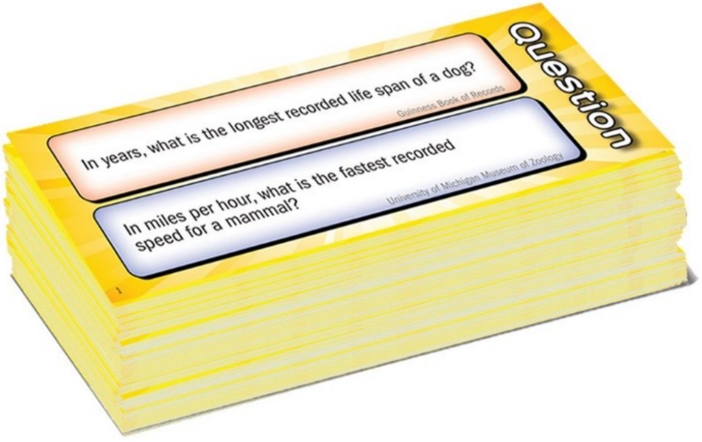
Confidence Intervals

For a **confidence level** of 95%...

95% of the **confidence intervals** will have the parameter in them



Wits and Wagers:
90% confidence interval estimator



I will ask 10 questions that have numeric answers

Please come up with a range of values that contains the true value in it for 9 out of the 10 questions

- i.e., be a 90% confidence interval estimator

Wits and Wagers…

Question 1: What is the diameter of the moon (in miles)?

Question 2: Formula Rossa in Abu Dhabi is the world's fastest roller coaster. What is its top speed in miles per hour (mph)?

Question 3: In what year did Alexander Graham Bell receive a patent for the invention of the telephone?

Wits and Wagers…

Question 4: How much does the average dog owner spend on dog food per year?

Question 5: In pounds, how heavy was the heaviest sumo wrestler?

Question 6: How many McDonalds Restaurants are there in the UK?

Question 7: How many songs did Elvis Presley have on the Billboard hot 100 chart?

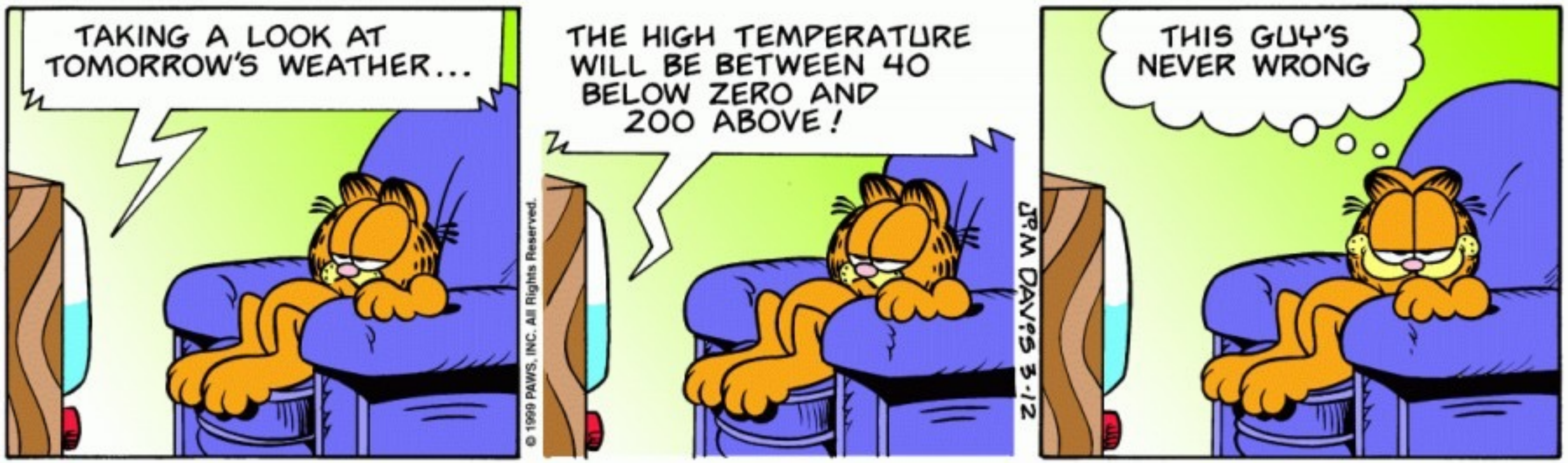
Wits and Wagers…

Question 8: How many verses does the Greek national anthem have?

Question 9: Including the antenna on the top, how many meters tall is the Eiffel Tower?

Question 10: What was the price of the first Ford Model T car?

Tradeoff between interval size and confidence level



There is a <u>tradeoff</u> between the **confidence level** (percent of times we capture the parameter) and the **confidence interval size**

Using hypothesis tests to construct confidence intervals

Constructing confidence intervals

There are several methods that can be used to construct confidence intervals including

- ”Parametric methods” that use probability functions
  - E.g., confidence intervals based on the normal distribution
- A “bootstrap method” where data is resampled from our original sample to approximate a sampling distribution

To learn more about these methods, take Introductory Statistics!

Constructing confidence intervals

We are going to use a less conventional method to get confidence intervals based on the relationship between confidence intervals and hypothesis tests

- The method we will discuss is valid, but can be more computationally expensive than other methods

What we will do is to run a series of hypothesis test with different null hypothesis parameter values

Our confidence interval will be all parameters values where we **fail to reject** the null hypothesis

$H_0: \pi = \pi_0$

Failure to reject $\pi = \pi_0$

means π_0 is plausible

Motivation: Bechdel Confidence Interval

From running a hypothesis test on the Bechdel data, we saw that H_0: $\pi = .5$ is unlikely

- i.e., it was not plausible that 50% of movies pass the Bechdel test

But what is a reasonable range of values for the population proportion of movies that pass the Bechdel test?

Let's create a confidence interval for H_0: $\pi_{Bechdel}$ to find out!

Let's explore this in Jupyter!