Hypothesis tests

Overview

Quick review of sampling distributions, confidence intervals and the bootstrap

Hypothesis tests for a single proportion

- Framework/terminology for hypothesis testing
- Hypothesis tests for a single proportion using randomization in R

If there is time: Hypothesis tests for two means

- Randomization tests for comparing two means in R

Announcements

Homework 1 feedback has been posted

- Any regrade requests need to be made within a week
  - Official grades are on Gradescope
- Future homework:
  - Be sure to mark pages for questions on Gradescope

How would describe the pace of the class this past week?

Way too slow		**0 %**
Too slow	4 respondents	5 %
About right	66 respondents	80 %
Too fast	10 respondents	12 %
Way too fast	2 respondents	2 %

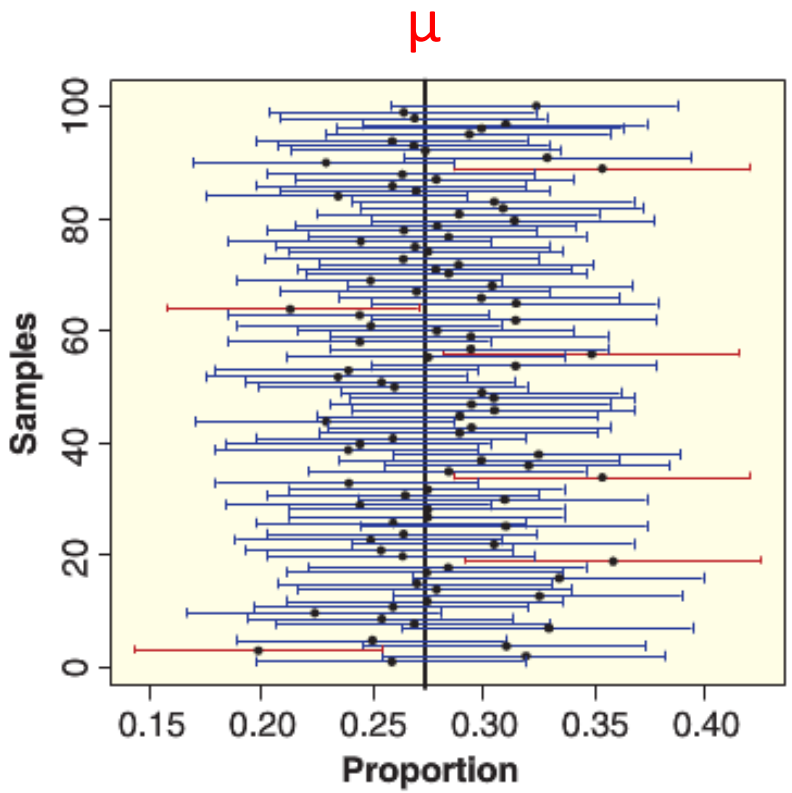
Review: Creating confidence intervals

What are Confidence intervals?

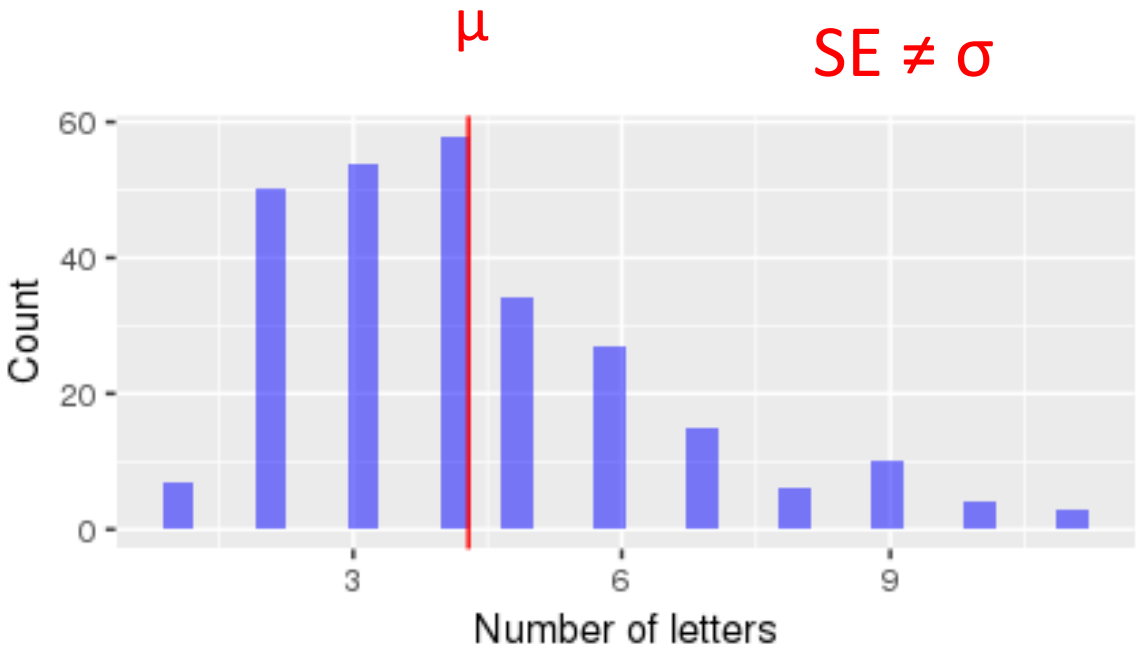
- Range of plausible values that capture the parameter a fixed % of the time

How can we create confidence intervals?

- Use the bootstrap (to estimate the SE)
- Use formulas (to estimate the SE)



Review of sampling distribution



10, 3, 3, 3, 4,
3, 2, 6, 10, 5

$\bar{x} = 5$

SE ≠ s_1

2, 6, 2, 6, 6,
2, 5, 3, 2, 9

$\bar{x} = 4.3$

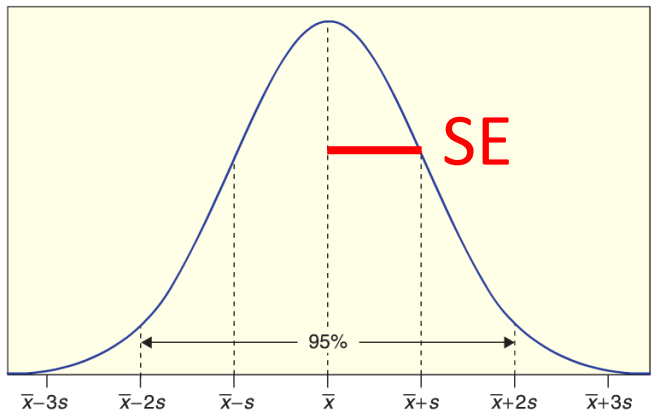
SE ≠ s_2

3, 9, 3, 4, 4,
3, 6, 6, 2, 2

$\bar{x} = 4.2$

SE ≠ s_3

The standard deviation of a sampling distribution is called the standard error (SE)

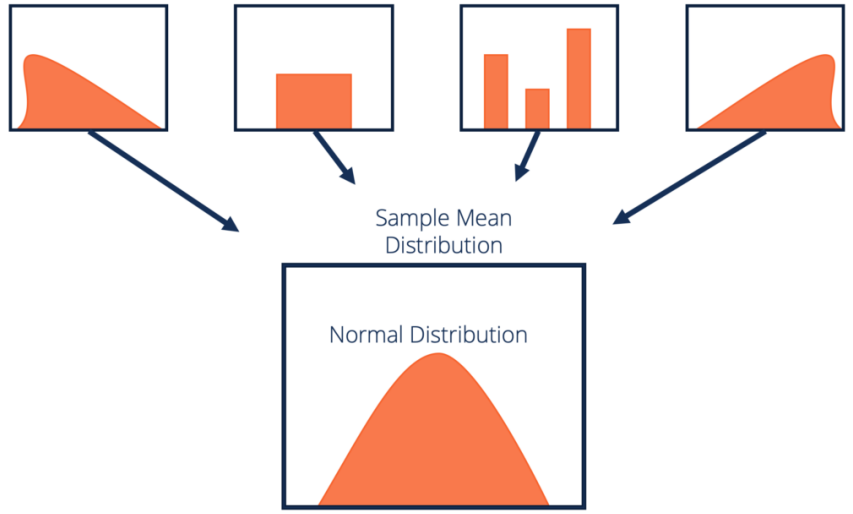


Sampling distribution!

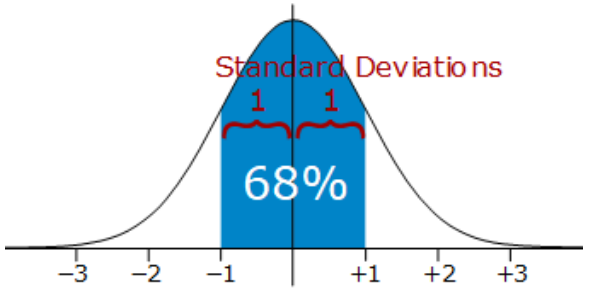
math "shortcut": $SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

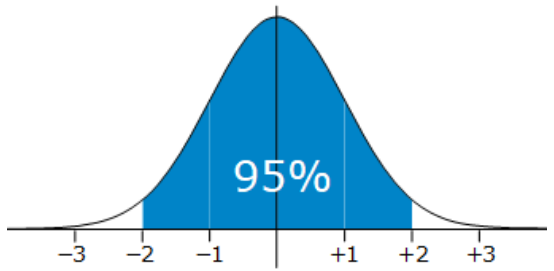
The central limit theorem

The **central limit theorem** establishes that when independent random variables are summed, the resulting statistic (sampling distribution) approaches a normal distribution as the sample size n increases.



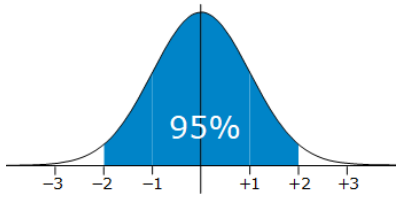
All these statistics, $\hat{p}$, $\bar{x}$, s, r, b, are sums of random data so their sampling distributions are normal





We can construct 95% confidence intervals using:

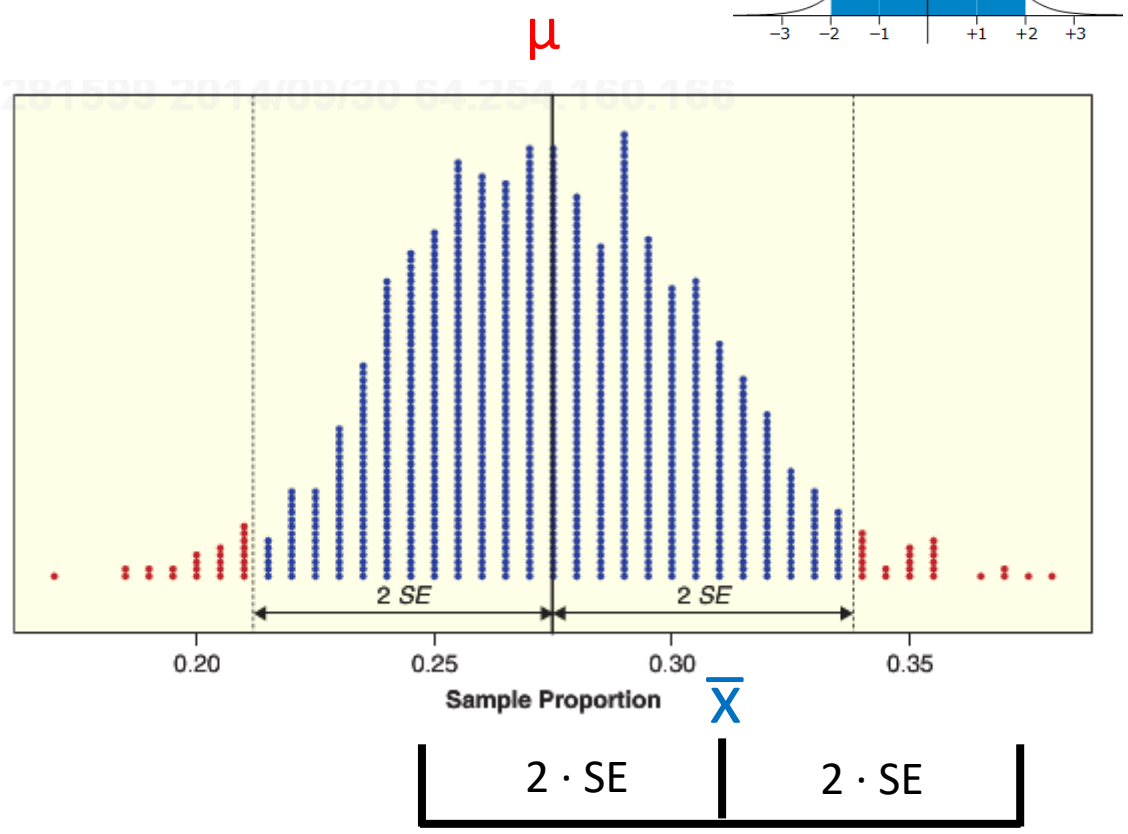
$$CI_{95} = \text{stat} \pm 2 \cdot SE$$



For example, a 95% confidence interval for the mean μ is:

$$\bar{x} \pm 2 \cdot SE$$

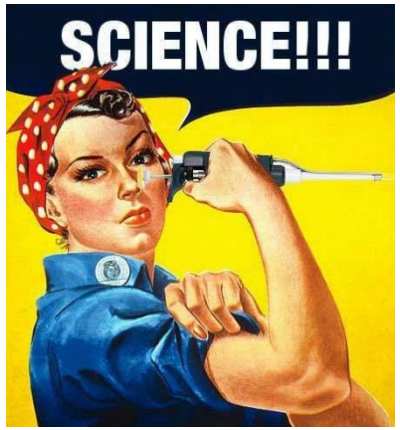
Why does this work?



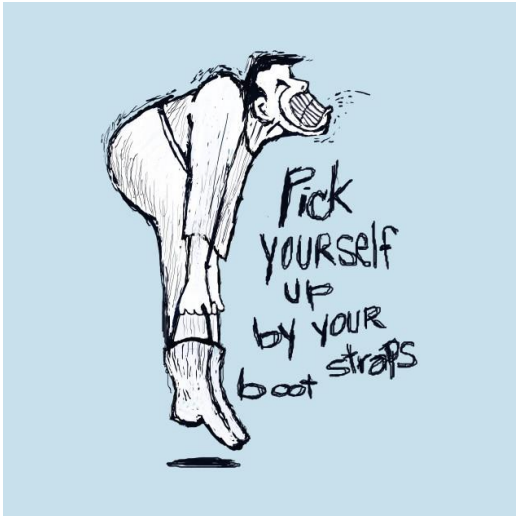
Confidence interval

Using the bootstrap to estimate SE

We can't calculate the sampling distribution by repeatedly sampling from the population ☹



Instead we nee do pick ourselves up from the bootstraps

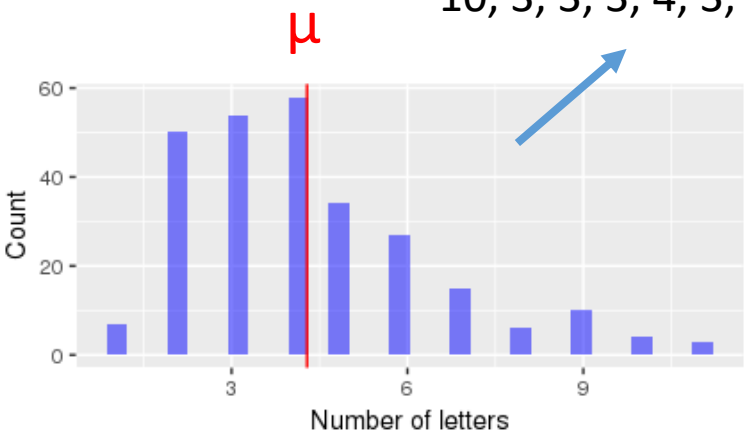


1. Estimate SE with $\hat{SE}$ from a single sample
2. Then use $\bar{x} \pm 2 \cdot \hat{SE}$ to get the 95% CI

Bootstrap distribution illustration

The sample (n = 10)

10, 3, 3, 3, 4, 3, 2, 6, 4, 5



3, 3, 3, 5, 3,
4, 5, 2, 2, 10

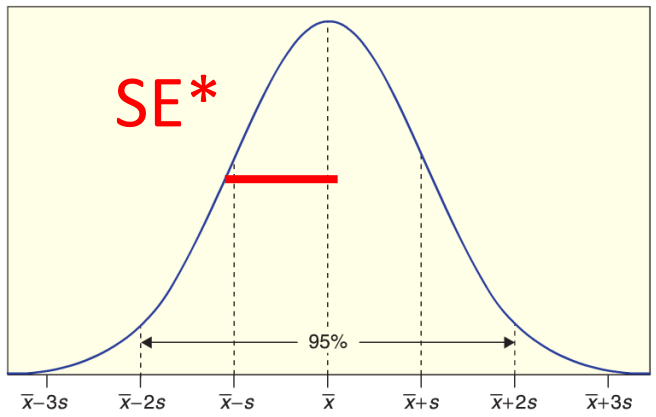
$\bar{x}^* = 4$

3, 3, 2, 3, 6,
4, 6, 5, 3, 6

$\bar{x}^* = 4.1$

5, 3, 2, 3, 3,
3, 10, 3, 4, 3

$\bar{x}^* = 3.9$



Bootstrap distribution!

Notice there is no 9's in the bootstrap samples

95% Confidence Intervals

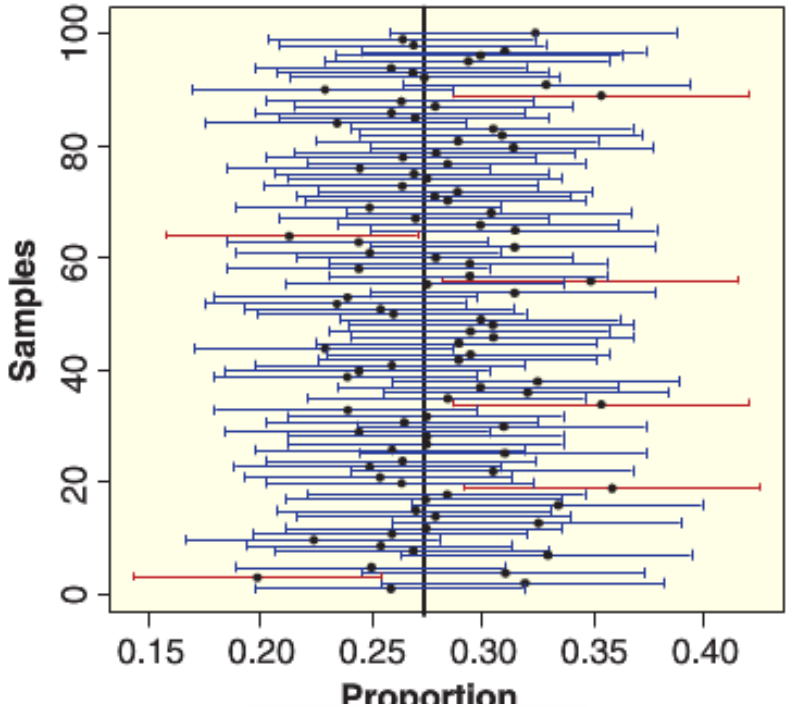
We can estimate a 95% confidence interval using: (provided the bootstrap distribution is reasonably normal)

$$Statistic \pm 2 \cdot SE^*$$

Where SE* is the standard error estimated using the bootstrap

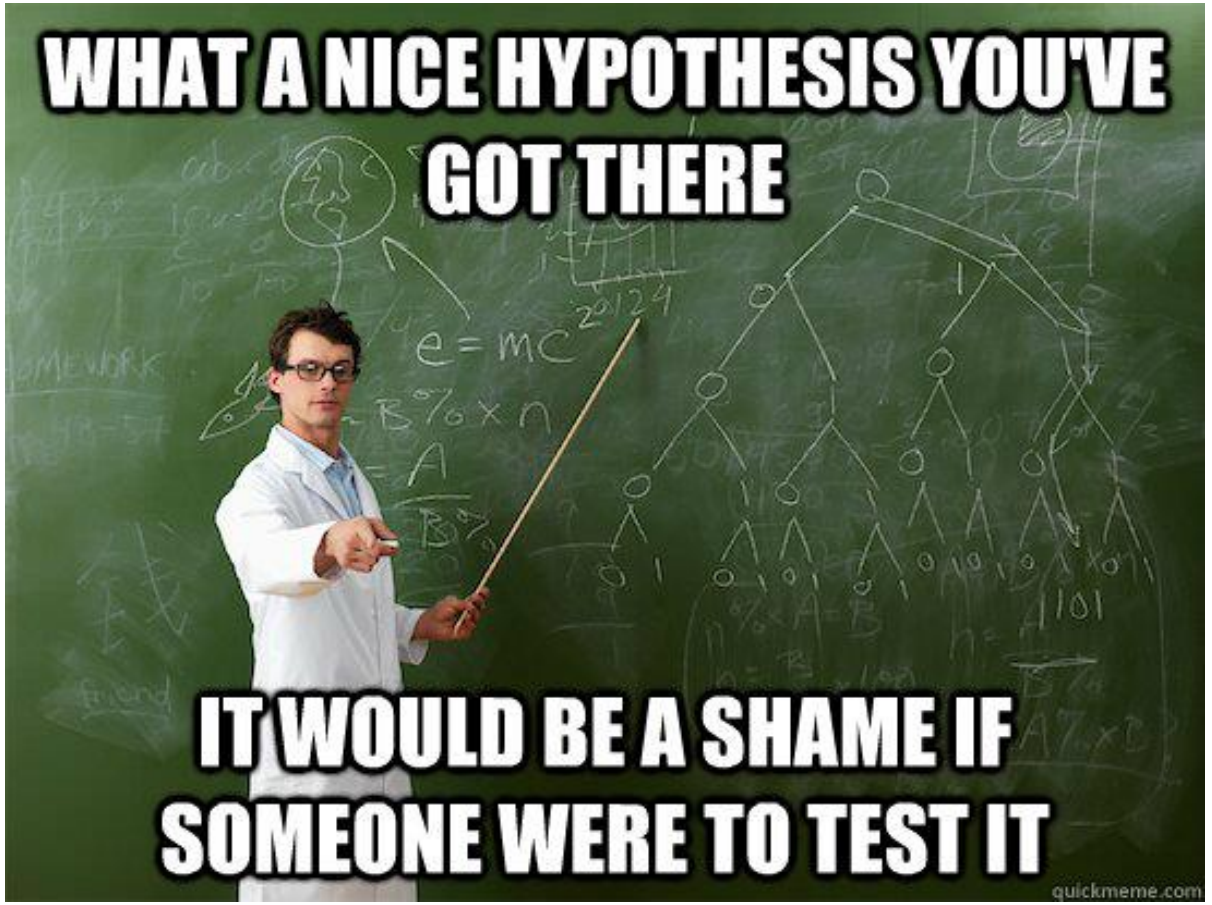
Q: Why use the bootstrap instead of…?

$$CI_{95} = \bar{x} \pm 2 \cdot \frac{s}{\sqrt{n}}$$




Questions?

Hypothesis tests



Overview

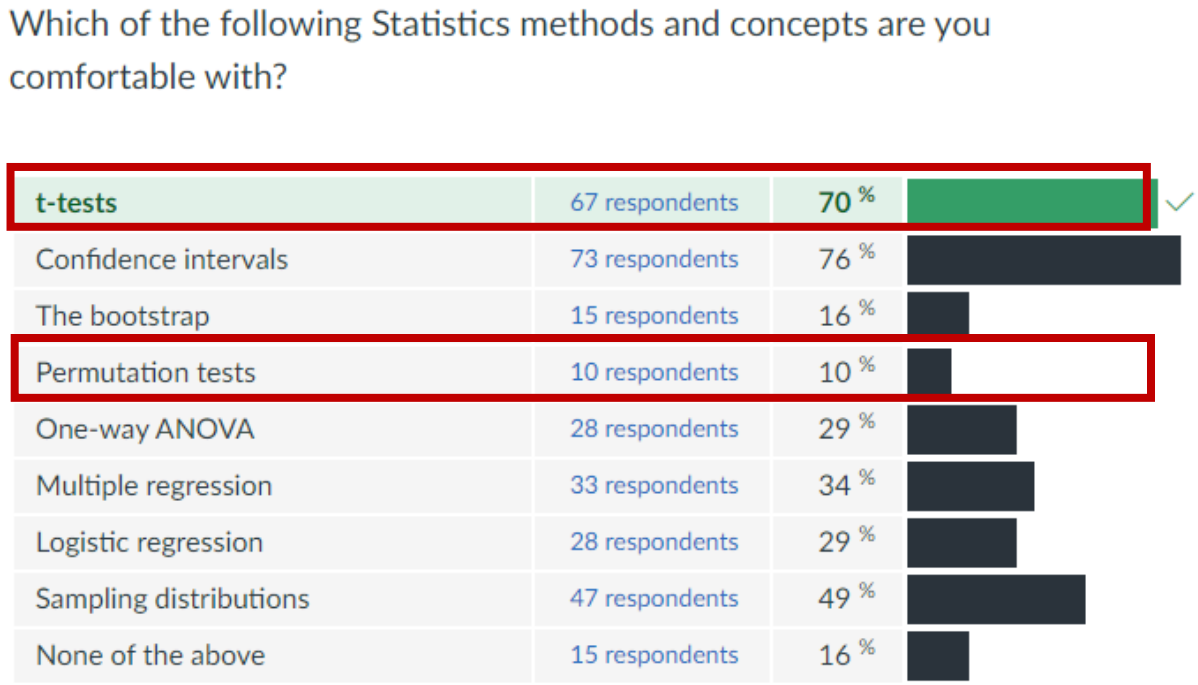
Assuming you are familiar with hypothesis tests from Intro Stats

- Particularly parametric hypothesis tests, such as the t-test

Quick review concepts of hypothesis test

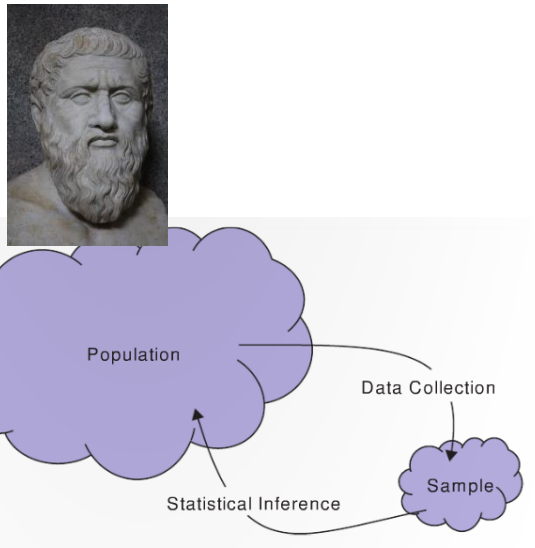
Introduce computational methods for hypothesis tests that use randomization

- These methods make fewer “assumptions” than parametric methods, so they can potentially work in more situations



Statistical tests (hypothesis test)

A **statistical test** uses data from a sample to assess a claim about a population



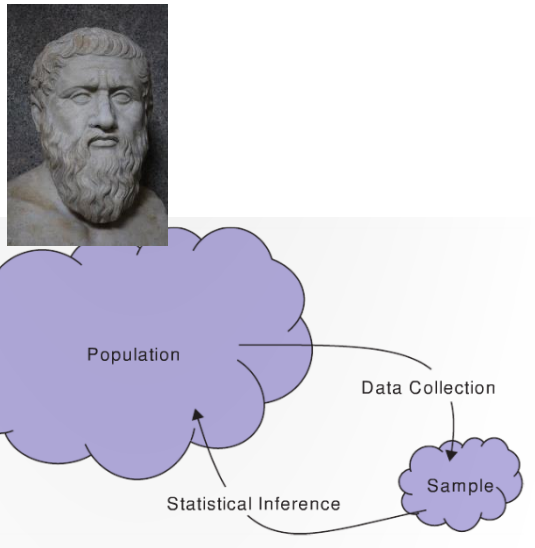
<u>Example 1</u>: we might make the claim that Biden's approval rating for all US citizens is 43%

How can we write this using symbols?

- $\pi = .43$

Statistical tests (hypothesis test)

A **statistical test** uses data from a sample to assess a claim about a population



<u>Example 2</u>: we might make the claim that the average height of a baseball player is 72 inches

How can we write this using symbols?

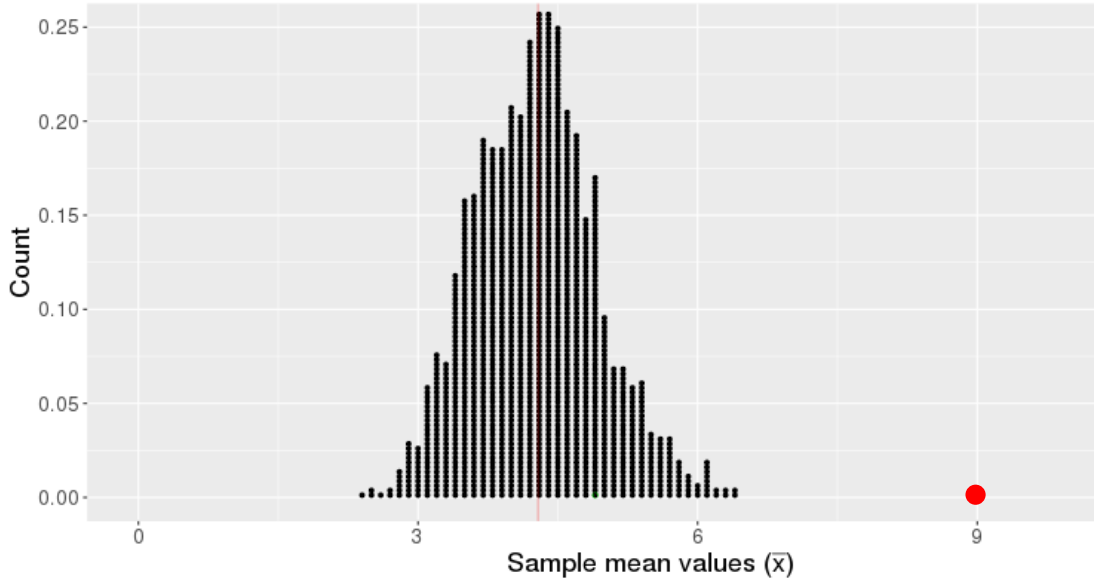
- $\mu = 72$

Basic hypothesis test logic

We start with a claim about a population parameter

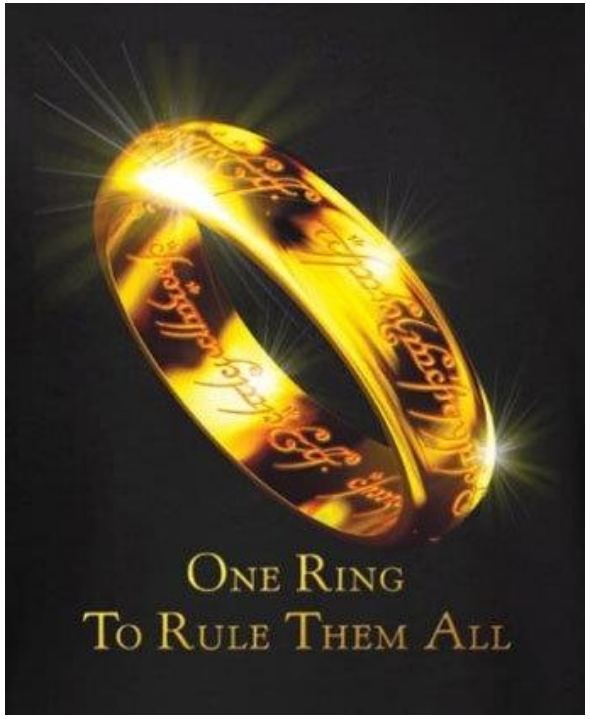
- E.g., μ = [illegible]

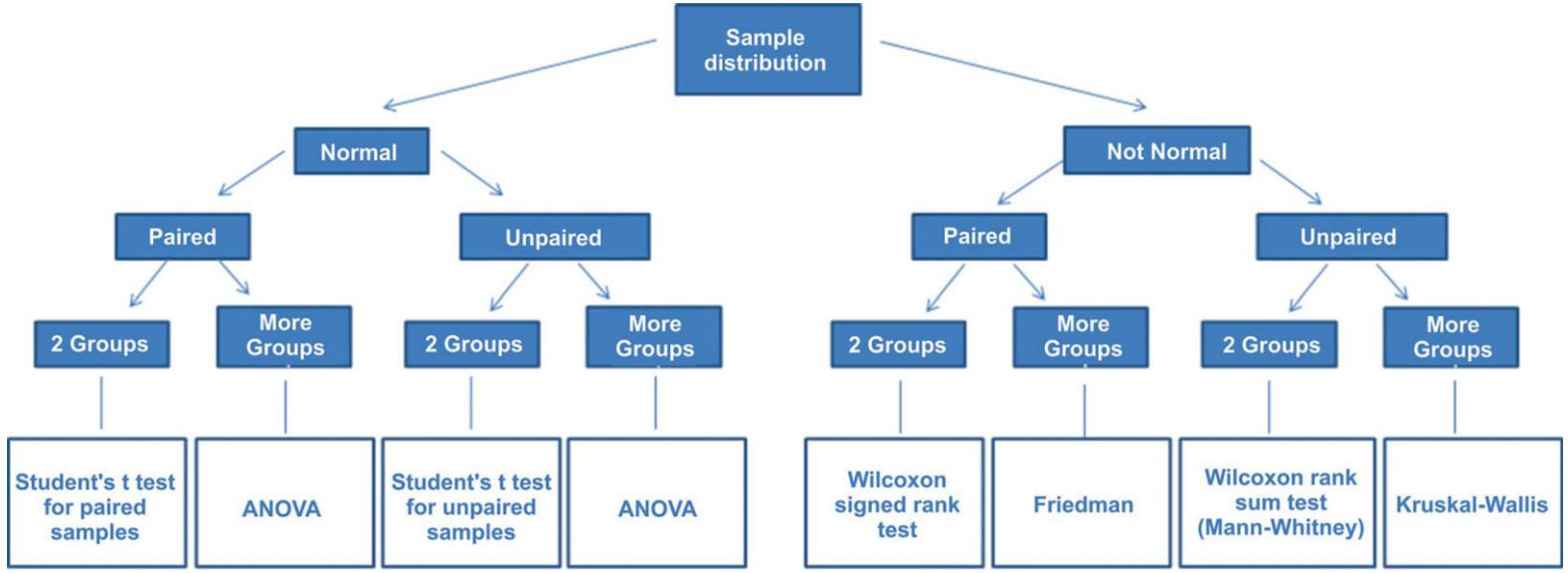
This claim implies we should get a certain distribution of statistics



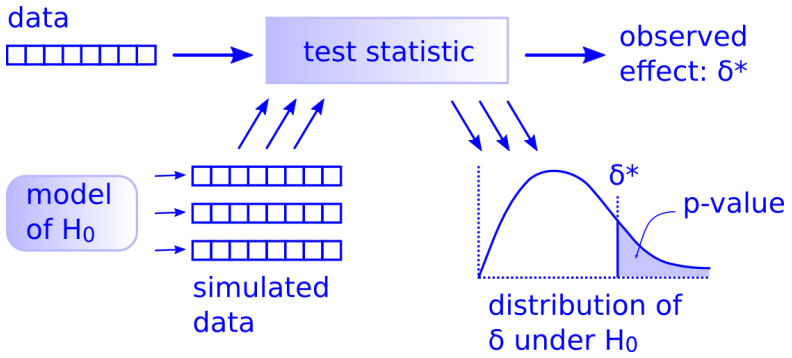
If our observed statistic is highly unlikely, we reject the claim

The big picture: There is only one hypothesis test!





Just need to follow 5 steps!



Example: Is it possible to smell whether someone has Parkinson's disease?

Joy Milne claimed to have the ability to smell whether someone had Parkinson's disease

To test this claim researchers gave Joy 6 shirts that had been worn by people who had Parkinson's disease and 6 shirts by people who did not

Joy identified 11 out of the 12 shirts correctly

Questions about the experiment

1. What are the cases in this experiment?
   - Hint: draw out what the data would look like

2-3. What is the variable of interest, and is it categorical or quantitative?

4-5. What is the observed statistic - and what symbols should we use to denote it?

6. What is the population parameter we are trying to estimate, and what symbol should we use to denote it?

7. Do you think the results are due to chance?
   - i.e., do you think Joy got 11 correct answers by guessing?

8. Do you believe Joy can really smell whether someone has Parkinson's disease?

Questions about the experiment

1. What are the cases in this experiment?
   - A: Each case is trial where Joy had to smell one shirt

2-3. What is the variable of interest, and is it categorical or quantitative?

   - A: The variable of interest is whether Joy correctly identified whether a shirt was worn by someone who had Parkinson's disease.
   - A: It is categorical (correct or incorrect)

4-5. What is the observed statistic - and what symbols should we use to denote it?

   - A: The observed statistic is the proportion of shirts Joy correctly identified
   - A: The symbols we use to denote this statistic is $\hat{p}$

Questions about the experiment

6-7. What is the population parameter we are trying to estimate, and what symbol should we use to denote it?

- A: The population parameter is how many shirts she would have correctly determined that were worn by people with/without Parkinson's disease if she had to smell infinitely many shirts.
- A: The symbol we would use to denote the population parameter is π

8. Do you think the results are due to chance?

- A: Opinions may vary. We will do more analyses to quantify this!

9. Do you believe Joy can really smell whether someone has Parkinson's disease?

- A: Opinions may vary, but it's good to have an opinion going in!

Smelling Parkinson's disease

If Joy was just guessing, what would we expect the value of the parameter to be?

$\pi = 0.5$

If Joy was not guessing, what would we expect the value of the parameter to be?

$\pi > 0.5$

Chance models

How can we assess whether 11 out of 12 correct trials ($\hat{p}$ = .916) is beyond what we would expect by chance?

If Joy was guessing, we can model his guesses as a coin flip:

Heads = correct guess

Tails = incorrect guess

We could flip 12 coins and see if we get 11 heads

Chance models

To really be sure, we should repeat flipping a coin 12 many times.

Any ideas how to do this?

Flipping coins in R

We can simulate coin flipping using the rbinom() function

```
flip_simulations <- rbinom(num_sims, size, prob)
```

num_sims: the number of simulations run

- Typically we do around 10,000 repeats

size: the number of trials on each simulation

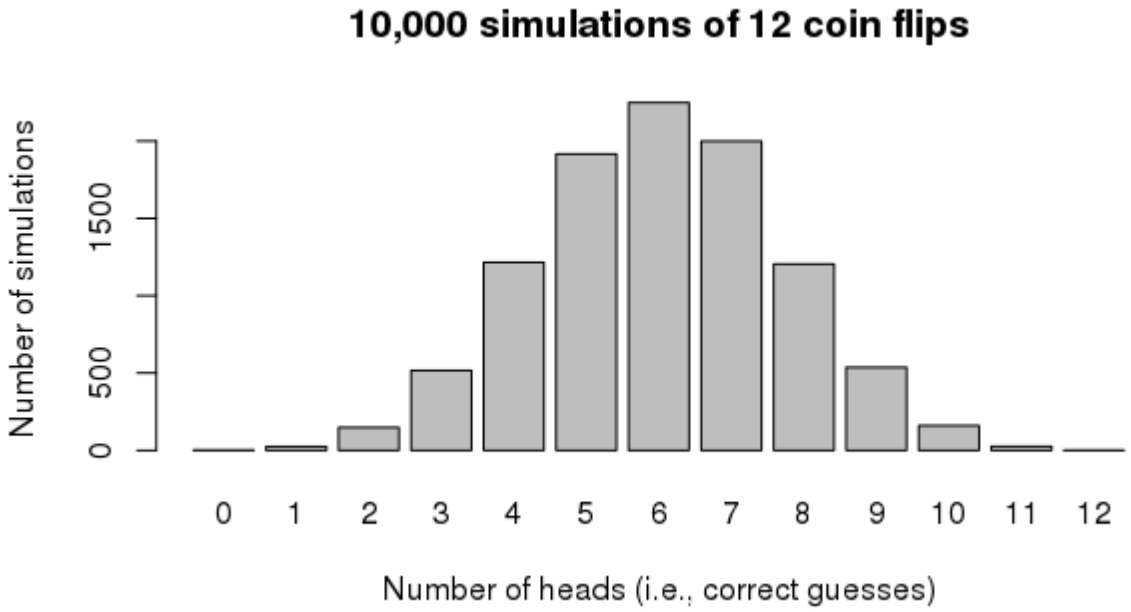
- 12 for simulating Joy's guesses

prob: the probability of success on each trial

- .5 if Joy was guessing

Simulating Flipping 12 coins 10,000

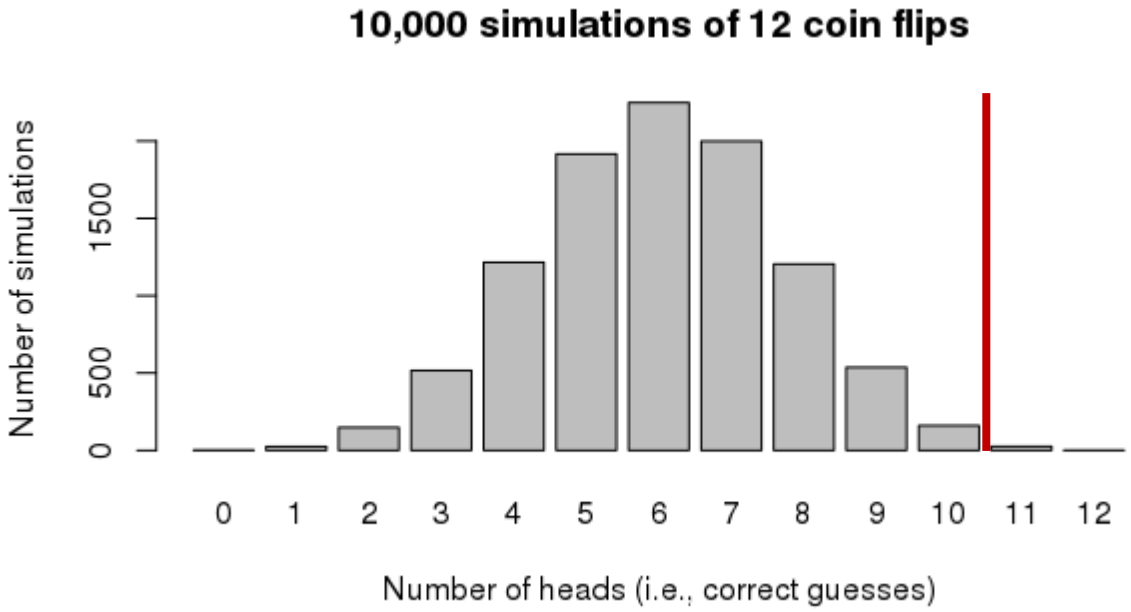
0	2
1	26
2	147
3	558
4	1269
5	1967
6	2310
7	1843
8	1142
9	537
10	162
11	33
12	4



Is it likely that Joy was guessing?

Simulating Flipping 12 coins 10,000

0	2
1	26
2	147
3	558
4	1269
5	1967
6	2310
7	1843
8	1142
9	537
10	162
11	33
12	4



Is it likely that Joy was guessing?

Do you believe Joy can really smell whether someone has Parkinson's disease?

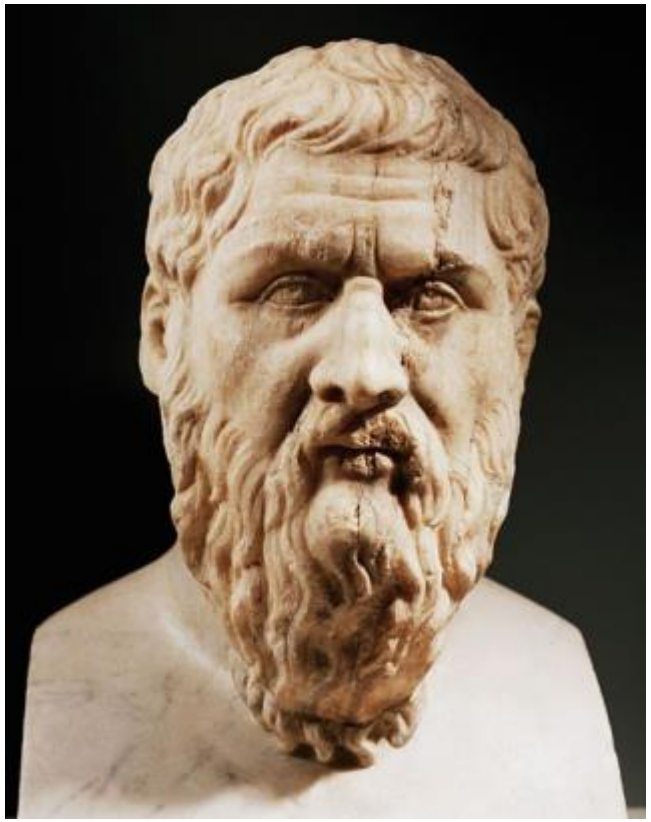
Is it possible to smell whether someone has Parkinson's disease?

We will examine this in R in a minute…

But first, let's review hypothesis testing terminology

Review of terminology and the 5 steps of hypothesis tests

Question: who is this?



A: Gorgias

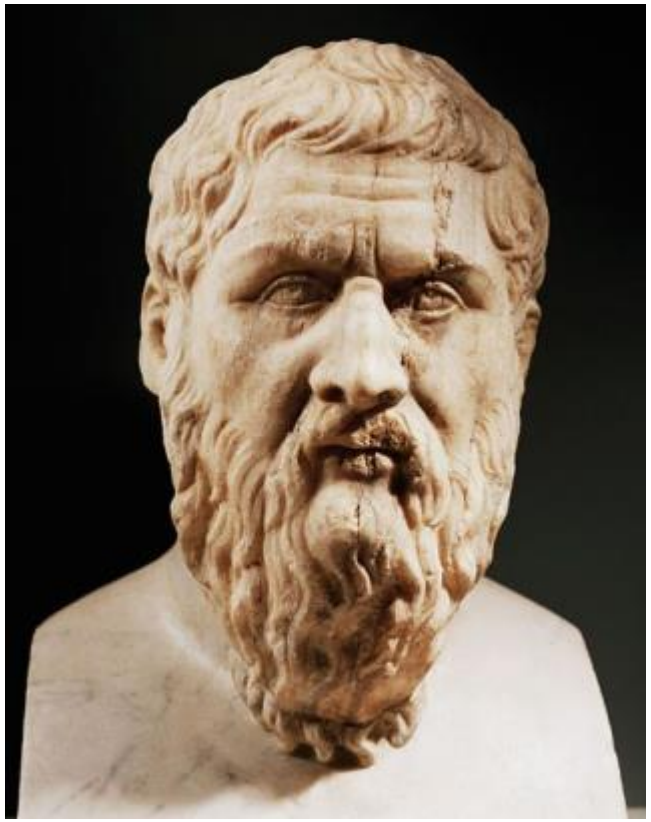
Question: Who is Gorgias?

A: a skeptic

Question: Does Gorgias believe Joy can smell Parkinson’s disease?

A: No!

Question: who is this?



Gorgias believes in the ***null hypothesis***

- that Joy was guessing

How can we write the null hypothesis in symbols?

$H_0: \pi = 0.5$

We believe in the ***alternative hypothesis***

- Joy can smell Parkinson's disease

How can we write the alternative hypothesis in symbols?

$H_A: \pi > 0.5$

Question: who is this?

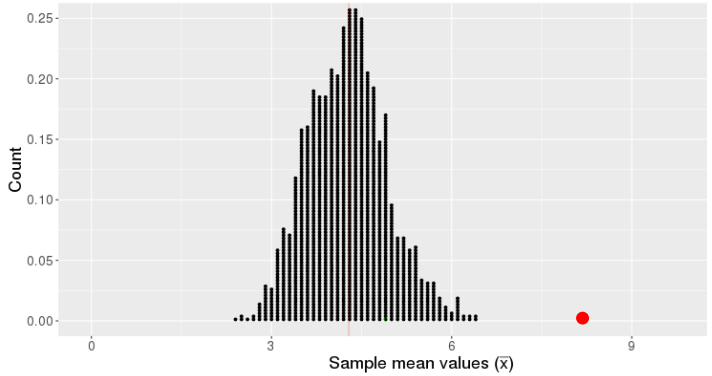
To prove Gorgias wrong, we will start by assuming he is right!

Namely, we will assume H_0 (that $\pi = 0.5$)

We will then generate a number of statistics ($\hat{p}$) that are consistent with H_0

- i.e., we will create a ***null distribution***

If our observed statistic looks very different from the statistics generated under we can reject H_0 and accept H_A

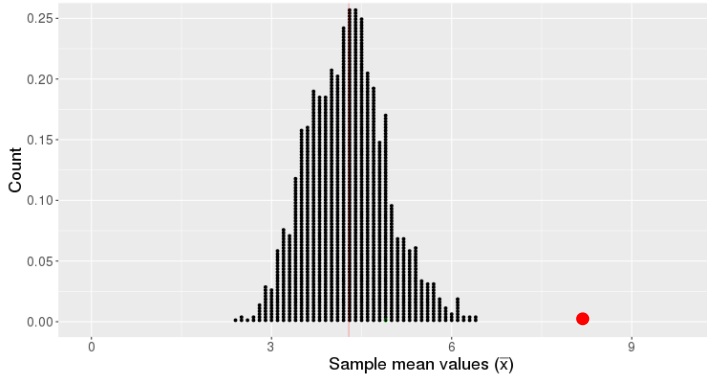


Terminology

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

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

The alternative hypothesis is established by observing evidence that inconsistent with the null hypothesis



Review: the Joy smelling Parkinson's disease

1. What is the null hypothesis?

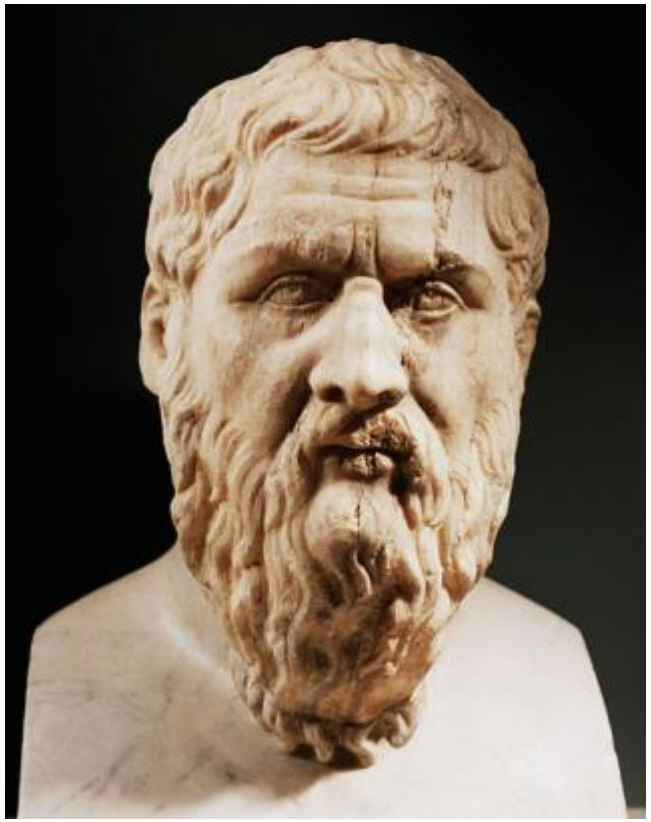
2. We can write this in terms of the population parameter as:

$H_0: \pi = 0.5$

3. What is the alternative hypothesis?

$H_A: \pi > 0.5$

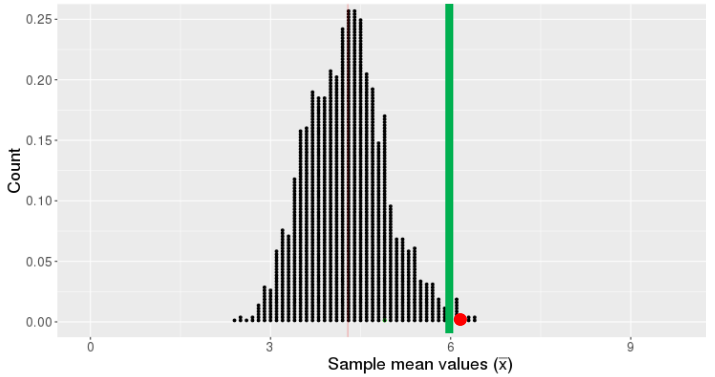
Setting the rules



Life wisdom: If you are going to make a bet with a nihilist, you'd better agree to the rules first!

Rules

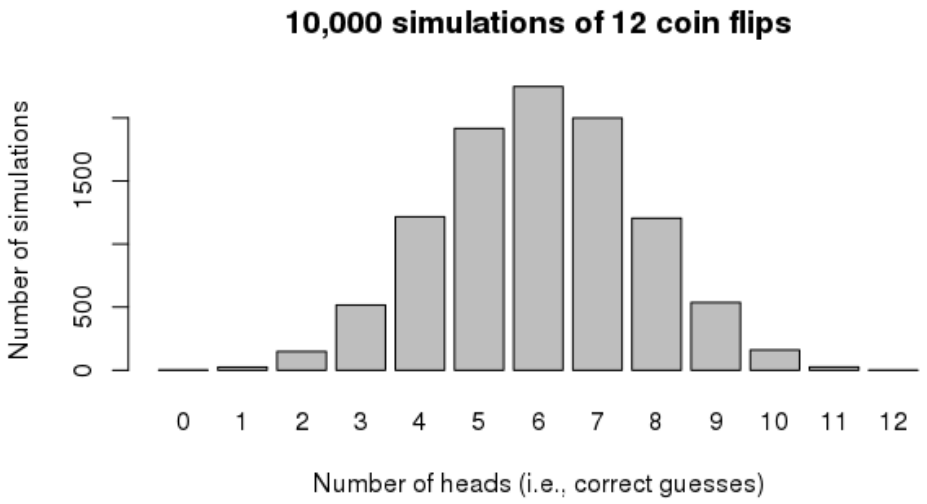
- If there is a less than 5% chance we would get a random statistic as or more extreme than the observed statistic (if H_0 is true) we will reject H_0
  - i.e., Gorgias loses the bet
- In symbols: $\alpha = 0.05$



Null Distribution

A **null distribution** is the distribution of statistics one would expect if the null hypothesis (H_0) was true

i.e., the null distribution is the statistics one would expect to get if nothing interesting was happening



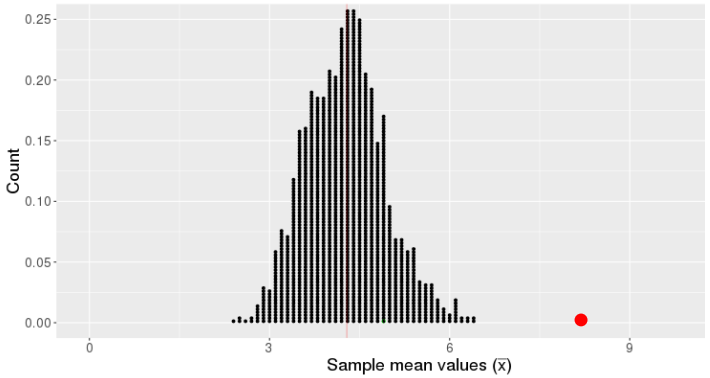
P-values

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

- i.e., the probability that we would get a statistic as extreme as our observed statistic from the null distribution

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

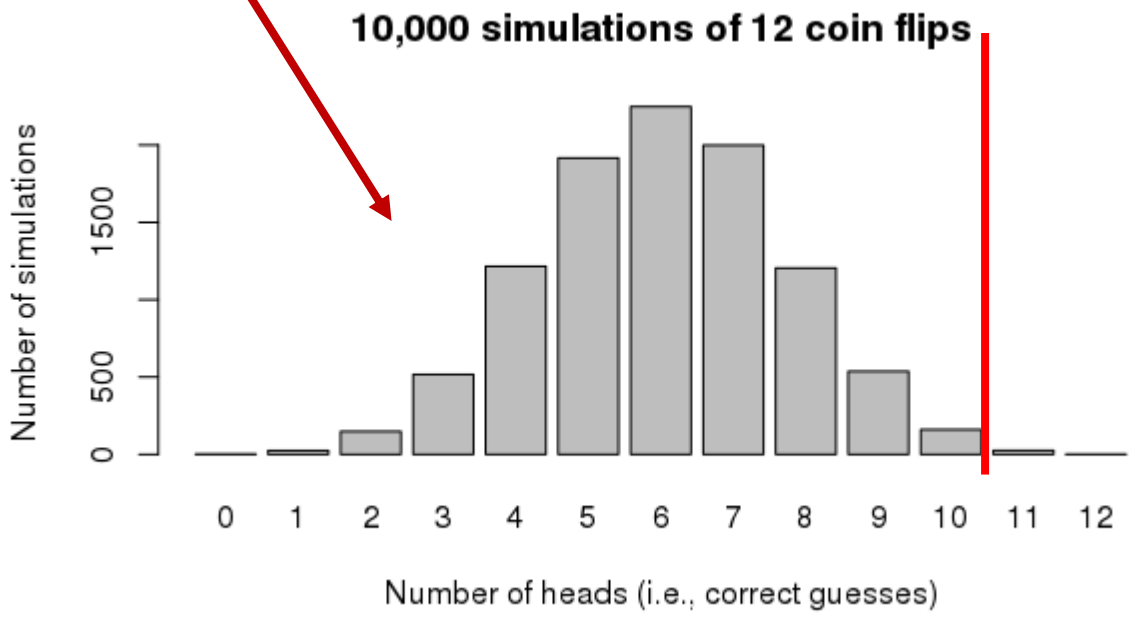
The smaller the p-value, the stronger the statistic evidence is against the null hypothesis



Joy example

0	2
1	26
2	147
3	558
4	1269
5	1967
6	2310
7	1843
8	1142
9	537
10	162
11	33
12	4

Null distribution



p-value = 37/10000 = 0.0037

Statistical significance

When our observed sample statistic is unlikely to co[illegible] the null distribution, people often say the results are **statistica[illegible]gnificant**

- i.e., our p-value is less than α
- i.e., Gorgias lost the bet!

'Statistically significant' results mean we have strong evidence against H_0 in favor of H_a

- The American Statistical Association rejects the phrase 'statistically significant'

5 steps for testing hypotheses

1. State the null hypothesis… and the alternative hypothesis
   - Joy was just guessing so the results are due to chance: $H_0: \pi = 0.5$
   - Joy is getting more correct results than expected by chance: $H_A: \pi > 0.5$

2. Calculate the observed statistic (and visualize the data)
   - Joy got 11 out of 12 guesses correct, or $\hat{p} = .917$

3. Create a null distribution that is consistent with the null hypothesis
   - i.e., what statistics would we expect if Joy was just guessing

4. Examine how likely the observed statistic is to come from the null distribution
   - What is the probability that the Joy would guess 11 or more correct?
   - i.e., what is the p-value

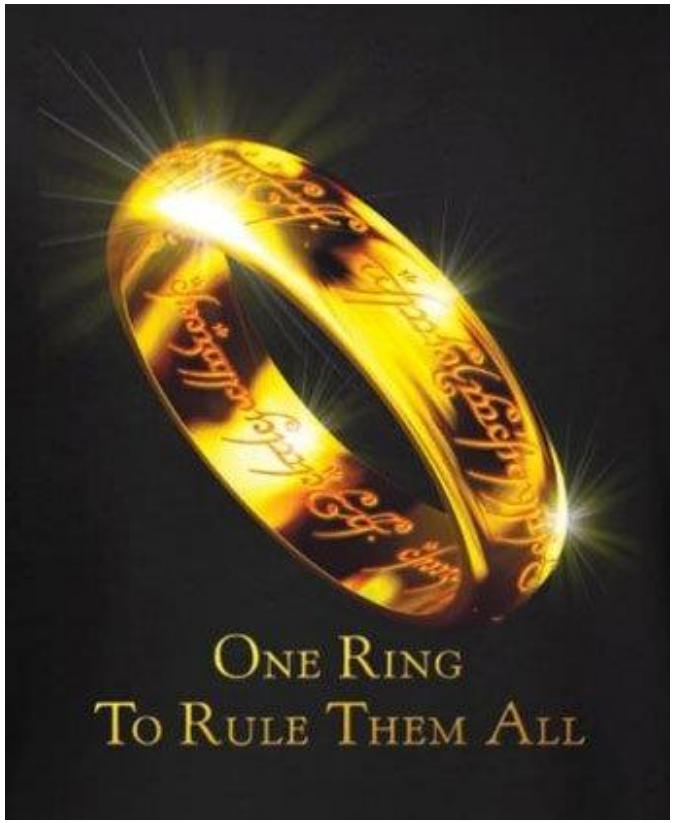
5. Make a judgement
   - If we have a small p-value, this means that $\pi = .5$ is unlikely and so $\pi > .5$
   - i.e., we could say our results are 'statistically significant'

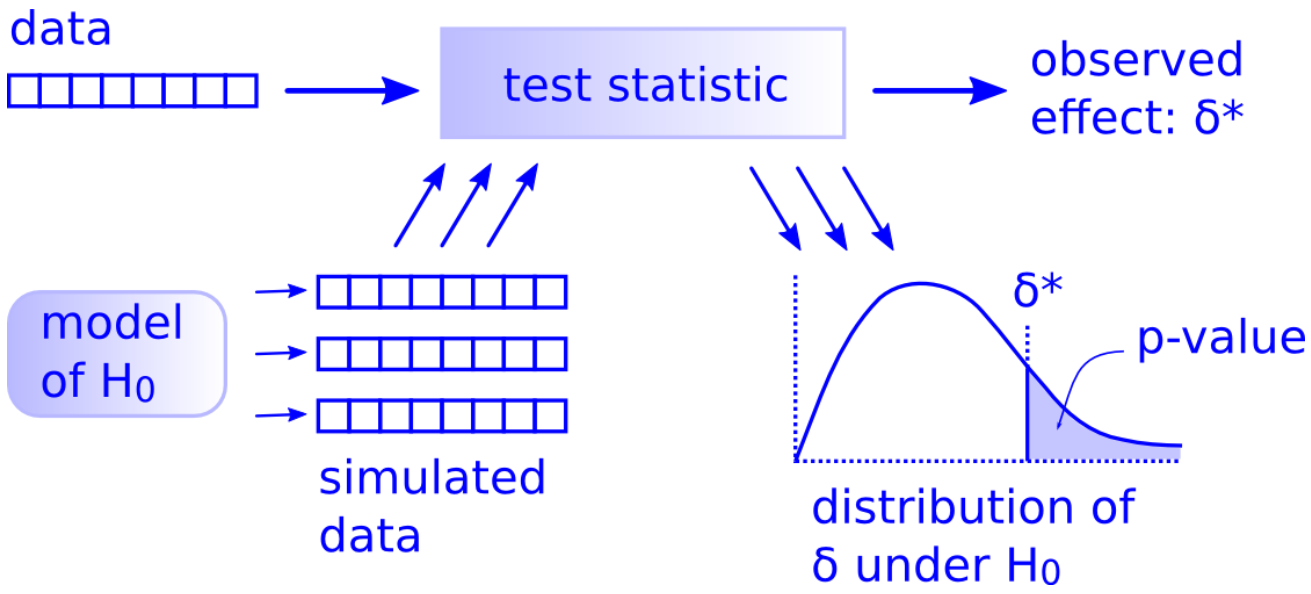
Is it possible to smell whether someone has Parkinson's disease?

Let's examine this in R!

Hypothesis tests comparing 2 means

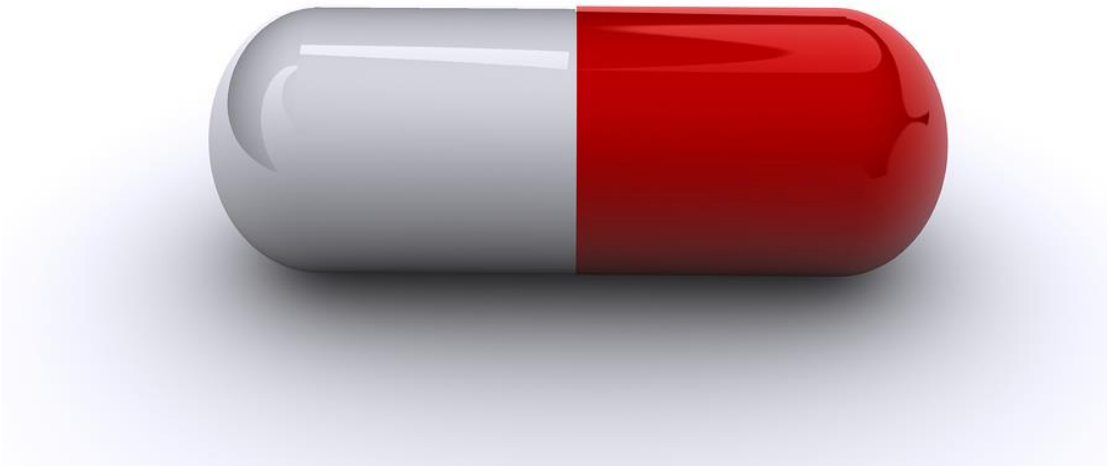
The big picture: There is only one hypothesis test!





Just need to follow 5 steps!

Hypothesis tests for comparing two means



Question: Is this pill effective?

Testing whether a pill is effective

How would we design a study?

What would the cases and variables be?

What would the parameter and statistic of interest be?

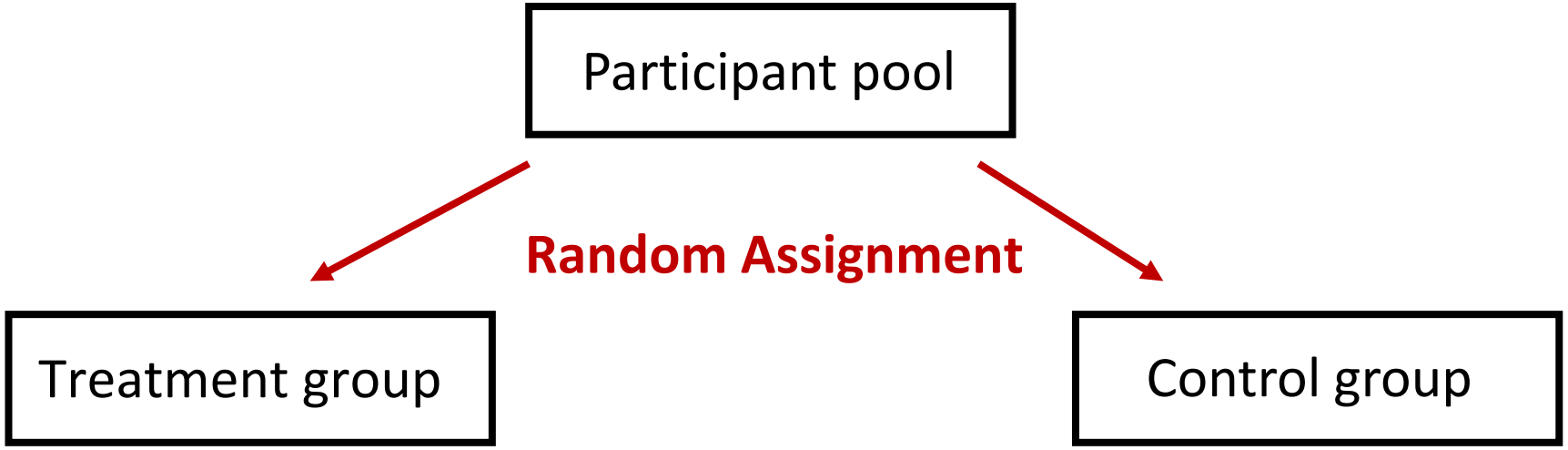
What are the null and alternative hypotheses?

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

Experimental design

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



Observational and experimental studies

An **experiment** is a study in which the researcher actively controls one or more of the explanatory variables

- Allows one to get at questions of **causation**!

An **observational study** is a study in which the researcher does not actively control the value of any variable but simply observes the values as they naturally exist

Question: Are the smelling Parkinson's disease and/or drug studies experimental or observational?

Hypothesis tests for differences in two group means

1. State the null and alternative hypothesis
   - H_0: $\mu_{Treatment} = \mu_{Control}$ or $\mu_{Treatment} - \mu_{Control} = 0$
   - H_A: $\mu_{Treatment} > \mu_{Control}$ or $\mu_{Treatment} - \mu_{Control} > 0$

2. Calculate statistic of interest
   - $\bar{x}_{Effect} = \bar{x}_{Treatment} - \bar{x}_{Control}$

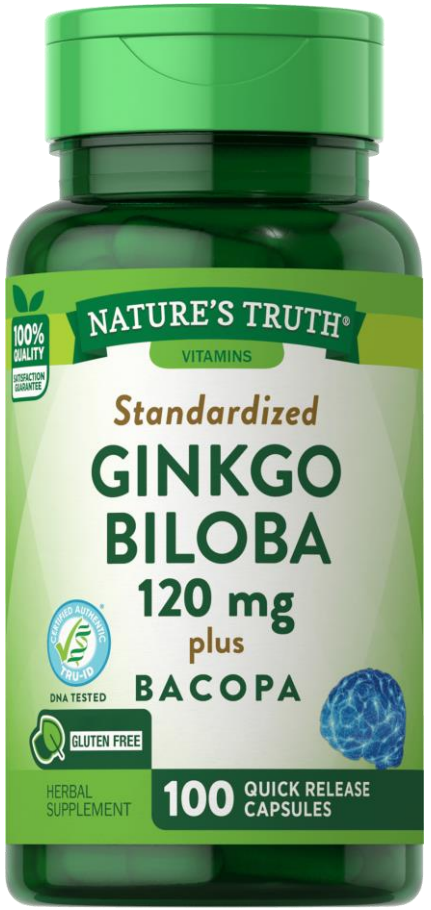
Example: Does Gingko improve memory?

A double-blind randomized controlled experiment by Solomon et al (2002) investigated whether taking a Ginkgo supplement could improve memory

- A treatment group of n = 104 participants took a Gingko supplement 3 times per day for 6 weeks
- A control group of n = 99 participants took a placebo 3 times per day for 6 weeks

Standardized neuropsychological tests of learning and memory, attention and concentration were measured at the end of the six week period.

Question: Was there a difference in a composite cognitive score between the treatment and control groups?



1. State the null and alternative hypothesis

In words:

- **Null hypothesis**: The average memory score will be the same for participants who took Gingko and the placebo
- **Alternative hypothesis**: The average memory score will be different for the two groups.

In symbols:

- H_0: $\mu_{Treatment} = \mu_{Control}$ or $\mu_{Treatment} - \mu_{Control} = 0$
- H_A: $\mu_{Treatment} \neq \mu_{Control}$ or $\mu_{Treatment} - \mu_{Control} \neq 0$

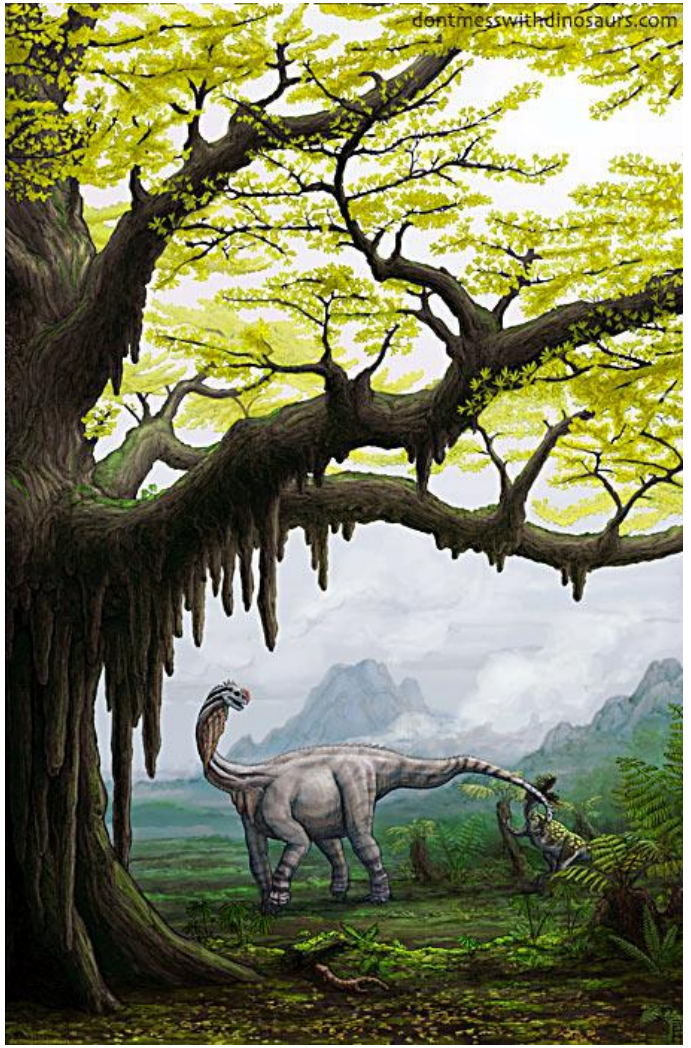
2. Visual the data can calculate the observed statistic

How could we visualize the data?

- We will try this in R soon…

What could we use for the observed statistic?

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



3. Create the null distribution!

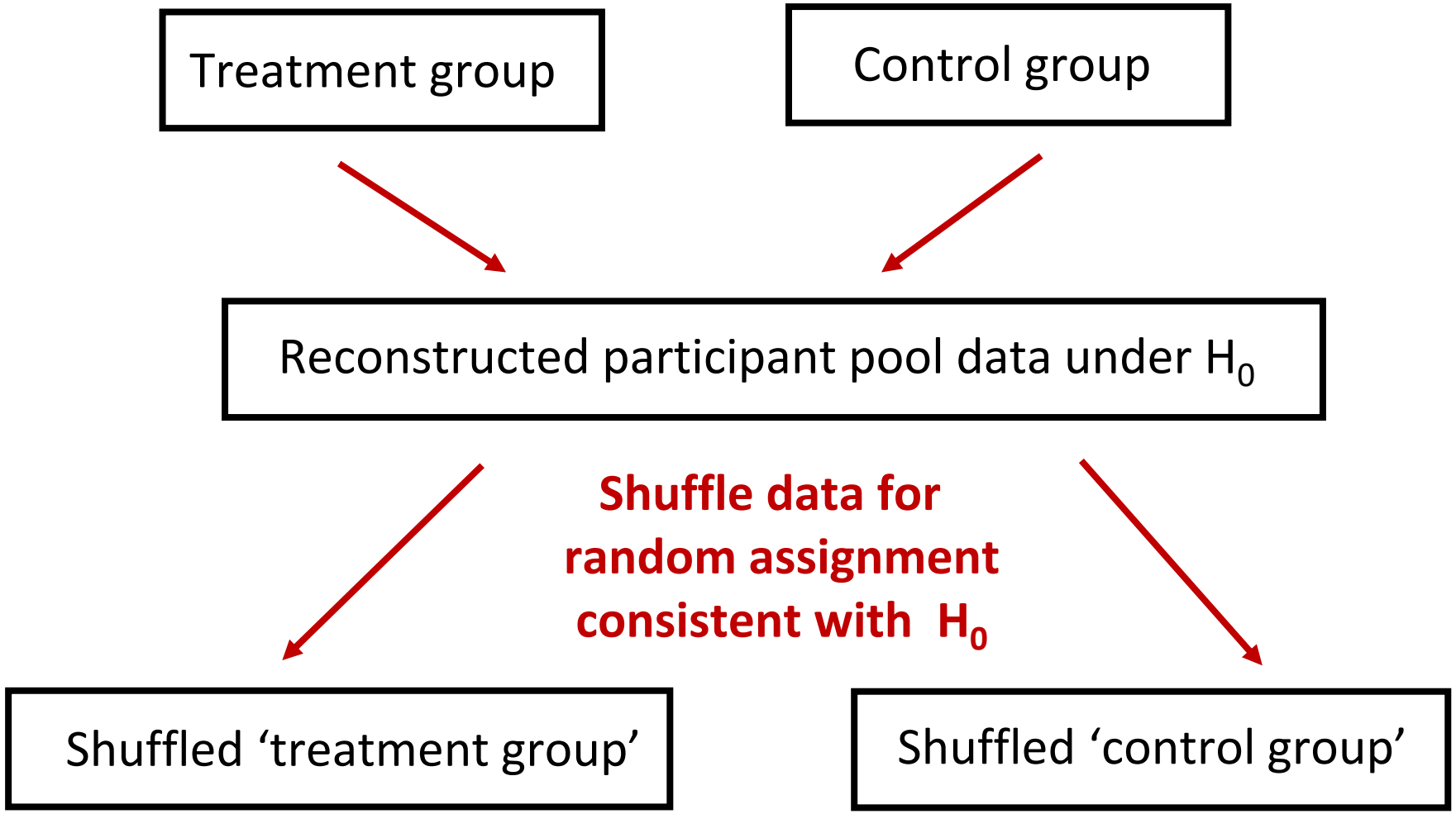
How could we create the null distribution?

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

- 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: $\bar{x}_{Shuff\_Treatment} - \bar{x}_{Shuff\_control}$

3. Create a null distribution

1. Combine data from both groups

2. Shuffle data

3. Randomly select 104 points to be the ‘null’ treatment group

4. Take the remaining 99 points to the ‘null’ control group

5. Compute the statistic of interest on these ‘null’ groups

6. Repeat 10,000 times to get a null distribution

Let’s try the rest of the hypothesis test in R…