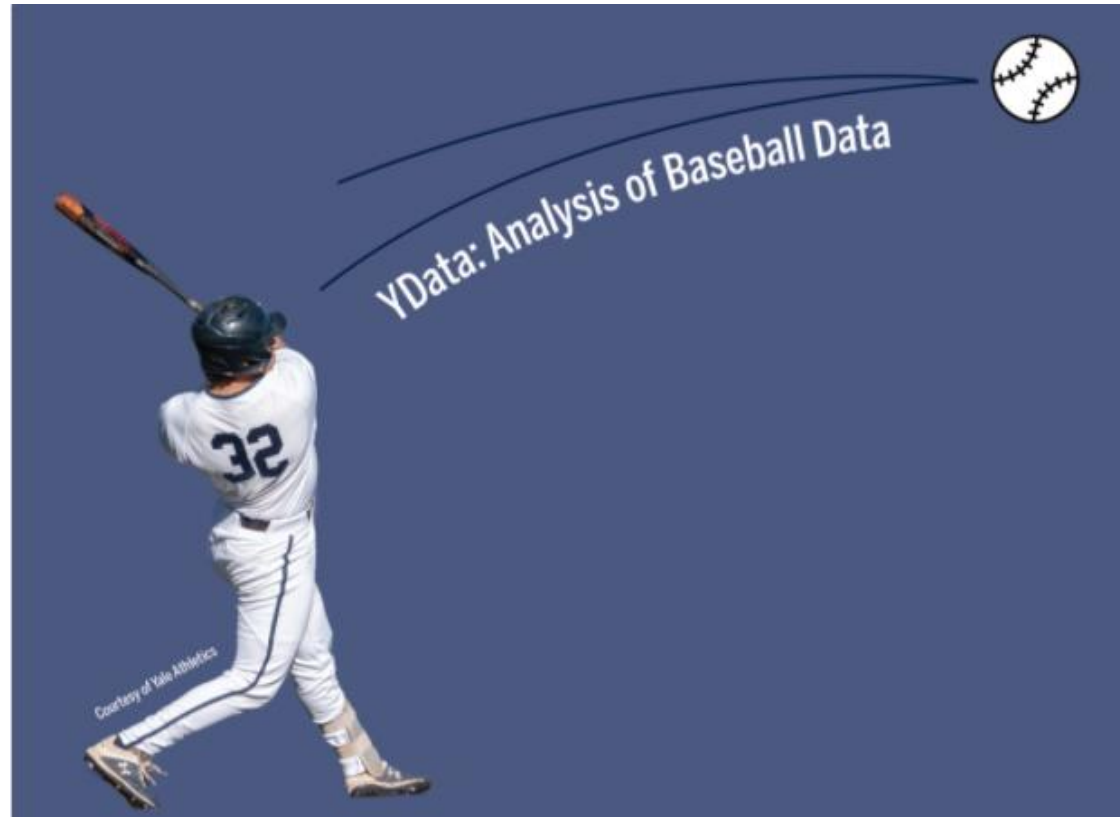


Hypothesis tests



Overview

Lab 5 discussion

Discussion of chapter 5 of Astroball

Quick example of Bayesian inference

Hypothesis tests for a single proportion

Hypothesis tests comparing two means

Lab 5: questions?

How did it go?

Class plan:

Next Wednesday is a break day so there will be no class

The midterm exam will be the following Wednesday

- Focus: data manipulation, Python programming, probability, hypothesis tests
- Last year's exam is on GitHub

Astroball discussion

Let's discuss the chapter for 7 minutes in breakout rooms and then have a larger conversation as a group

- Discuss your quote and reaction to chapter 5

Prospects Astros were considering drafting 1-1

Astros considering 6 prospects

- Less seriously: Aaron Nola, Nick Gordon
- More seriously: Carlos Rodon, Alex Jackson, Brady Aiken, Tyler Kolek

Is anyone familiar with these players?

Prospects Astros were considering drafting 1-1

Astros considering 6 prospects

- Less seriously: Aaron Nola, Nick Gordon



Aaron Nola

Position: Pitcher

Bats: Right • **Throws:** Right

6-2, 200lb (188cm, 90kg)

Team: [Philadelphia Phillies](#) (majors)

[More bio, uniform, draft, salary info ▼](#)

SUMMARY	WAR	W	L	ERA	G	GS	SV	IP	SO	WHIP
2020	2.2	5	5	3.28	12	12	0	71.1	96	1.079
Career	21.7	58	40	3.47	139	139	0	842.2	922	1.164



Nick Gordon

Positions: Shortstop and Second Baseman

Bats: Left • **Throws:** Right

6-0, 160lb (183cm, 72kg)

Team: [Minnesota Twins](#) (majors)

[More bio, uniform, draft, salary info ▼](#)

Prospects Astros were considering drafting 1-1

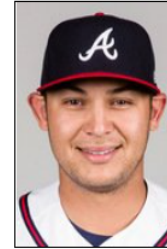


Carlos Rodón

Position: Pitcher
Bats: Left • **Throws:** Left
6-3, 250lb (190cm, 113kg)
Team: [Chicago White Sox](#) (majors)

[More bio, uniform, draft, salary info ▼](#)

SUMMARY	WAR	W	L	ERA	G	GS	SV	IP	SO	WHIP
2020	-0.4	0	2	8.22	4	2	0	7.2	6	1.565
Career	6.5	29	33	4.14	97	92	0	536.2	525	1.379



Alex Jackson

Position: Catcher
Bats: Right • **Throws:** Right
6-2, 215lb (188cm, 97kg)
Team: [Atlanta Braves](#) (majors)

[More bio, uniform, draft, salary info ▼](#)

SUMMARY	WAR	AB	H	HR	BA	R	RBI	SB	OBP	SLG	OPS	OPS+
2020	0.0	7	2	0	.286	0	0	0	.286	.429	.714	84
Career	-0.4	20	2	0	.100	0	0	0	.182	.150	.332	-13



Brady Aiken


Position: Starting Pitcher
Bats: Left • **Throws:** Left
6-4, 205lb (193cm, 92kg)
Team: [Cleveland Indians](#) (minors)

[More bio, uniform, draft, sal](#)



Tyler Kolek

Position: Pitcher
Bats: Right • **Throws:** Right
6-5, 260lb (196cm, 117kg)

Born: December 15, 1995 (Age: 25-090d) in Shepherd, TX 
Draft: Drafted by the [Miami Marlins](#) in the [1st round](#) (2nd) of the 2014 from [Shepherd HS \(Shepherd, TX\)](#).

High School: [Shepherd HS \(Shepherd, TX\)](#)

Full Name: Tyler Frank Kolek

Twitter: [@tylerkolek](#)

MLB

The Astros Are Trying To Dick Draft Picks Out Of Their Money



Tom Ley
7/15/14 11:58AM



2



2014 daft results



Alex Bregman

Positions: Third Baseman, Shortstop and Leftfielder

Bats: Right • **Throws:** Right

6-0, 192lb (183cm, 87kg)

Team: [Houston Astros](#) (majors)

[More bio, uniform, draft, salary info ▼](#)

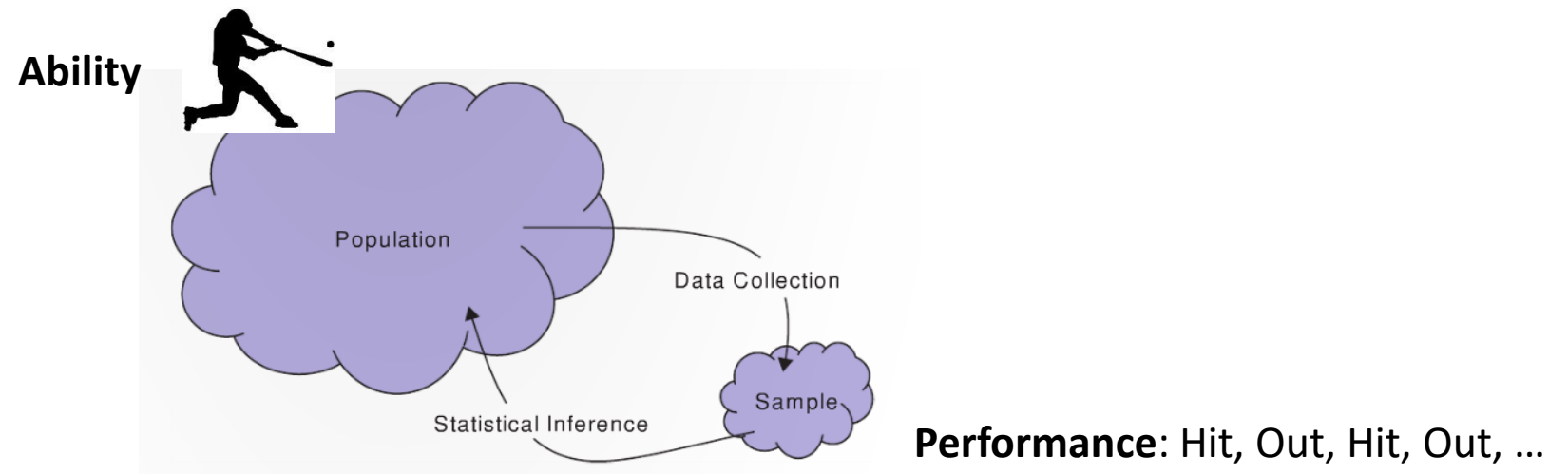
SUMMARY	WAR	AB	H	HR	BA	R	RBI	SB	OBP	SLG	OPS	OPS+
2020	1.0	153	37	6	.242	19	22	0	.350	.451	.801	116
Career	23.4	2058	582	105	.283	365	342	34	.381	.521	.902	142

Statistical inference

Statistical inference: use sample of data to deduce properties of an underlying population or stochastic process

In the context of baseball this usually means: looking at a player's **performance** to tell something about the player's **ability**

- **Ability:** innate talent
- **Performance:** outcomes from playing a number of games



Statistical inference

We've seen many cases of simulating a player's performance based on pre-specified abilities (probability)

- E.g., we can simulate a .333 OBP by rolling a die (or using an Python) to generate random data consistent with a .333 OBP



Hit, Out, Hit, Out, Out, Out, ...

With **statistical inference** we go in the other direction: we take a collection of outcomes and estimate the probability model parameters

Hit, Out, Hit, Out, Out, Out, ...



Estimate π_{hit}



Parameters vs. statistics

A **statistic** is a number that is computed from ***data in a sample***

- i.e., a number summarizing a player's performance
- We denote these with Latin characters (\hat{p} , \bar{x} , etc.)

A **parameter** is a number that describes some aspect of a ***population or process***

- i.e., a number summarizing a player's ability
- We denote parameters with Greek characters (π , μ , etc.)

Proportions for categorical data

The sample of say size $n = 100$

- out, out, hit, out, walk, out, ..., out

The proportion for a **sample** is a **statistic** denoted by \hat{p} (pronounced “p-hat”)

- $\hat{p}_{\text{hit}} = 28/100 = 0.28$

The proportion for a **population/process** is a **parameter** denoted π

- π_{hit} could be the proportion for:
 - Proportion of hits over all plate appearances in a player’s career
 - Proportion of future hits a player will have (e.g., a prospect)
 - Proportion of hits if a player batted an infinite number of times

\hat{p} is a **point estimate** of π

- i.e., \hat{p} our best guess of what π is

Bayesian Inference: Determining the ability of a player

Suppose there are 3 players with different true OBP abilities:

- Player H_1 's true OBP is .200 $H_1: \pi = .200$
- Player H_2 's true OBP is .333 $H_2: \pi = .333$
- Player H_3 's true OBP is .500 $H_3: \pi = .500$

One player is selected at random and we observe the player for 10 plate appearances

Can we tell whether it was player H_1 , H_2 , H_3 , who was picked?

Let's simulate this with a 4, 6, and 10 sided die

- 1 or 2 is on base
- Higher numbers are outs
- I will roll the die 10 times...

Simulating different sided dice rolls

Bayesian Inference: Determining the ability of a player

Suppose we got 5 on base events out of 5 plate appearances

Question: What die was chosen?

- i.e., what value is π ?

Determining the ability of a player

Here are the simulation results from 1000 simulations of rolling the different dice 10 times:

	0	1	2	3	4	5	6	7	8	9	10
0.200	95	271	315	199	80	35	5	0	0	0	0
0.333	17	81	206	270	226	121	59	16	4	0	0
0.500	2	15	36	108	205	248	196	124	53	21	1

Total number of simulations that produced 5 hits = $35 + 121 + 248 = 404$

$$\Pr(\pi = .200 \mid 5 \text{ hits}) = 35/404 = .087$$

$$\Pr(\pi = .333 \mid 5 \text{ hits}) = 121/404 = .299$$

$$\Pr(\pi = .500 \mid 5 \text{ hits}) = 205/404 = .614$$

Bayesian inference


Gives a probability distribution over ability (parameters) given that we have observed some performance (data)

$$Pr(H_i|data) = \frac{Pr(data|H_i) \cdot Pr(H_i)}{Pr(data)}$$

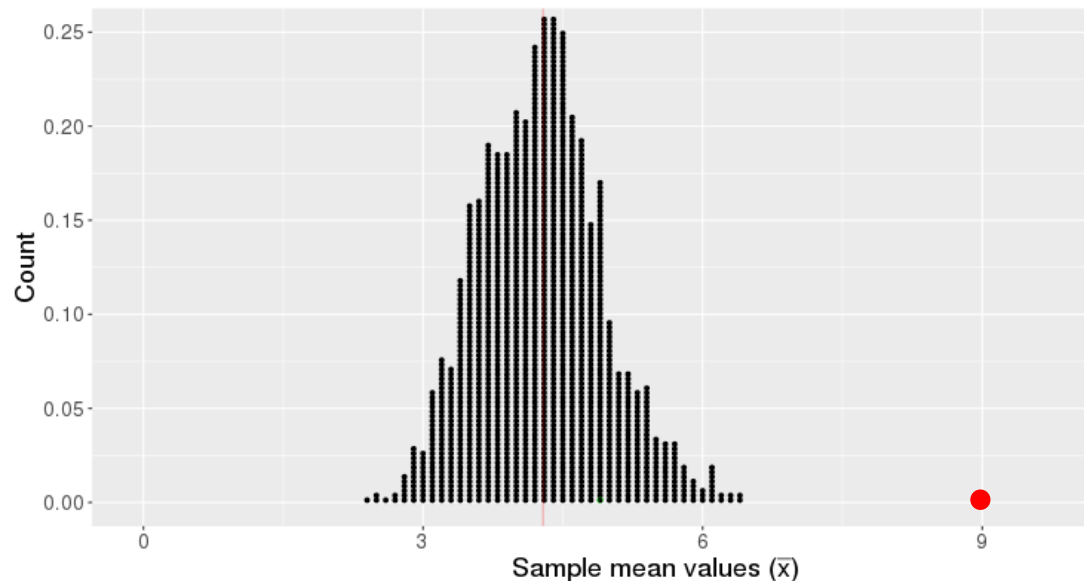
Hypothesis testing (frequentist inference)

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

Hypothesis tests in soccer

In the 2010 World Cup, Paul the Octopus (in a German aquarium) became famous for correctly predicting 11 out of 13 soccer games



Question: is Paul psychic?

Let's do the analyses in the [class 7 Jupyter notebook](#)

Paul the Octopus



Question: If Paul was not psychic, what proportion of games would we expect him to guess correctly?

- Answer: $\pi = .5$

Question: How could we calculate the probability Paul would **guess** 11 or more games correctly?

- Answer 1: We could flip a fair coin 13 times and see how many times we get 11 or more heads. We could then repeat this process 10,000 times.

Paul the Octopus



We can use the `sample_proportions()` function in the datascience package to simulate the proportion of heads we would get for flipping n coins.

```
p_hats = sample_proportions(n, [prob_heads, prob_tails])
```

- n : the number coin flips
- $[prob_heads, prob_tails]$: probability of heads and tails

For simulating if Paul was guessing, what should the code be?

```
prop_correct_guesses = sample_proportions(13, [.5, .5]).take(0)
```

Paul the Octopus



This gives us one simulation of flipping 13 coins

```
prop_correct_guesses = sample_proportions(13, [.5, .5]).take(0)
```

Q: What should we do next?

A: repeat this many times to get a “null distribution”.

We can then see how often 11/13 heads occurs in this “null distribution”

Paul the Octopus



A function to generate one point in the null distribution:

```
def generate_flip_proportion_heads(num_flips, prob_h):  
    return sample_proportions(13, [.5, .5]).take(0)
```


Paul the Octopus



We can then generate a large number of statistics consistent with Paul guessing using

```
null_distribution = make_array()
```

```
for i in range(0, 10000):
```

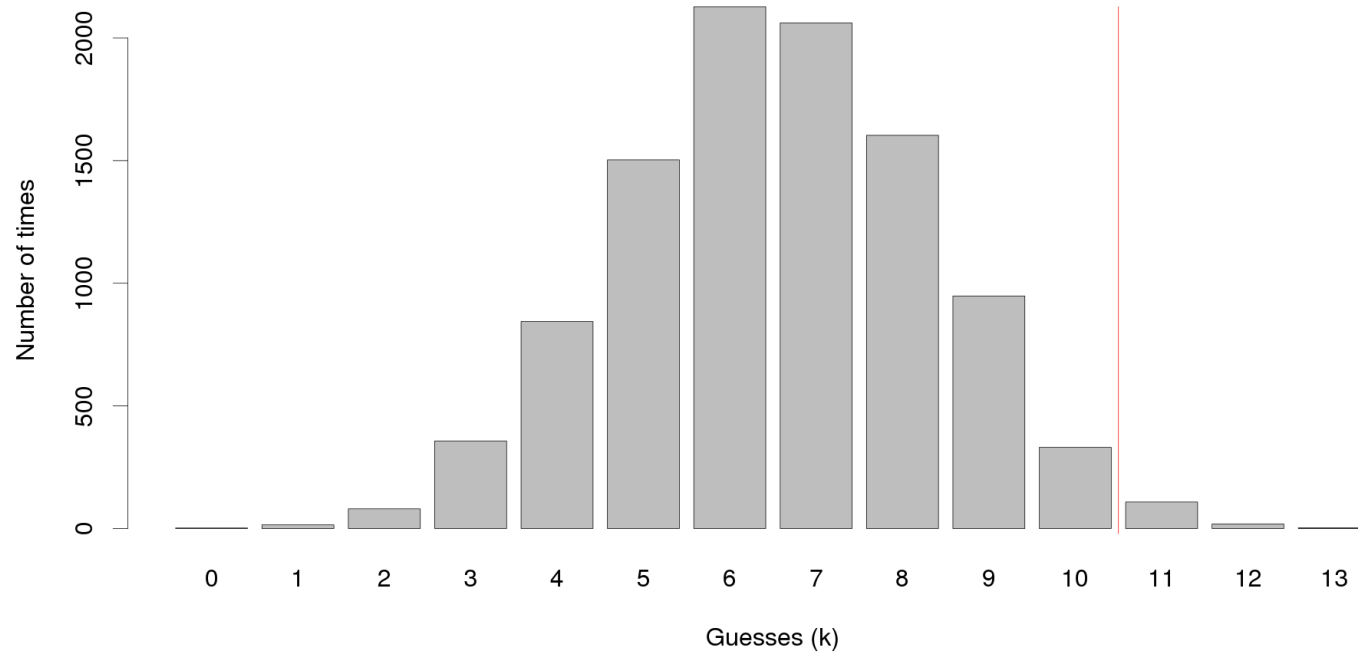
```
    curr_p_hat = generate_flip_proportion_heads(13, .5)
```

```
    null_distribution = np.append(null_distribution, curr_p_hat)
```

```
null_table = Table().with_column('Null Distribution', null_distribution)
```

```
null_table.hist(bins = np.arange(0, 1, 1/14))
```

Paul the Octopus



From looking at this figure, approximately how often an octopus that was randomly guessing, guess 11 out of 13 games correctly?

Paul the Octopus



We can calculate proportion of our simulated guesses that were as great or greater than the actually proportion of guesses Paul got correct using:

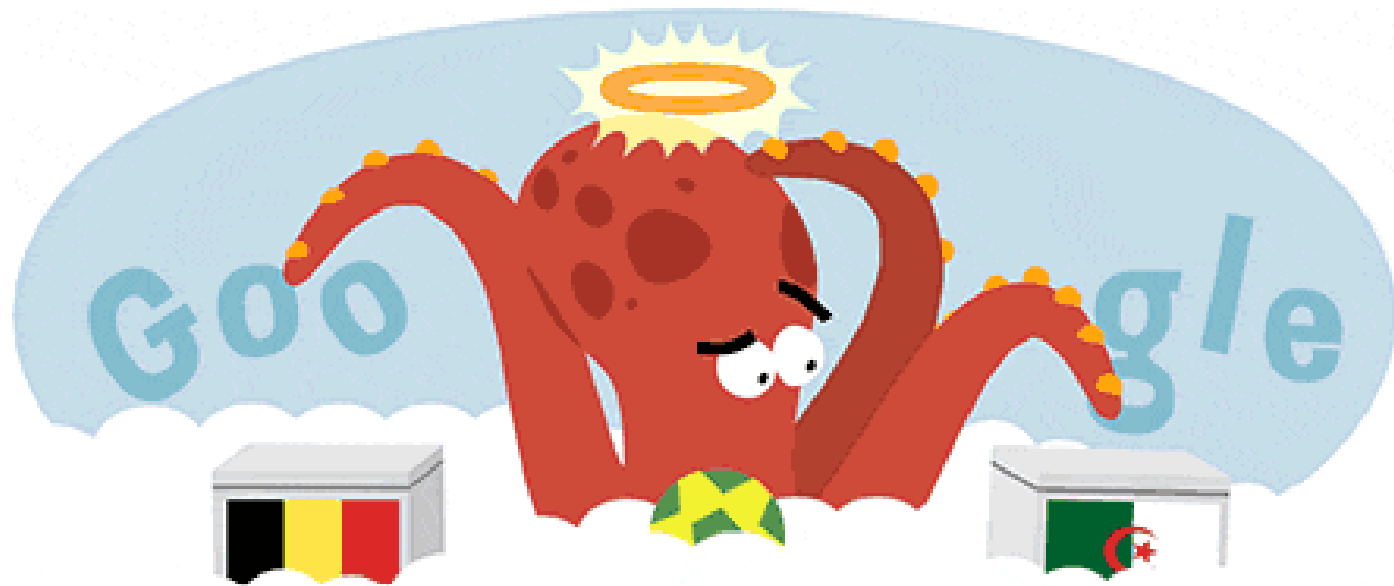
```
np.count_nonzero(null_distribution >= 11/13) / 10000
```

I got: 129 of 10,000 simulated experiments that have 11 or more correct guesses

- If Paul was guessing, he would only get 11 right $129/10,000 = 1.2\%$ of the time

Paul the Octopus

Do you think Paul is psychic?



Formalizing hypothesis testing

Let's describe what we just did...

1. First we stated two hypotheses which were:

- Null hypothesis: Paul is guessing $H_0: \pi = .5$
- Alternative hypothesis: Paul is psychic $H_A: \pi > .5$

2. Next we created a ***null distribution*** of \hat{p} 's that are consistent with what we would expect if the null hypothesis H_0 was true.

3. Finally we examined the probability we would get a random statistic from the null distribution that was greater than the observed \hat{p} statistic

- $\Pr(X \geq \hat{p} = 11/13 \mid \text{from the null distribution})$

P-values

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

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

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

How good is A-Rod's really?

In 2012, Alex Rodriguez had a .353 OBP based on 529 plate appearances

- Let us denote this observed performance with the symbol \hat{p}

This observed performance (\hat{p}) is due to both:

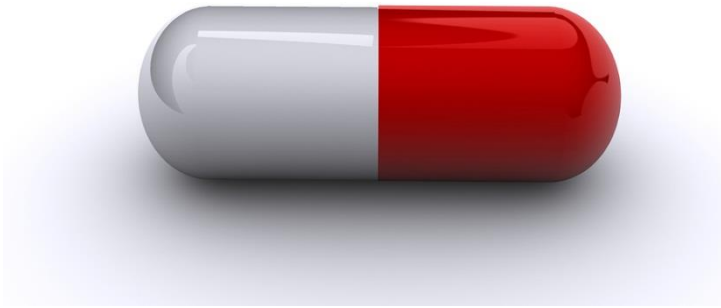
- a) A-Rod's innate ability or skill (π)
- b) Luck, randomness, chance

So how good is A-Rod really?

- Is it plausible that A-Rod's OBP ability π was really .300 and he just got lucky to get a \hat{p} of .353?

Lab 6!

Hypothesis tests for two means (or proportions)



Question: Is this pill effective?



Question: Does this steroid work?

Question: What does it mean for a pill to work?

Answer: On average the people who take the pill will be better than the people who do not

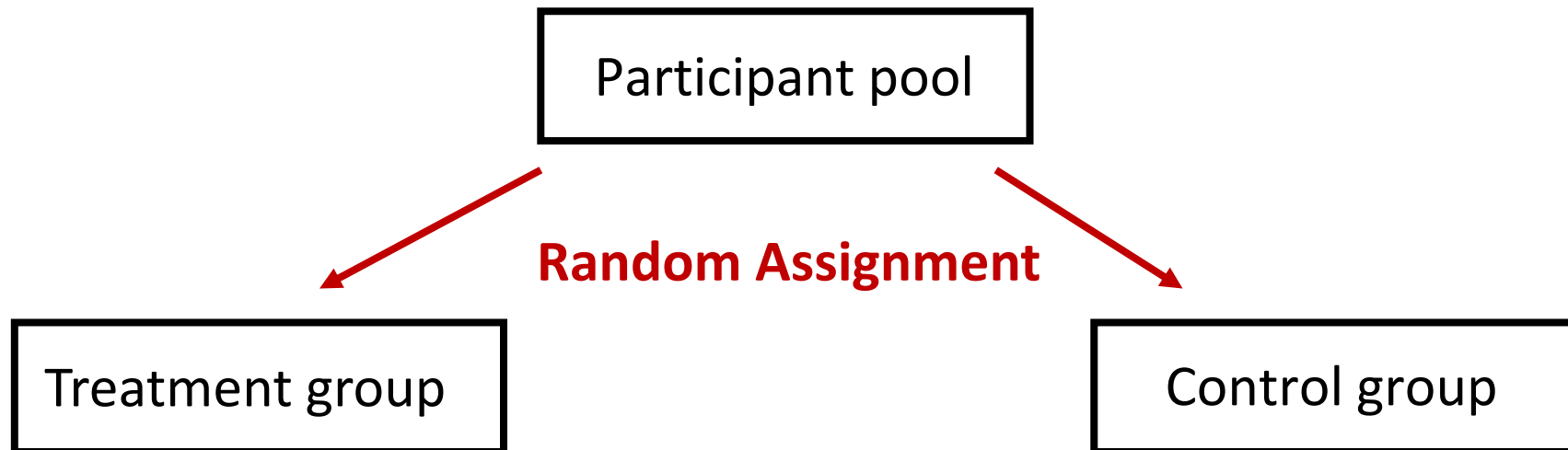
- $\mu_{\text{pill}} > \mu_{\text{no-pill}}$

Experimental design

Question: How can we design a study to test whether a pill ***causes*** an improvement in health?

Take a group of participants 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



Experimental design

Question: If the pill did not work, would you expect better outcomes in the treatment or control groups?

If we find that the treatment group performs better, then this could be due to:

- A systematic improvement due to the pill
- The treatment group is better due to the random assignment of people

If we can show that random assignment of people is not likely to account for the differences between the groups, then the pill must have caused the improvement in the treatment group

Assessing whether the random assignment could account for our results

Question: what is the first we do when testing hypotheses?

- 1. State the null and alternative hypotheses in symbols and words
 - $H_0: \mu_T = \mu_C$ or $\mu_T - \mu_C = 0$
 - $H_A: \mu_T > \mu_C$ or $\mu_T - \mu_C > 0$

What do we do next?

- Let's look at a real experiment!

2. Compute the statistic of interest

Kamath, et al, 2003, suspected that drinking tea might help boost one's immune system

To test this hypothesis recruited 21 healthy volunteers:

- 11 volunteers were assigned to drink 5 to 6 cups of tea a day
- The remaining 10 were assigned to drink that much coffee

After two weeks:

- Blood samples were taken from all participants and exposed to an antigen
- The production of and interferon gamma (which is related to an immune response) was measured

The data showed the following interferon gamma levels:

Tea	5	11	13	18	20	47	48	52	55	56	58
Coffee	0	0	3	11	15	16	21	21	38	52	

State the null and alternative hypotheses for this experiment and compute the statistic of interest

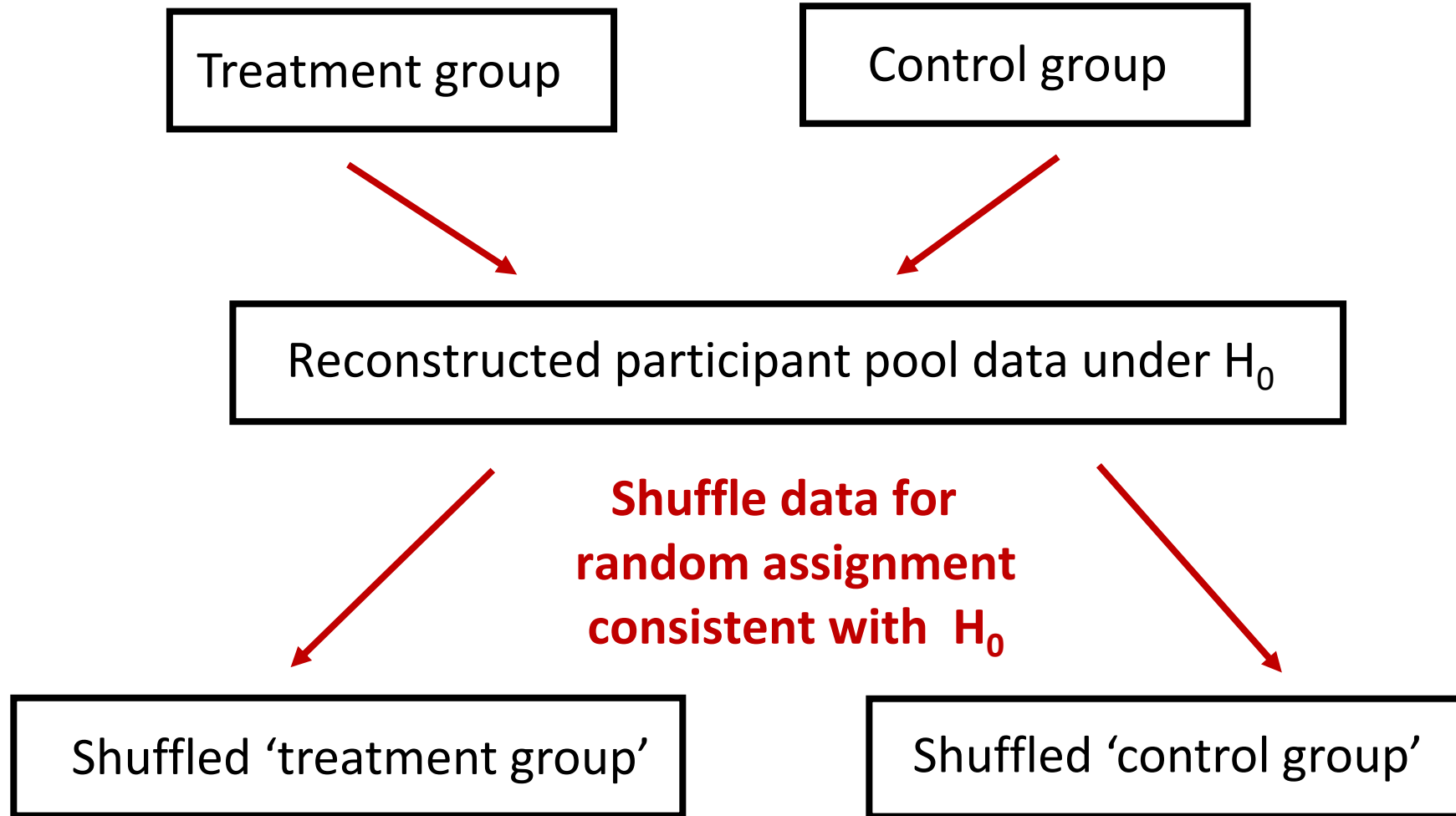
3. Creating a null distribution

Question: how can we create a null distribution here?

Under that null hypothesis there is no difference between the tea and coffee drinks, so we can generate a null distribution by:

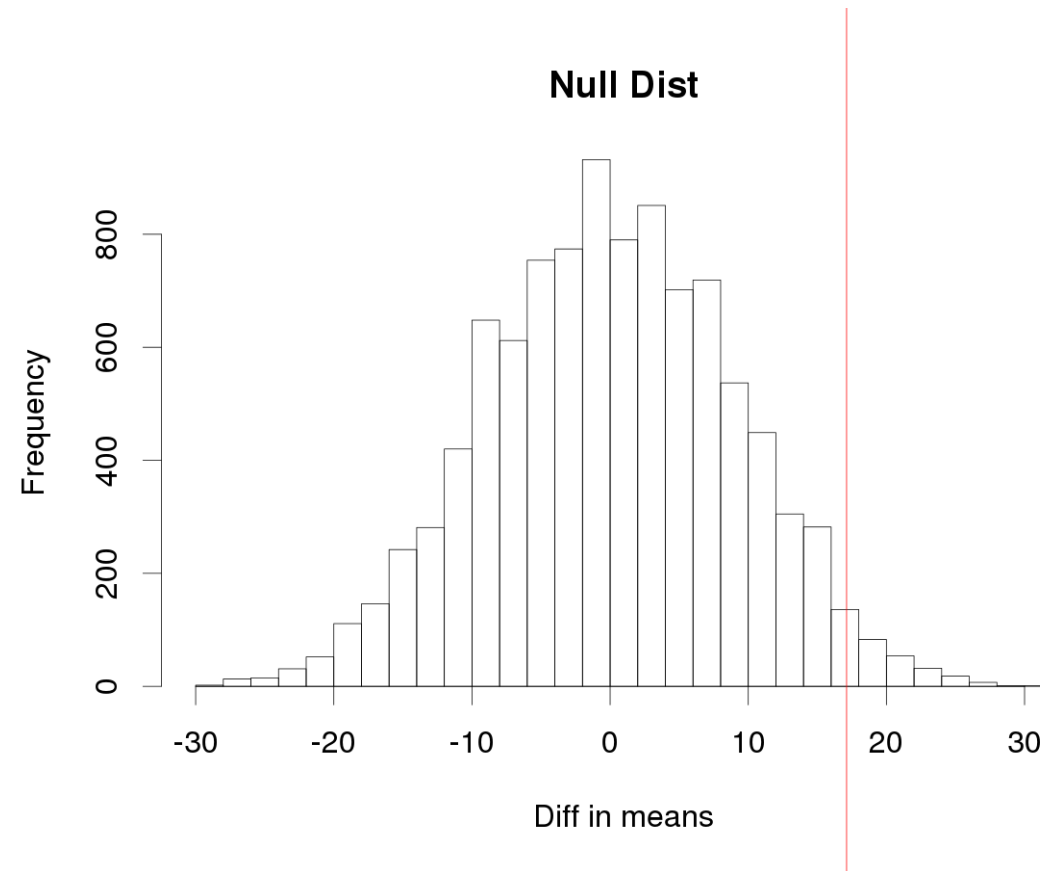
1. Combining the data from both groups together
2. Randomly selecting **11** subjects to simulate the tea drinkers and the remaining **10** subjects to simulate the coffee drinkers
3. Calculating the difference between the means of these two shuffled groups
4. Repeating this process 10,000 times to get a full null distribution

3. Create the null distribution!



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

3. Creating a null distribution



4. p-value = .0255

5. Conclusions?

Implementing a permutation test in Python

Lab 6, part 2: examining whether more triples were hit by the American League or the National League

Any thoughts a priori on whether the AL or NL have a higher number of triples hit on average?

Example 2: Comparing two means

Have baseball games gotten longer in the past 50 years?

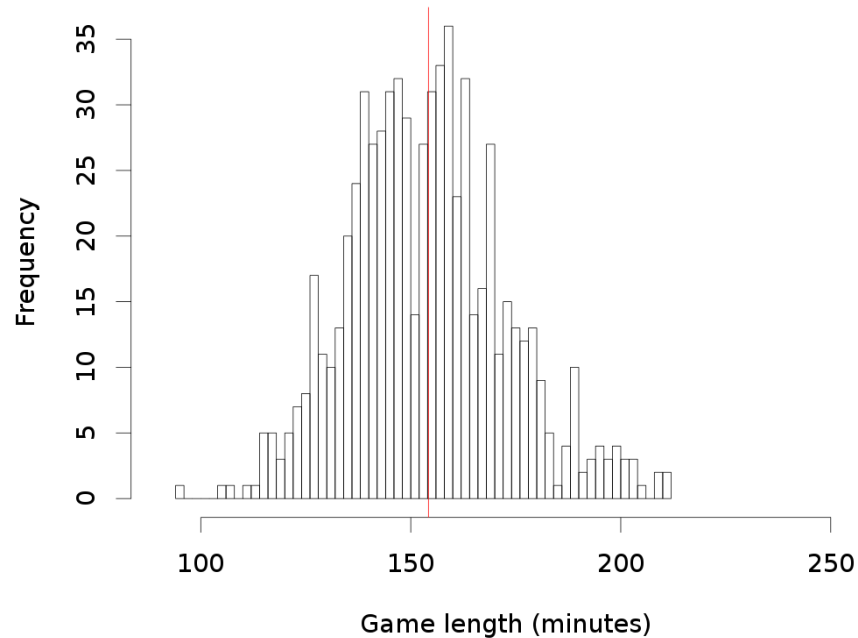
How could we examine this?

- Compare mean lengths of games in 1964 to those in 2014

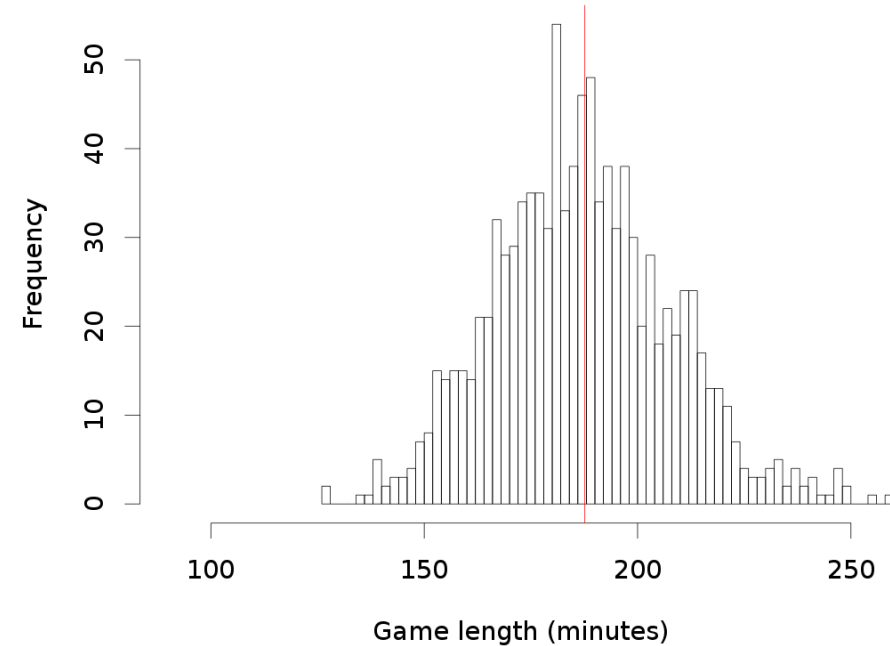
What would be a good first thing to do?

0. Plot the data

1964 game lengths (54 outs)



2014 game lengths (54 outs)



Average game length 1964 is: $\bar{x}_{1964} = 154.21$ minutes
• (based on $n = 684$ games with 54 outs)

Average game length 2014 is: $\bar{x}_{2014} = 187.64$ minutes
• (based on $n = 1021$ games with 54 outs)

1. Null and Alternative Hypotheses

1a. State the null and alternative hypotheses in words

- **Null hypothesis:** Baseball games on average are the same length in 1964 as they are in 2014
- **Alternative hypothesis:** Baseball games on average are longer in 2014 than in 1964

1b. State the null and alternative hypotheses using symbols

- $H_0: \mu_{2014} = \mu_{1964}$ or $\mu_{2014} - \mu_{1964} = 0$
- $H_A: \mu_{2014} > \mu_{1964}$ or $\mu_{2014} - \mu_{1964} > 0$

What do we do next?

- 2. Compute the statistic of interest

What is the statistic of interest?

2. Compute the statistic of interest

Average game length 1964 is: $\bar{x}_{1964} = 154.21$ minutes

- (based on $n = 684$ games with 54 outs)

Average game length 2014 is: $\bar{x}_{2014} = 187.64$ minutes

- (based on $n = 1021$ games with 54 outs)

2. So the statistic of interest is...?

- $\text{observed_stat} = 187.64 - 154.21 = 33.42$ minutes

What do we do next?

- 3. Create a null distribution

Hypothesis tests for two means

3. Calculate the null distribution

- How can we create a null distribution???

One way: under the null hypothesis all games lengths from 1964 and 2014 are equally likely

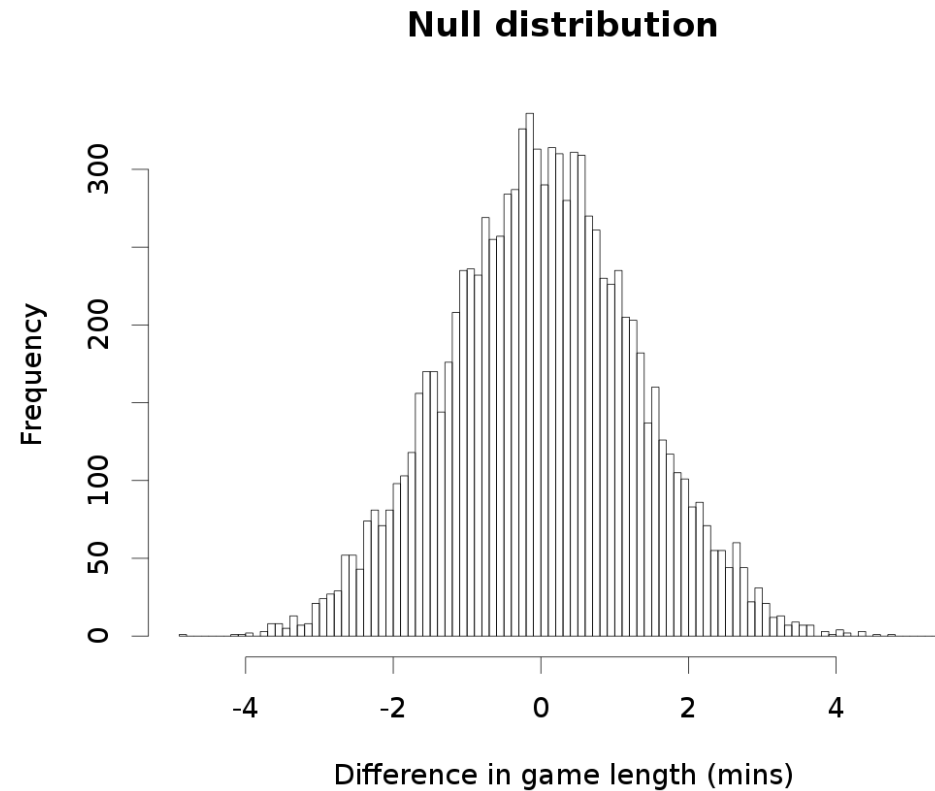
Thus combine all the games lengths from the 1964 and 2014 seasons into one vector

We can then randomly select **684** games to simulate the 1964 season and take the remaining **1021** to simulate the 2014 season

The difference in these means of these 684 and 1021 games gives us one point in the null distribution

If we repeat this 1,000 times we will get a full null distribution

Hypothesis tests for two means



4. Do the results seem statistically significant?

- Observed difference of 33 minutes is not even close to being on this figure
- 5. Conclusions?