

STAT 400 - Final Project

Using Monte Carlo to Simulate the 2019 MLB Postseason

Group 3 - Quinn Johnson, Jonathan Olds, Andrew Wiggans

12/12/2019

Introduction & Motivation

About Us

- ▶ Upperclassmen with statistics majors
- ▶ Planning to pursue careers in the sports industry

Motivation

- ▶ Expand our knowledge of statistics and computing within the context of a specific and relevant sports-related problem

Presentation Outline

Background

- ▶ Description of MLB Postseason play
- ▶ Common Terms

Original Study

- ▶ Description
- ▶ Methodology
- ▶ Results

Our Project

- ▶ Description + Broader comparison to original project
- ▶ Methodology
- ▶ Results

Further Study/Limitations

Background Information

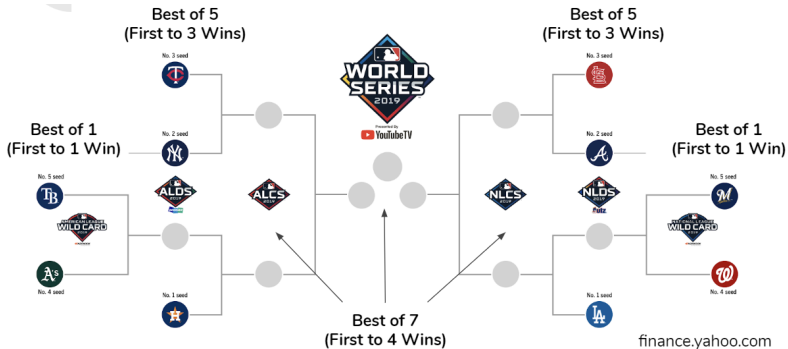
MLB Regular Season Format

- ▶ 30 teams (divided into 2 leagues) play 162 games each
- ▶ The top 5 teams by record in each league advance to the postseason
 - ▶ 3 division winners, 2 wild card teams

MLB Postseason Format

- ▶ 4 Round Single- Elimination Tournament
 - ▶ *Wild Card, League Division Series, League Championship Series, World Series*
 - ▶ Wild Card (Single Match)
 - ▶ League Division Series (Best of 5)
 - ▶ League Championship Series (Best of 7)
 - ▶ World Series (Best of 7)

Example of MLB Postseason Bracket



Description of Terms

- ▶ Single = Batter reaching 1st base on a hit
- ▶ Double = Batter reaching 2nd base on a hit
- ▶ Triple = Batter reaching 3rd base on a hit
- ▶ Home Run = Batter scoring a run on a hit
- ▶ Walk = Batter reaching 1st base via the pitcher throwing 4 balls, or the batter being hit by a pitch
- ▶ Run = Anytime a batter or runner reaches home plate safely
- ▶ AL/NL Wild Card = American/National League Wild Card
- ▶ ALDS/NLDS = American/National League Division Series
- ▶ ALCS/NLCS = American/National League Championship Series
- ▶ World Series = Championship Round of MLB Postseason
- ▶ $P(A)$ = Probability of Event A occurring

Original Study - Description

- ▶ “Did The Best Team Win? Analysis of the 2010 Major League Baseball Postseason Using Monte Carlo Simulation” published in 2012
 - ▶ Aimed to estimate the probability with which each of the 8 teams in the 2010 MLB Postseason could be expected to win the World Series
 - ▶ *Prior to 2012, the Wild Card round did not exist; only 8 teams qualified for the postseason each year*
 - ▶ Additionally looked to compare study results with the actual results of the postseason that year

Original Study - Methodology

- ▶ Utilized 2010 regular season batting & pitching statistics of the 8 postseason teams, as well as the log5 method to calculate probabilities of different events occurring.
 - ▶ $P(\text{Single})$, $P(\text{Double})$, $P(\text{Triple})$, $P(\text{Home Run})$, $P(\text{Walk})$
- ▶ These probabilities were then included in a model designed to simulate each possible matchup of each round of the postseason
 - ▶ Each possible series matchup simulated 2,000,000 times, probability of either team winning was calculated

Original Study - Results

Probabilities of each team
winning a particular round

Team	Wins
Rangers	90
Yankees	95
Rays	96
Twins	94
Giants	92
Braves	91
Phillies	97
Reds	91

of regular season wins
for each team

Team	P(Win LDS)	P(Win LCS)	P(Win WS)
Rangers	.602	.374	.233
Yankees	.564	.275	.146
Rays	.398	.186	.092
Twins	.436	.165	.063
Giants	.516	.314	.161
Braves	.484	.276	.136
Phillies	.554	.245	.104
Reds	.446	.165	.064

- Rangers & Giants most likely to reach World Series, correctly reflected actual events.
 - Giants beat Rangers (4-1) in 2010 WS.
- However, not what might would be predicted given regular season wins

Our Study - Description

- ▶ We aimed to replicate the methods of the 2010 study and apply them to simulate the 2019 MLB Postseason.
 - ▶ Goals:
 - 1 - Determine the probabilities that each team in the 2019 postseason would win the World Series
 - 2 - Compare these results to the actual events that unfolded in October 2019.
 - ▶ Adjusted to include 10 postseason teams, instead of 8 from the original study

Methodology - Collecting & Cleaning Data

- ▶ Gathered all data from BaseballReference.com
 - ▶ Team pitching data, team batting data, American League batting data
- ▶ Created lineups of 9 batters for each team
 - ▶ 8 fielders & Pitcher (for National League) or DH (American League)
- ▶ Selected each team's top 4 starting pitchers and combine top 5 relievers into one, average reliever

Methodology - Collecting & Cleaning Data (graphic only)

Batting Lineup

George Springer

Jose Altuve

Michael Brantley

Alex Bregman

Yuli Gurriel

Yordan Alvarez

Carlos Correa

Robinson Chirinos

Josh Reddick

Pitching Rotation

Justin Verlander

Gerrit Cole

Wade Miley

Brad Peacock

Brad Peacock (R)

Chris Devenski (R)

Will Harris (R)

Josh James (R)

Hector Rondon (R)

Methodology - Collecting & Cleaning Data (Cont.)

Calculated individual probabilities for batters:

- ▶ $P(\text{Walk}) = (\# \text{ Walks} + \# \text{ HBP}) / \text{Plate Appearances}$
- ▶ $P(\text{Single}) = (\# \text{ Hits} - (\# \text{ Doubles} + \# \text{ Triples} + \# \text{ Home Runs})) / \# \text{ Plate Appearances}$
- ▶ $P(\text{Double}) = \# \text{ Doubles} / \text{Plate Appearances}$
- ▶ $P(\text{Triple}) = \# \text{ Triples} / \text{Plate Appearances}$
- ▶ $P(\text{Home Run}) = \# \text{ Home Runs} / \# \text{ Plate Appearances}$

Similar for pitchers, but lower values are better

Log5 Method of Creating “Simulation-Ready” Probabilities

$$P(\text{Single}) = \frac{\frac{P(\text{Single})_{\text{Batter}} \cdot P(\text{Single})_{\text{Pitcher}}}{P(\text{Single})_{\text{LgAvg}}}}{\frac{P(\text{Single})_{\text{Batter}} \cdot P(\text{Single})_{\text{Pitcher}}}{P(\text{Single})_{\text{LgAvg}}} + \frac{(1 - P(\text{Single})_{\text{Batter}})(1 - P(\text{Single})_{\text{Pitcher}})}{1 - P(\text{Single})_{\text{LgAvg}}}}$$

Where:

- ▶ $P(\text{Single})_{\text{Batter}}$ is the probability that the batter hits a single in any given plate appearance
- ▶ $P(\text{Single})_{\text{Pitcher}}$ is the probability that the pitcher allows a single in any given plate appearance
- ▶ $P(\text{Single})_{\text{LgAvg}}$ is the league-wide probability of a single occurring in any given plate appearance

Log5 Example

Name (Pitcher)		p1B	p2B	p3B	pHR	pBB
Stephen Strasburg		0.1272295	0.0297265	0.0059453	0.0285375	0.078478
Name (Batter)		p1B	p2B	p3B	pHR	pBB
Cody Bellinger		0.1301059	0.0514372	0.0045386	0.0711044	0.1482602
Tm		p1B	p2B	p3B	pHR	pBB
AL	0.1390774	0.0462912	0.0041274	0.037286	0.0948981	
Matchup		p1B	p2B	p3B	pHR	pBB
Matchup	0.1189167	0.033095	0.0065364	0.0548732	0.1238706	

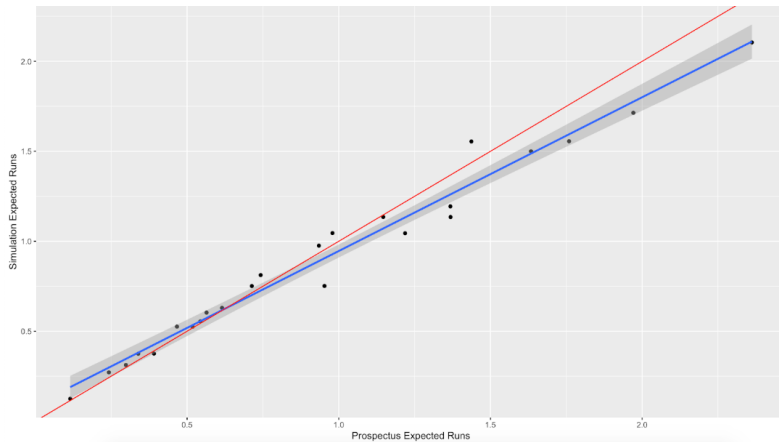
Methodology - Building and Testing an Inning Simulator

- ▶ Built a function that read in the number of desired inning simulations, initial bases and outs conditions
 - ▶ 24 different base-runner and outs combinations, each ran 10 million times
- ▶ For each play, sample one occurrence of out, single, double, triple, home run or walk
- ▶ Given base-runner positions and number of outs, move runners accordingly or end inning
- ▶ Track runs scored per inning until 3 outs have been recorded, then rerun the inning given same initial conditions.
- ▶ Compare average runs scored to real-life expected runs to test accuracy of model
 - ▶ Linear regression model, simulations vs 2019 data
 - ▶ If perfect match, slope = 1, intercept = 0
 - ▶ Results: slope = 0.850, intercept = 0.089

■

	0 - Pros	0 - Sim	1 - Pros	1 - Sim	2 - Pros	2 - Sim
000	0.5439	0.5546517	0.2983	0.3120666	0.1147	0.1244533
001	0.9345	0.9753566	0.5641	0.6035033	0.2422	0.2713257
010	1.1465	1.1350232	0.7134	0.7511442	0.3391	0.3747959
100	1.3685	1.1344901	0.9528	0.7514747	0.3907	0.3754299
011	1.4371	1.5543923	0.9792	1.0454250	0.4666	0.5256891
101	1.7591	1.5555543	1.2186	1.0444235	0.5182	0.5252382
110	1.9711	1.7137594	1.3679	1.1936163	0.6151	0.6295055
111	2.3618	2.1047169	1.6337	1.4995961	0.7426	0.8119866

$$R^2=0.974$$



Methodology - Building a Game Simulator

- ▶ Built function to take an offensive and defensive team, starting pitcher place in rotation and which league (AL/NL, important for use of DH)
- ▶ Calculated probabilities of single, double, etc. given pitcher vs each batter matchup
- ▶ Run innings 1-6 using starting pitcher, at 7th use the compiled reliever to close game
- ▶ Return runs at the end
- ▶ To simulate full game, switch offensive team to defense and defensive to offense, compare runs scored
 - ▶ Pitching matchup (place in rotation) holds the teams even so that we can compare the runs
- ▶ If score is tied, home team wins with 53.1% probability, away with 46.9% (Baseball Reference), rather than run extra innings

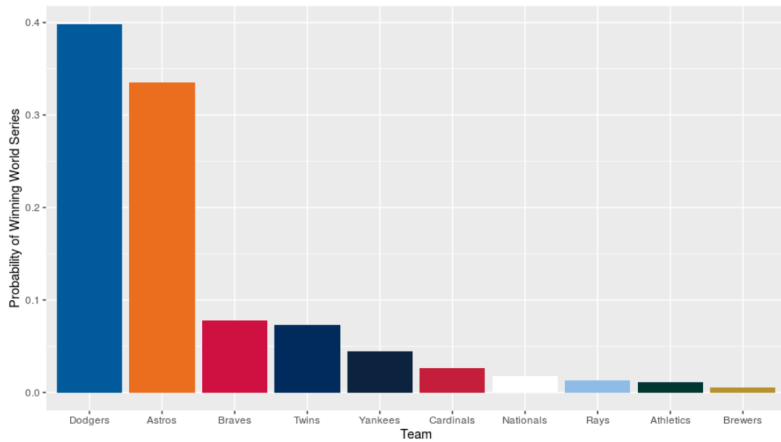
Methodology - Simulate Series

- ▶ Each game simulated 100,000 times
- ▶ Probability of each team winning the series: $\#$ of simulated series won/100,000
- ▶ Wild Card: One game series. Pitcher rotation spot #1 (ace)
- ▶ Divisional: Each series iteration ran until a team won 3 games. Pitching rotation started with ace, repeated once 4th pitcher pitched a game. Host: H,H,A,A,H for extra innings
- ▶ Championship: Each series iteration ran until a team won 4 games. Pitching rotation same as Divisional. Host: H,H,A,A,A,H,H
- ▶ World Series: Alternate NL and AL hosts forcing lineup changes to accomodate for DH. Hosting same as Championship. First game home team determined by regular season record

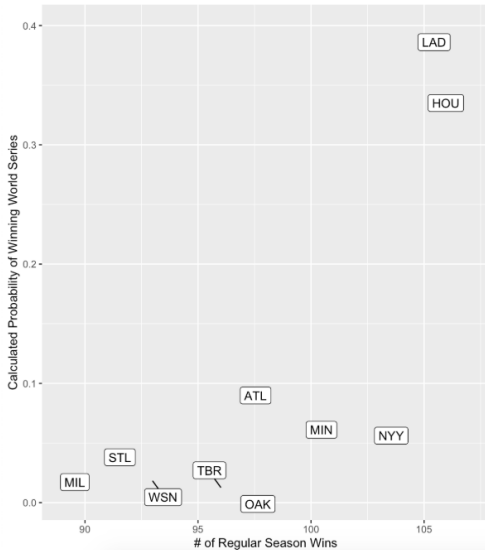
Results

Team	P(Advance to LDS)	P(Advance to LCS)	P(Advance to WS)	P(Win WS)
Houston Astros	1	0.786	0.583	0.335
Minnesota Twins	1	0.554	0.199	0.073
New York Yankees	1	0.446	0.137	0.044
Tampa Bay Rays	0.477	0.115	0.043	0.013
Oakland Athletics	0.523	0.099	0.039	0.011
Los Angeles Dodgers	1	0.831	0.628	0.398
Atlanta Braves	1	0.608	0.202	0.078
St. Louis Cardinals	1	0.392	0.098	0.026
Washington Nationals	0.321	0.08	0.045	0.018
Milwaukee Brewers	0.679	0.089	0.027	0.005

■



Strong correlation between regular season wins and probability to win World Series, unlike in the original study.



Summary of Results

- ▶ **Simulation:** Los Angeles Dodgers & Houston Astros heavily favored to compete in the World Series.
- ▶ **Real Life:** Washington Nationals defeated Astros 4-3 in the World Series (1.8% chance according to our simulations)
 - ▶ Nationals beat the Dodgers 3-2 in the NLDS (24.8% chance to defeat LAD)
 - ▶ Nationals won 8 straight games entering 2019 postseason
 - ▶ *Momentum a potential effect*
- ▶ **Our Simulation:** Favored teams with more regular season wins, 8 teams with $< 10\%$ chance of winning World Series.
- ▶ **Original Study:** Did not favor successful regular season teams as much, only 3 teams with $< 10\%$ chance of winning World Series.

Limitations & Further Study

- ▶ Our model produced results that followed a drastically different pattern from the original study - however, it was used to simulate events from a different year with different teams.
 - ▶ Only able to run 100,000 simulations of each series, compared to 2,000,000 in the original study
 - ▶ Difference between the team with the most regular season wins and the team with the least regular season wins in 2019 was 18 (107-89); only 7 in 2010 (97-90)
 - ▶ Evidence that there was more statistical inequality between the playoff competitors in 2019

Works Cited

- ▶ “Baseball Prospectus: Statistics: Custom Statistics Reports: Run Expectations.” Baseball Prospectus | Statistics | Custom Statistics Reports: Run Expectations, MLB Advanced Media, 2019, legacy.baseballprospectus.com/sortable/index.php?cid=1918829.
- ▶ “Did the Best Team Win? Analysis of the 2010 Major League Baseball Postseason Using Monte Carlo Simulation,” Journal of Quantitative Analysis in Sports, De Gruyter, vol. 8(1), pages 1-13, March.
- ▶ “MLB Stats, Scores, History, & Records.” Baseball, Sports Reference LLC, 2019, www.baseball-reference.com/.