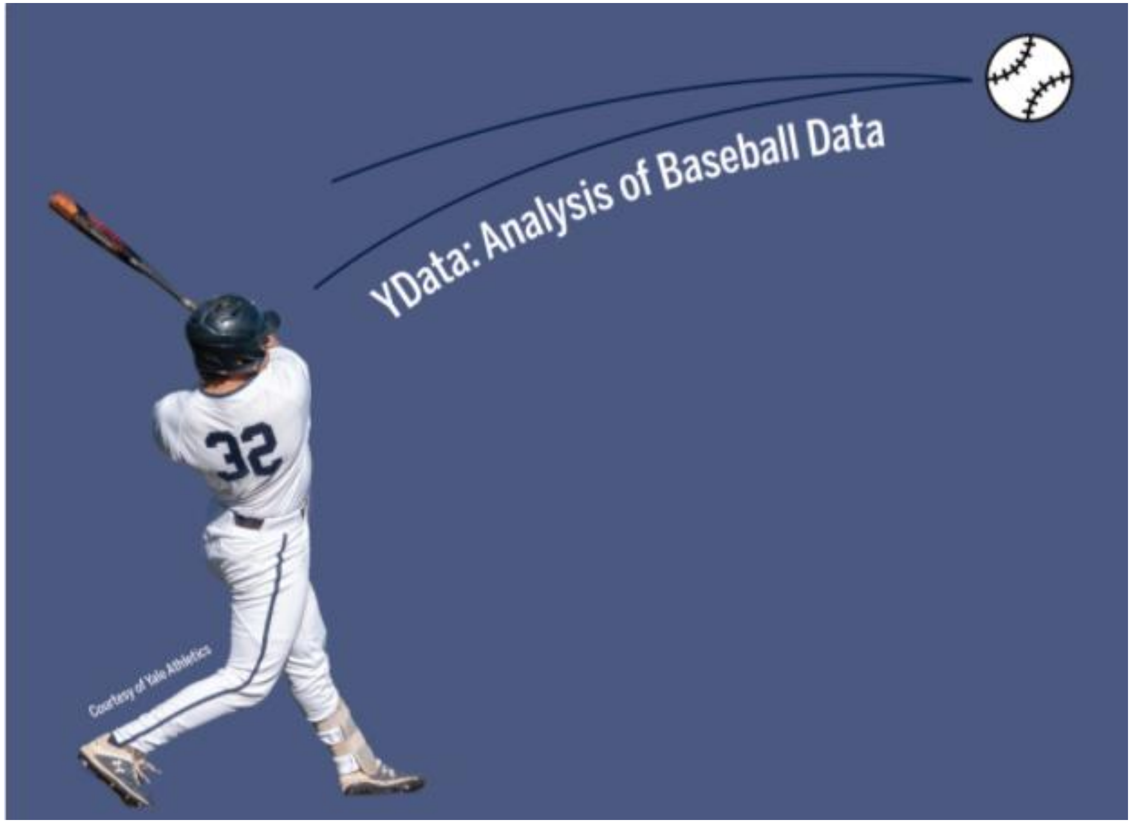
Parametric hypothesis tests and confidence intervals



Overview

Questions about the midterm exam and final projects

Discussion of chapter 6 of Astroball

Binomial distribution and parametric hypothesis tests

The logic of confidence intervals and the bootstrap

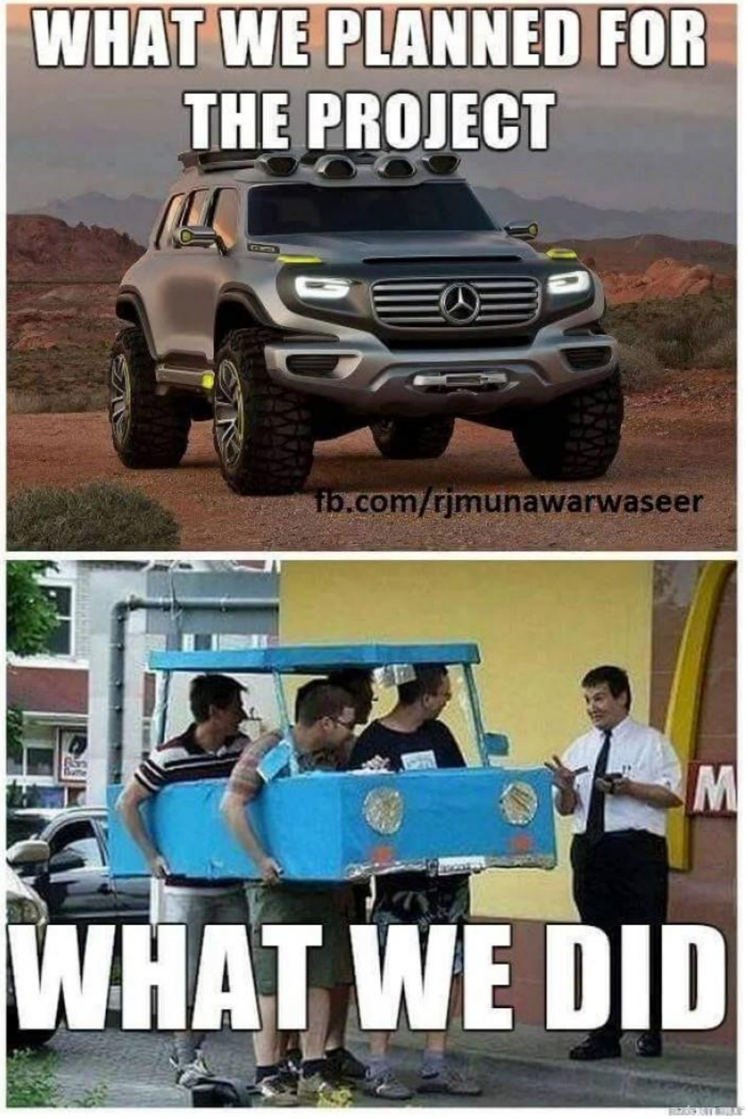
Midterm exam: questions?

The answers have been posted to Canvas

Final projects!

Use the Data Science methods you have learned in this class to answer a question related to baseball!

- Ex 1: Is it true that some baseball players are clutch hitters?
- Ex 2: If all baseball players on a team were Mike Trout, how many runs would they score in a season?



Final projects!

You will create a 5 minute video presentation due on Sunday May 2nd

- Q&A about your project in class on May 5th

All final project written reports are due at 11:30pm on May 13th (last day of reading period)

- Report should be a Jupyter notebook approximately 10 pages long
- A final project Jupyter notebook template has been uploaded to the class GitHub site

The season has started!

American League | National League

AL East

Team	W	L	Pct	GB	Home	Away	L10
Orioles	3	2	.600	-	0-0	3-2	3-2
Blue Jays	3	2	.600	-	0-0	3-2	3-2
Yankees	3	2	.600	-	3-2	0-0	3-2
Red Sox	2	3	.400	1.0	2-3	0-0	2-3
Rays	2	3	.400	1.0	0-0	2-3	2-3

538 baseball predictions

TEAM	DIVISION	TEAM RATING	1-WEEK CHANGE	AVG. SIMULATED SEASON: RECORD	AVG. SIMULATED SEASON: RUN DIFF.	POSTSEASON CHANCES: MAKE PLAYOFFS	POSTSEASON CHANCES: WIN DIVISION	POSTSEASON CHANCES: WIN WORLD SERIES
Dodgers 5-1	NL West	1603	+4	105-57	+228	96%	77%	28%
Yankees 3-2	AL East	1572		97-65	+151	84%	63%	13%
Padres 4-2	NL West	1562	+1	94-68	+121	76%	21%	8%
Astros 5-1	AL West	1561	+9	97-65	+157	87%	76%	11%
Twins 3-2	AL Central	1542	+4	91-71	+106	69%	52%	6%
Mets 1-1	NL East	1540		88-74	+69	55%	37%	5%

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 6
  - Do you remember chapter 6?

Sports Illustrated cover jinx hitting 2014 Astros?

DEADSPIN | Sports News Without Fear, Fav... Shop Subscribe

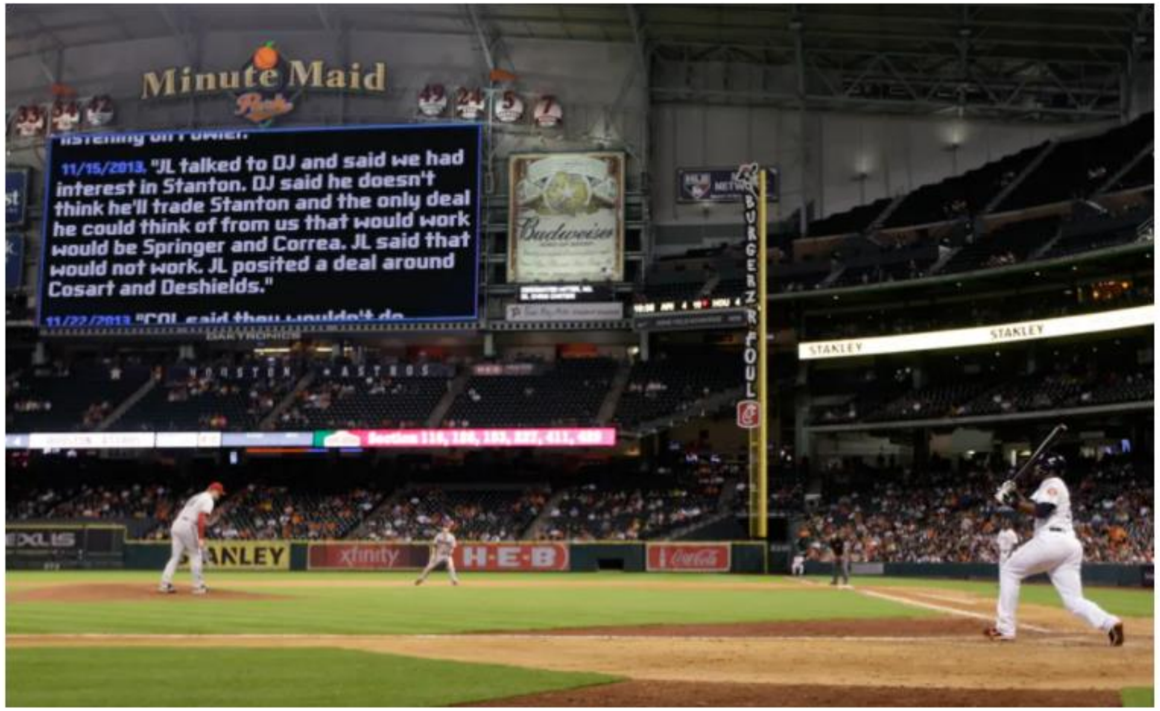
HOME LATEST NBA MLB NFL NHL SOCCER GLOBAL PANDEMIC THE MOURNING AFTER

MLB

Leaked: 10 Months Of The Houston Astros' Internal Trade Talks

Barry Petchesky
6/30/14 1:19PM

17

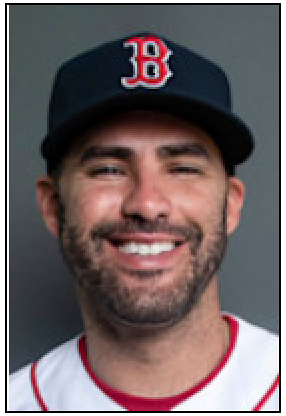


Two years ago, the Houston Astros constructed "Ground Control"—a built-from-scratch online database for the private use of the Astros front office. It is by all accounts a marvel, an easy-to-use interface giving executives instant access to player statistics, video, and communications with other front offices around baseball. All it needs, apparently, is a little better password protection.

What Happened to the Houston Astros' Hacker?

As the 2017 spring season wore on, it became increasingly difficult to score against one of the four clubs in the softball league at the federal prison camp in Cumberland, Md. The Dogs played the standard 10 fielders, but it seemed as if they had twice as many….

The bail project



J.D. Martinez

Positions: Outfielder and Designated Hitter

Bats: Right • **Throws:** Right

6-3, 230lb (190cm, 104kg)

Team: Boston Red Sox (majors)

More bio, uniform, draft, salary info ▼

SUMMARY	WAR	AB	H	HR	BA	R	RBI	SB	OBP	SLG	OPS	OPS+
Career	24.0	4183	1213	238	.290	622	738	25	.354	.530	.883	135

Martinez acknowledged that every athlete who has ever been cut from a team feels aggrieved. What hurt most was how the Astros had done it. "They had all this data, all these nerds and geeks, and I think they forgot that at the end of the day, everyone is still human," he said. "And a human can adapt, and a human can adjust."

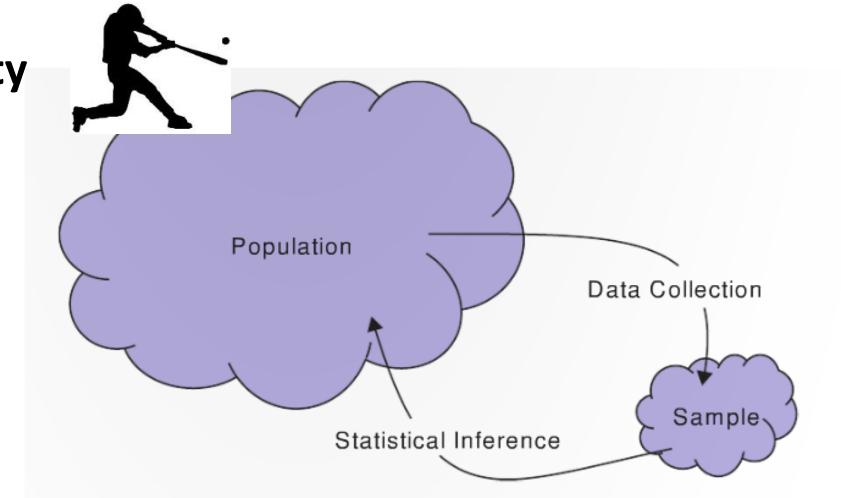
Statistical inference

Statistical inference: use <u>*sample*</u> of data to deduce properties of an underlying <u>*population*</u> or stochastic <u>*process*</u>

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

Ability



Performance: Hit, Out, Hit, Out, …

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

For a proportion the ***population*** is a ***parameter*** denoted $\boldsymbol{\pi}$

$\hat{p}$ is a ***point estimate*** of π

- i.e., $\hat{p}$ our best guess for the value of π

For a mean the ***population*** is a ***parameter*** denoted $\boldsymbol{\mu}$

$\bar{x}$ is a ***point estimate*** of μ

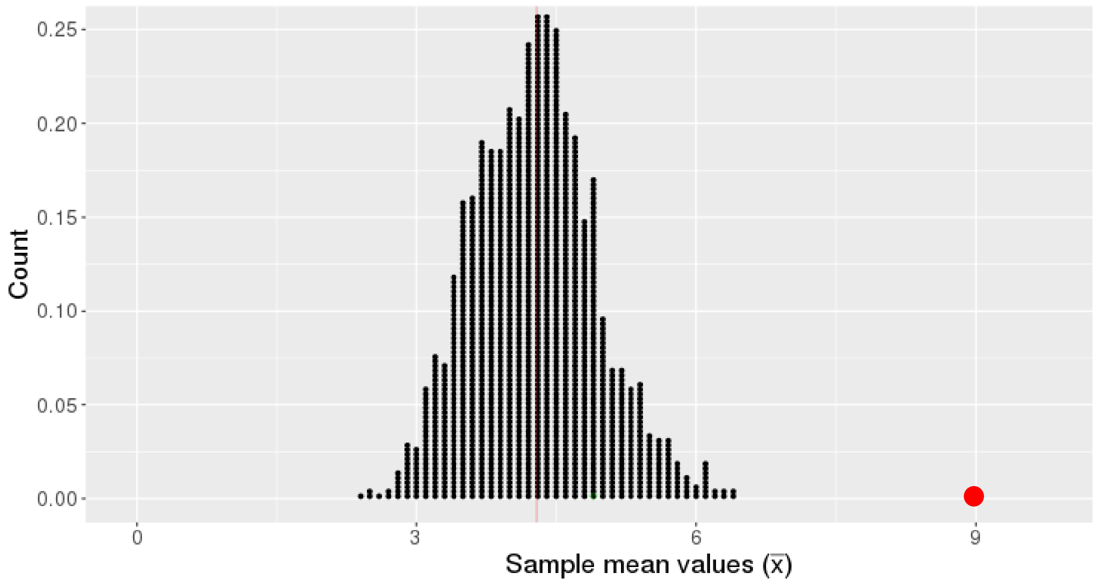
- i.e., $\bar{x}$ our best guess for the value of μ

Hypothesis testing: basic logic

We start with a claim (null hypothesis) about a population parameter

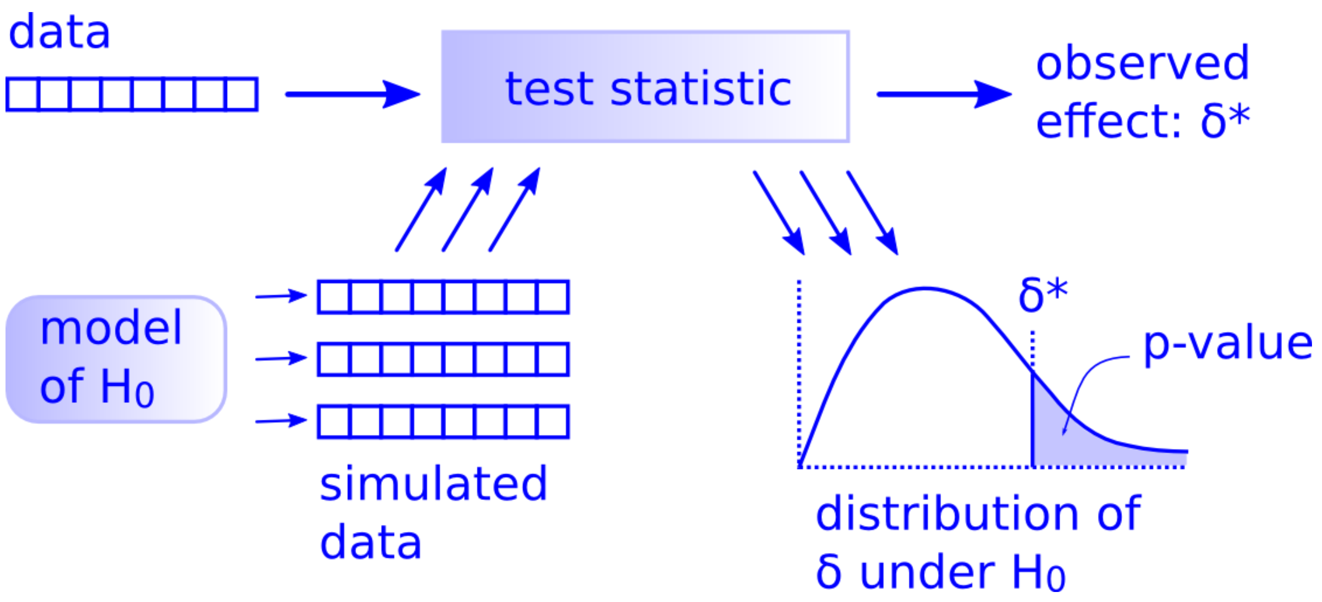
- E.g., H_0: μ = ~~[illegible]2~~

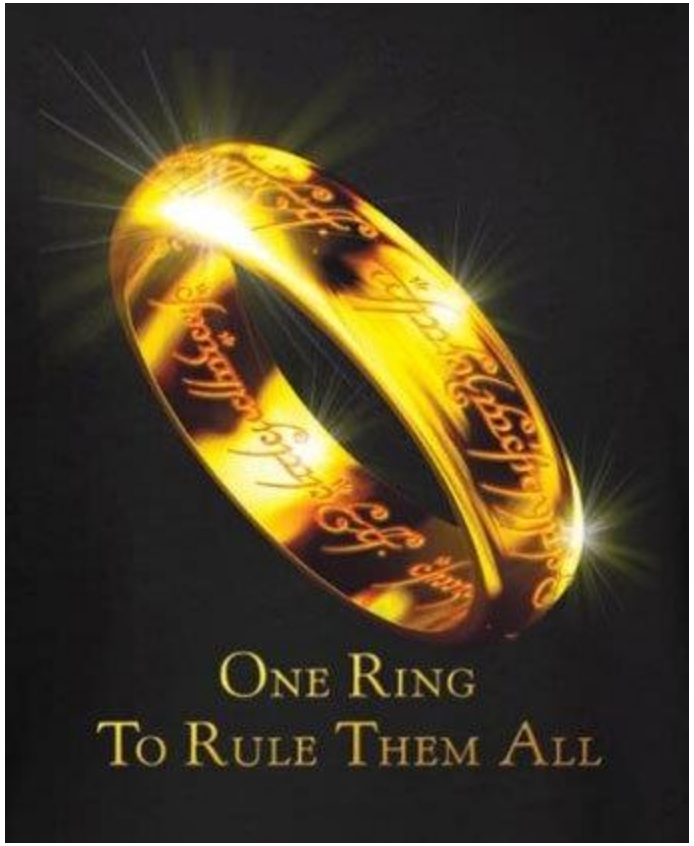
This claim implies we should get a certain distribution of statistics



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

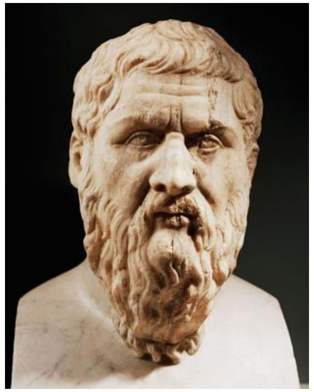
There is only one hypothesis test!





Just follow the 5 hypothesis tests steps!

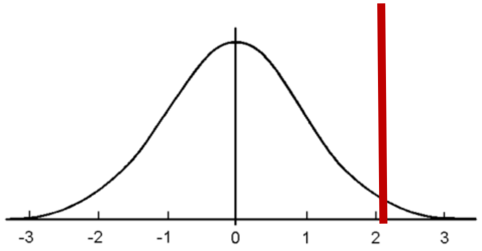
Five steps of hypothesis testing



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

2. Calculate the actual observed statistic

3. Create a **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

Example: How good is A-Rod's really?

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

- Let us denote this <u>observed performance</u> 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?

1. State null and alternative hypotheses

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

1. We can state this question in terms of a ***null*** and ***alternative*** hypothesis

- H_0: A-Rod's ability is $\pi = .300$
- H_A: A-Rod's ability is $\pi > .300$

2. Calculate the observed statistic

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

1. We can state this question in terms of a ***null*** and ***alternative*** hypothesis
   - H_0: A-Rod's ability is $\pi = .300$
   - H_A: A-Rod's ability is $\pi > .300$

2. Our observed statistic is:
   - $\hat{p} = .353$

3. Create the null distribution

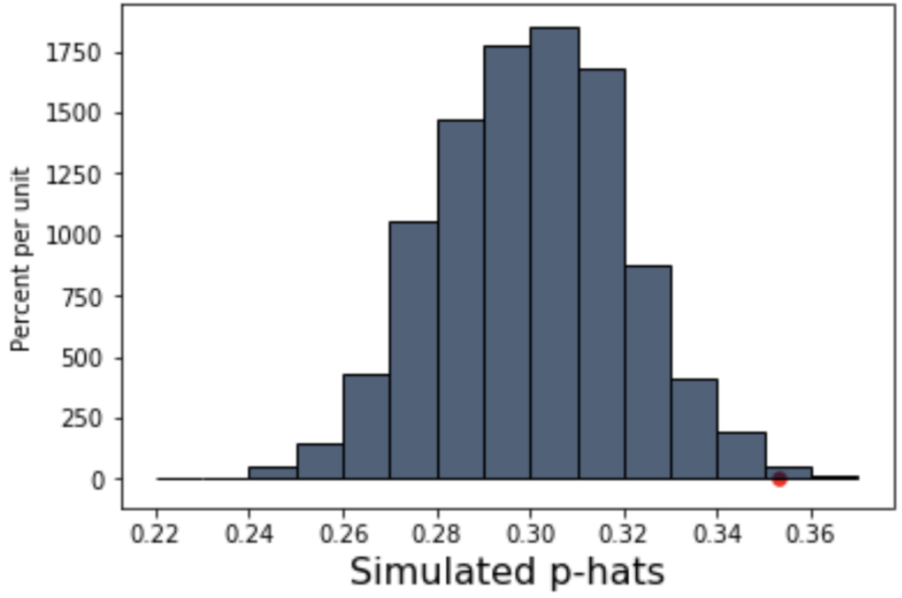
3. We create a "null distribution" using Python by generating statistics consistent with the null hypothesis

- π = .300 based on 529 plate appearances

```
def generate_flip_proportion_heads(num_flips, prob_heads):
    return sample_proportions(              ).take(0)
```

What is the value of the first argument?

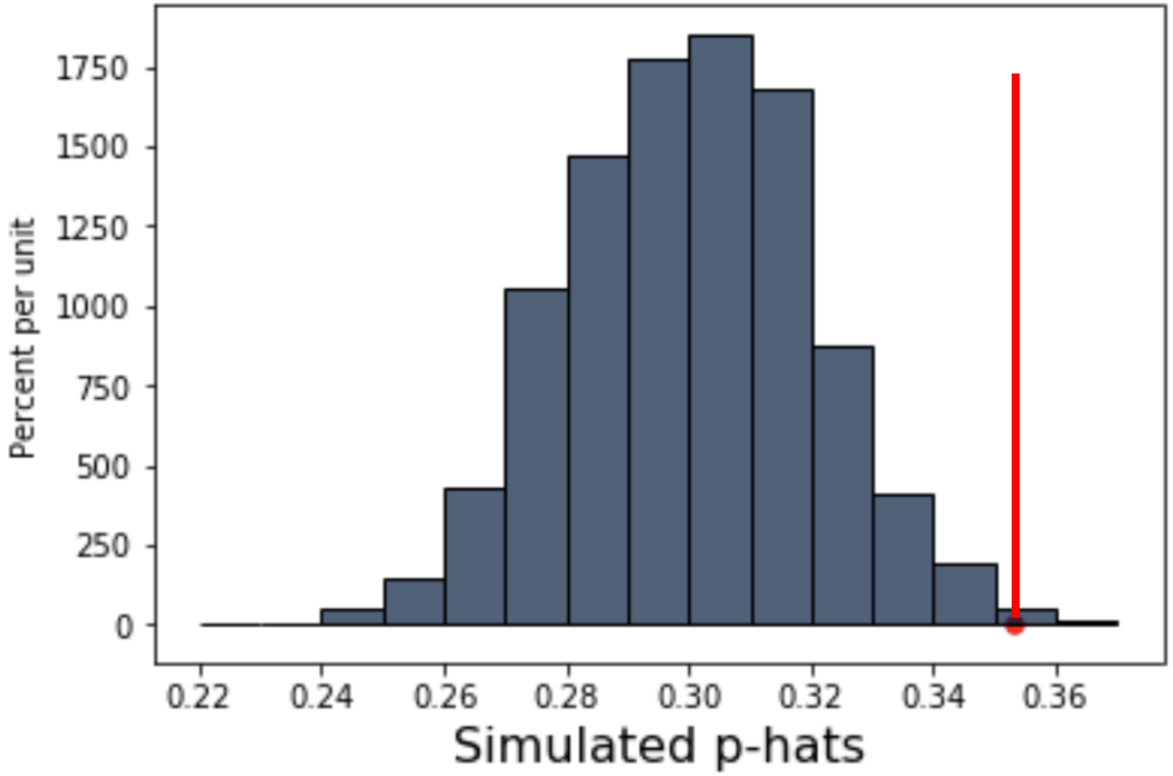
What is the value of the second argument?



4. Calculate the p-value

4. We can then see that it is unlikely that our observed statistic ($\hat{p}$ = .353) came from the null distribution

p-value = 0.0047



5. Make a decision

The p-value = 0.0047, what is your decision?

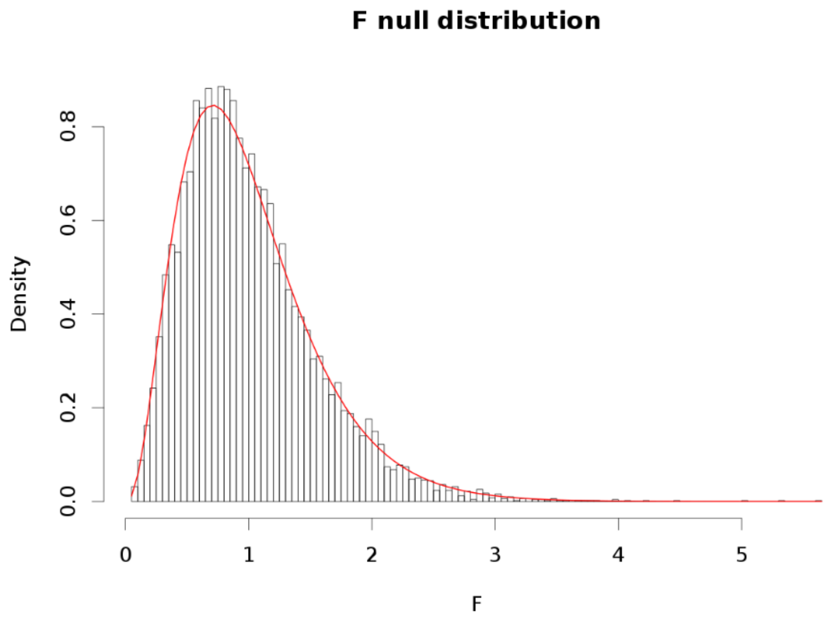
Parametric hypothesis tests

In parametric hypothesis tests, one uses mathematical functions (probability mass/density functions) for the null distribution

- i.e., we use probability functions instead of using simulations for the null distribution

The same logic and 5 steps of hypothesis testing are used

- Only real difference is step 4



Parametric probability models

We have explored probability using:

- Probability rules to calculate the probability of an event
- Dice/spinners

We often use mathematical formulas to calculate the probability of different events

- $\Pr(\mathbf{X} = x) = f(x; \theta)$

Random variates

We can think of a **random variate** as a random number

- Random variates are denoted with capital letters, e.g., X
  - We can think of it as: “what is behind the door marked with an X”

We can use probability distributions to assess the probability that a random variate X will have a value between two other numbers

- Notation: $\Pr(a < X < b)$

Random variates

Random variates can be either:

- Discrete: X takes on integer values
- Continuous: X takes on real values

Same properties of probability distributions apply:

- $Pr(a < X < b) \geq 0$
- $\Sigma Pr(X = x_i) = 1$

Bernoulli Distribution

Models the probability of two outcomes:

- Success: X = 1
- Failure: X = 0
- E.g., getting a head (X = 1) or a tail (X = 0) for flipping a coin

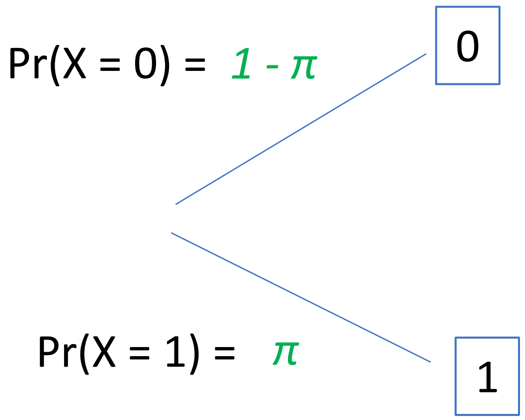
Model has one parameter π, which is the probability of getting a 1

- E.g., the probability of getting head on a coin flip
- Pr(X = 1) = ?   Pr(X = 0) = ?

Bernoulli Distribution

Can you draw the Bernoulli distribution as a tree diagram?

- Pr(X = x; *π)*



Bernoulli Distribution

Probability mass functions is a function that gives the probability of each outcome in the sample space of a *discrete distribution*

For Bernoulli Distributions we can write the probability mass function as:

Parameter π

$$f(k;\pi) = \begin{cases} \pi & \text{if } k = 1 \\ 1 - \pi & \text{if } k = 0 \end{cases} \qquad f(k;\pi) = \pi^k (1-\pi)^{1-k} \quad \text{for } k \in \{0, 1\}$$

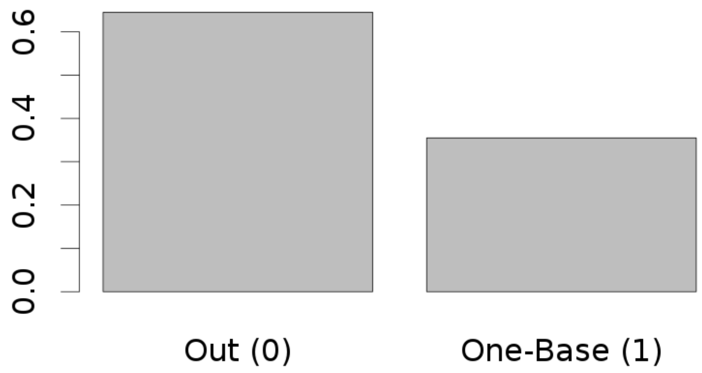
We can also plot the probability mass function

- If $\pi = .5$ what would it look like?
- If $\pi = .9$ what would it look like?

Bernoulli Distribution

Can you think of baseball example that can be modeled using a Bernoulli distribution?

- Example: The probability that a player gets on base for a given at bat
- If we were modeling A-Rod in 2012, what would a good estimate of π be?
  - OBP = .353, so a good number to use would be π = .353



Binomial distribution

Models the probability of having k successes out of the n trials

- Probability of success on each trial is π
- E.g., this models flipping a coin k times

Example: if a player comes to bat 4 times in a game, what is the probability of getting on-base k times

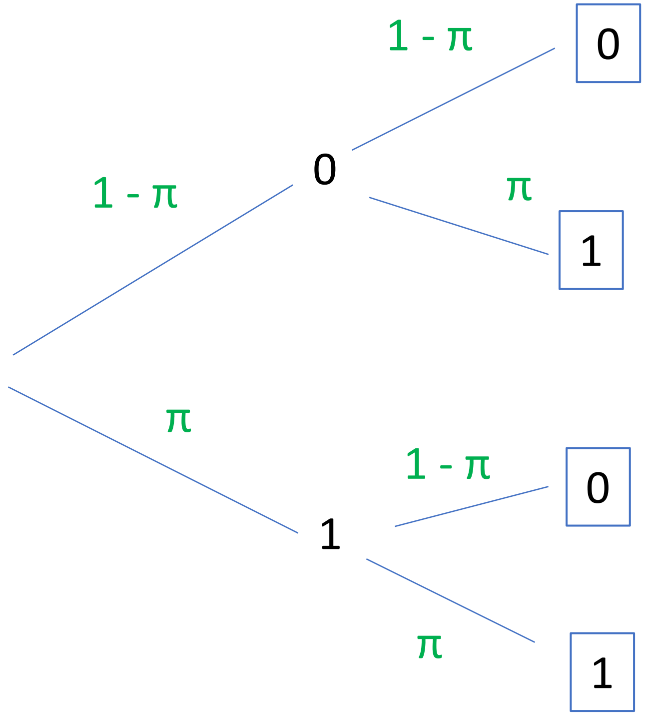
- What is the sample space?

Assumes the same probability of getting on-base each plate appearance (π)

- Are the assumptions reasonable for this model?
    - No streakiness

Tree diagrams of binomial distribution

The binomial distribution can be thought of sequences of Bernoulli trials.
For n = 2 trials we have:



$$Pr_{Bernoulli}(X_1 = 0; \pi) \cdot Pr_{Bernoulli}(X_2 = 0; \pi) = (1 - \pi)^2$$

$$= Pr_{Binomial}(X = 0; n = 2, \pi)$$

$$Pr_{Bernoulli}(X_1 = 0; \pi) \cdot Pr_{Bernoulli}(X_2 = 1; \pi) =$$

$$Pr_{Bern}(X_1 = 1; \pi) \cdot Pr_{Bern}(X_2 = 0; \pi) = \pi \cdot (1 - \pi)^2$$

$$Pr_{Binomial}(X = 1; \quad n = 2, \pi) = \quad 2 \cdot \pi \cdot (1 - \pi)^2$$

$$Pr_{Bernoulli}(X_1 = 1; \pi) \cdot Pr_{Bernoulli}(X_2 = 1; \pi) = \pi^2$$

$$= Pr_{Binomial}(X = 2; n = 2, \pi)$$

Binomial distribution

Probability of getting ***k*** successes out of ***n*** trials is:

Parameters: π and n

$$Pr(X = k)$$

where
$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

What values can X take on (i.e., what is the sample space)?

- k ranges from 0, 1, …, n

n choose k

N choose k function tells us how many ways there are to pick ***k*** items out of ***n*** total

- i.e., how many unique sets there are of k items

Q: How many ways are there to choose 3 things out of a total of 8?

A: $(8 \cdot 7 \cdot 6)/ (3 \cdot 2) = 56$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Binomial distribution example 1

In the 2013 season, Koji Uehara struck out around 33% of the batters he faced

If he pitched to 5 batt … stribution to calculate the probabil … ?

$Pr(X = k$ … $)^{n-k}$

$\binom{n}{k} = \frac{}{k!(}$ …

What are n, k and π?

Binomial distribution

Koji Uehara typically struck out about 33% of the batters he faced

If he pitched to 5 batters in a game, use the binomial distribution to calculate the probability he would strike out 2 of them?

What are the parameters?

n = 5, π = .333 k = 2

$$Pr(X = k) = f(k; n, \pi) = \binom{n}{k} \pi^k (1 - \pi)^{n-k}$$

$$Pr(X = 2) = f(k; 5, .333) = \binom{5}{2} .333^2 (1 - .333)^3$$

= .329

```
from scipy.stats import binom
binom.pmf(k = 2, n = 5, p = 1/3)
```

Binomial distribution

Suppose a Uehara pitched to 20 batters.

How could we calculate the probability that he would get more than 10 strike outs?

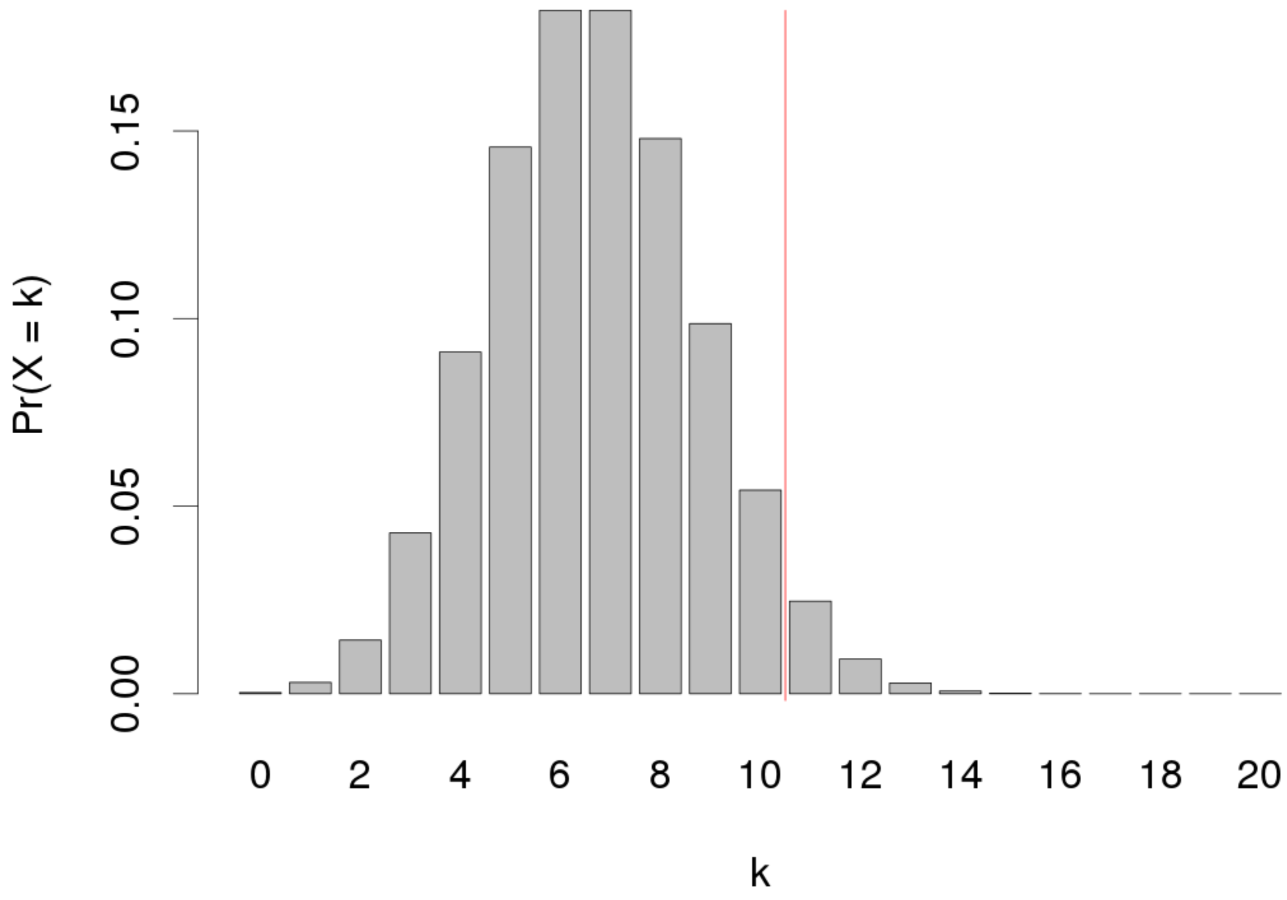
$$Pr(X > 10) = \sum_{k=11}^{20} f(k; n = 20, \pi = .333)$$

$$= \sum_{k=11}^{20} \binom{20}{k} .333^k (1 - .333)^{20-k}$$

= .0376

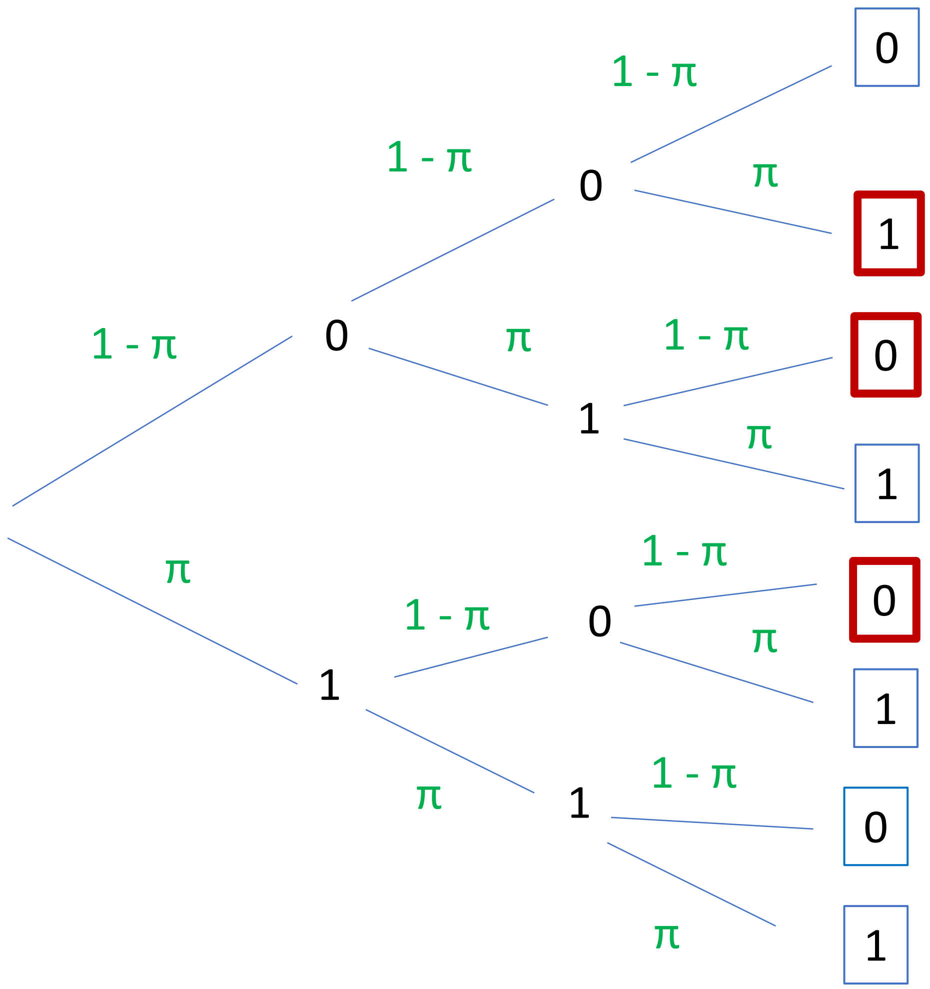
```
np.sum(binom.pmf(np.arange(11, 21), n = 20, p = 1/3))
```

Binomial distribution



Side note: Notice this distribution looks like a normal distribution

Verify binomial distribution agrees with tree diagrams



Checking for:

$Pr(X = 1; n = 3, \pi)$

What are the possible sequences for X = 1?

- 0, 0, 1
- 0, 1, 0,
- 1, 0, 0

$$Pr(X = k) = \binom{n}{k}\pi^k(1-\pi)^{n-k}$$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Back to how good is A-Rod's really?

<u>Question</u>: Is it plausible that A-Rod's OBP ability π was really .300 and he was just got lucky to get a p̂ of .353?

1. We can state this question in terms of a ***null*** and ***alternative*** hypothesis
    - H_0: A-Rod's ability is $\pi = .300$
    - H_A: A-Rod's ability is $\pi > .300$

2. Our observed statistic is:
    - $\hat{p} = .353$

3. Create the null distribution

For lab 7 problem 1 you will use a binomial distribution to redo the hypothesis test to see if his ability was really 0.300

Other parametric hypothesis tests

Many other commonly used hypothesis tests involve using parametric null distributions

- E.g., t-tests, ANOVAs, etc.

These null distributions are based on mathematical derivations (similar to what we did to get the binomial distribution).

Often "assumptions" need to be made to determine particular parametric distributions

- E.g., assume the data is normally distributed to use a t-distribution for quantitative data

While these tests are often robust to violations to these assumptions, it is possible to come to an incorrect conclusions if the assumptions are badly violated.

Parametric hypothesis tests: the t-test

An example of a parametric probability distribution we can use to test if two population means are the same is a t-test.

To run a t-test we use a t-statistic:

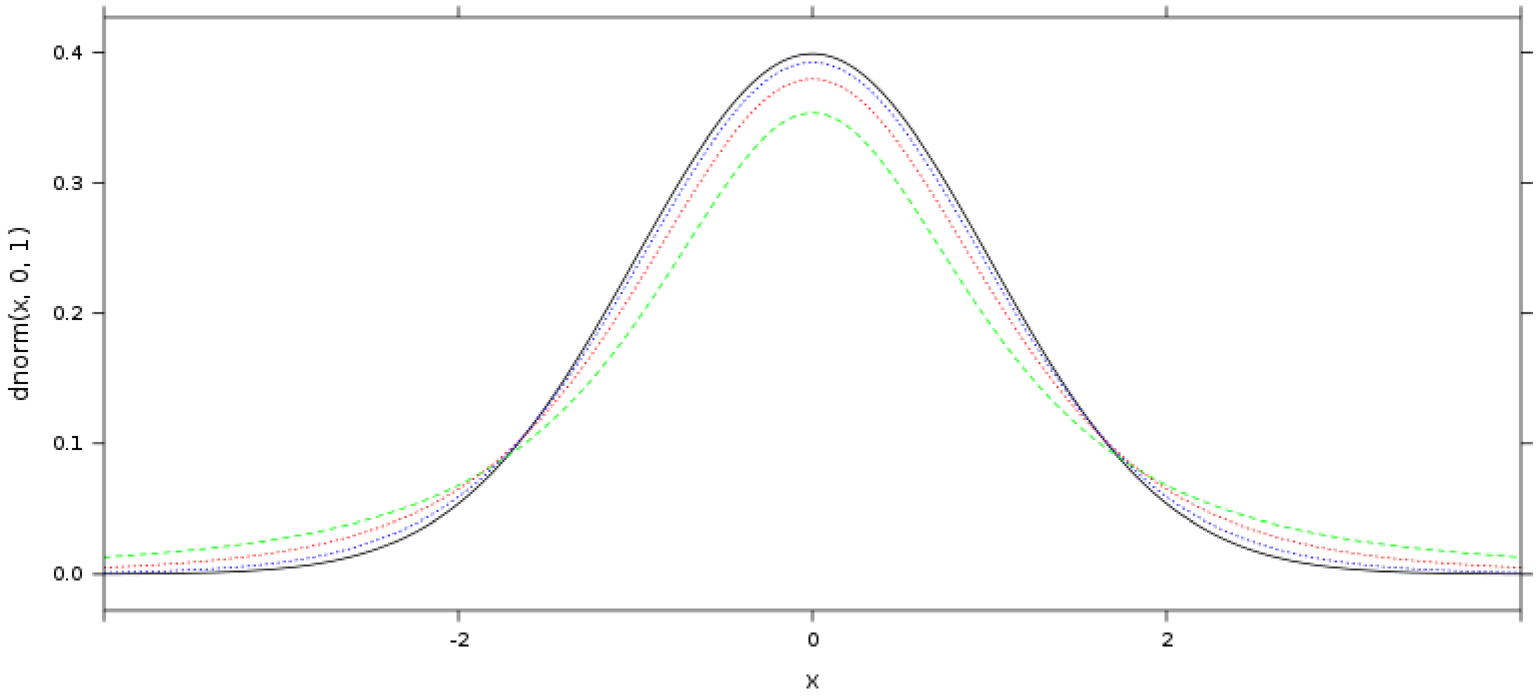
- $\overline{x}_1$ and $\overline{x}_1$ are the sample means
- s_1^2 and s_2^2 are the sample variances
- n_1 and n_2 are the sample sizes

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

We can use a t-distribution for the null distribution

- An appropriate t-distribution to use as the parameter '*degrees of freedom*' set to the minimum of n_1 -1 and n_2 -1

t-distributions



N(0, 1), df = 2, df = 5, df = 15

Homework problem 7, part 2

You will rerun the analyses comparing whether the mean number of triples hit in the AL and the NL are the same using a t-test

Confidence intervals

In hypothesis test, we test whether it is plausible that parameter is equal to a particular value

- E.g., $\pi = .300$

Rather then just testing whether a parameter is equal to a particular value, it could be useful to come up with a range of plausible values for a parameter

A ***confidence interval*** aims to do just that!

Confidence Intervals

A **confidence interval** is an interval computed by a method 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

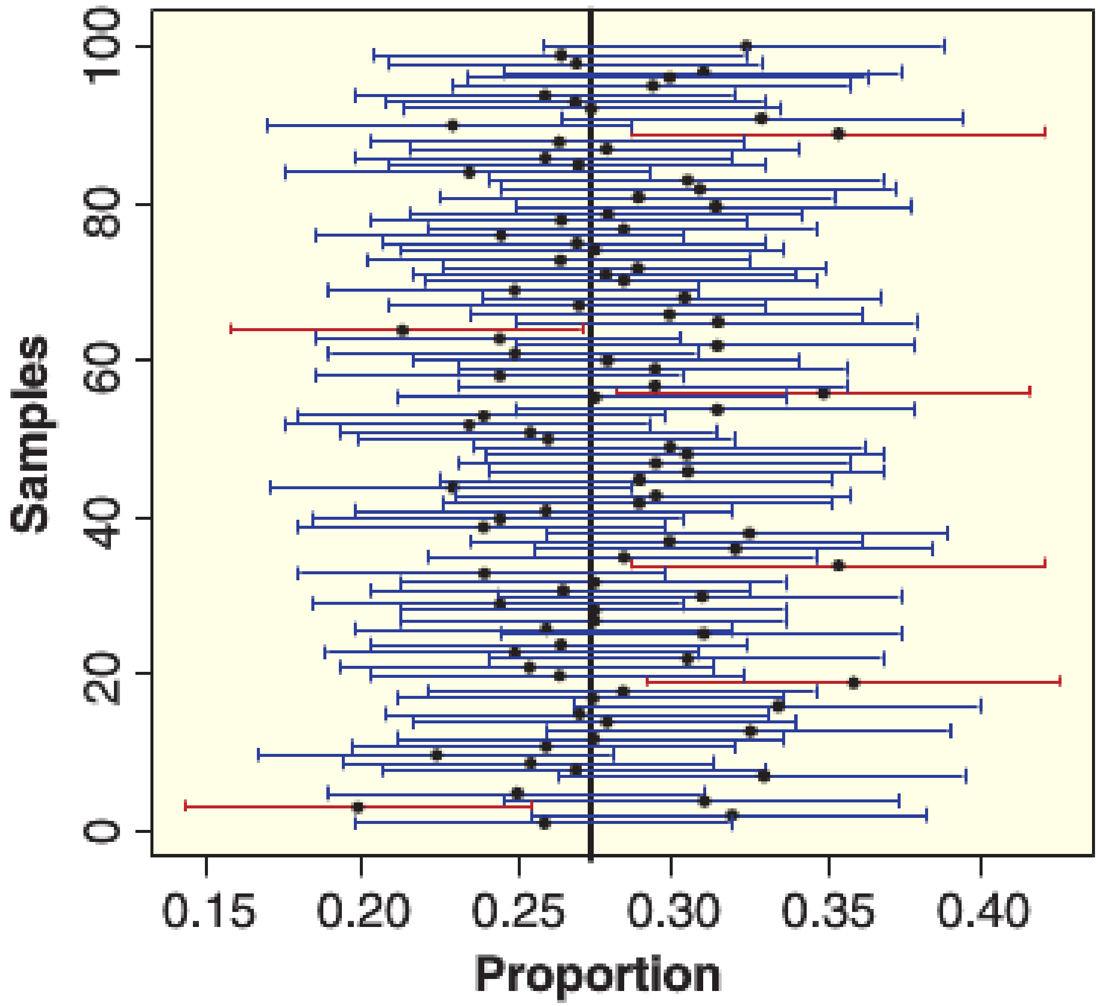
We toss intervals at it

95% of those intervals capture the parameter

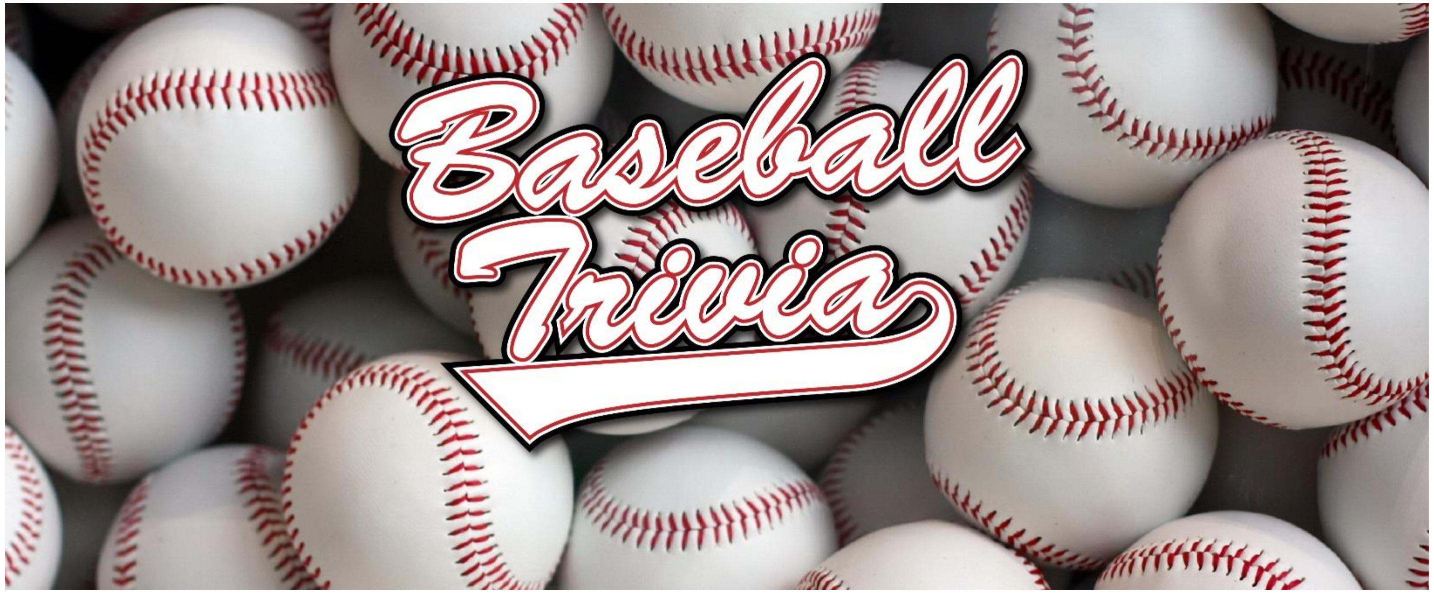
Confidence Intervals

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

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



Time to be a 90% confidence interval estimator



I will ask you 10 questions with numerical answers

You need to come up with ranges of plausible values (intervals) such that 9 out of 10 intervals contain the true answer

Baseball interval estimates

Question 1: How many stiches does a major league baseball have?

Question 2: What year was the oldest baseball stadium that is still in use built?

Question 3: In hours and minutes, how long was the longest baseball game?

Baseball interval estimates

Question 4: Cy Young award is given to the best pitcher of the year. It is named the pitcher Cy Young. How many wins did Cy Young have over his career?

Question 5: How many times did Roger Clemens win the Cy Young award?

Question 6: What is the most consecutive games that a player has played in?

Question 7: In what year were shin guards first used by Major League catchers?

Baseball interval estimates

Question 8: In today's dollars (i.e., inflation adjusted dollars) how much was Babe Ruth paid in the season he made the most money?

Question 9: In pounds, how much did the lightest baseball player weigh?

Question 10: Over Derek Jeter's career 260 home runs, how many were grand slams?

Answers…

Baseball interval estimates

Question 1: How many stiches does a major league baseball have?

- 108

Question 2: What year was the oldest baseball stadium that is still in use built?

- 1912
- Fenway Park

Question 3: In hours and minutes, how long was the longest baseball game?

- 8 hours, 6 minutes.
- 1984 game between the white sox and brewers

Baseball interval estimates

Question 4: Cy Young award is given to the best pitchers of the year. It is named the pitcher Cy Young. How many wins did Cy Young have over his career?

- 511

Question 5: How many times did Roger Clemens win the Cy Young award?

- 7 times

Question 6: What is the most consecutive games that a player has played in?

- 2,632 games
- Cal Ripkin: April 30th 1982 to September 19th 1998

Question 7: In what year were shin guards first used by Major League catchers?

- 1907. First used by Roger Bresnahan

Baseball interval estimates

Question 8: In today's dollars (i.e., inflation adjusted dollars) how much was Babe Ruth paid in the season he made the most money?

- $1.1 million
- When he was asked how he deserved to make more than the U.S. president, he replied, "I had a better year."

Question 9: In pounds, how much did the lightest baseball player weigh?

- 65 pounds. Eddie Gaedel

Question 10: Over Derek Jeter's career 260 home runs, how many were grand slams?

- One. It came against Cubs pitcher Joe Borowski on June 19th 2005.

How did we do?

Did everyone have the right answers in their intervals 9 out of 10 times?