Analyzing MLB play-by-play data to determine players' out distribution

Selah Dean
CSDS 234: Structured and Unstructured Data
December 16, 2022

Abstract
This project looks at baseball play-by-play data and provides a case study for Juan Soto's out distribution. While watching Juan Soto play, I felt that whenever he came up to bat, he grounded out to second base. Since baseball has many different statistics available, I looked online to find the answer to my question, but I was unable to find what I was looking for. In this project, I used data from Retrosheet which has play-by-play data for every game since 1915. I used this text data and a command line program to convert it into a structured table to be uploaded into a MySQL database. Once I created the database which contains tables for games, events, players, and teams, I queried the data to find every play by Juan Soto. Using these results, I ran a program I created in Python that produced a bar graph that displayed the frequency of all his outs. I found that after strikeouts, his most common out produced was a groundout to second. I compared these results with a few other conditions including all players since 1950 and all right and left-handed batters separately and found different distributions. The next step for this project would be to create a user interface, so individuals can write their own queries to find different out distributions with different conditions.

Introduction
Baseball is a game of numbers. Since Major League Baseball (MLB) was founded in 1876 there has been many different statistics that have been measured. In recent years, as technology has increased the number of aspects of the game that can be measured has increased as well. This has resulted in more statistics being measured. Data analytics has begun to take a larger role in the game which is changing how the game is being played.

Problem Statement
This project stems from a problem I ran into when I had a question watching a baseball game and could not find the answer in a simple google search. Juan Soto was a player for the Washington Nationals, and I spent the summer watching almost every game that the Nationals played. I felt that almost every time that Juan Soto came up to bat, he grounded to second base. I tried to find a way to look up a player's out distribution but was unable to find anything. Juan Soto ended being traded to the San Diego Padres in the middle of the summer, so I stopped watching him play every day. However, when the Padres were in the playoffs, I was watching the game and Soto grounded out to second base. This is when I decided to look into how to find a player's out distribution and whether he was truly grounding out to second base that often.

Related Work

There are numerous websites such as Baseball Reference, Baseball Savant, and Fan Graphs that have data for fans to view. However, when it comes to answering specific questions, such as the one I have, there is no straight forward way to find the answer. On FanGraphs, for each player there is a spray chart hat display all batted balls and what the result was (such as a certain hit, groundout, or flyout). The spray chart provides a visual for all of Soto's batted balls, but it does not specifically answer my question on Soto's out distribution. It also does not allow for further analysis or comparison with a group of players since the charts are for one player. This project looks specifically at outs produced and the distribution for different conditions. Based on the work done in this project, in the future the goal is to have a search engine where specific questions can be answered based on play-by-play data.

Methods

Finding the Data

Retrosheet is an organization that has digitized play-by-play data for every MLB game since 1915. I used their data to create a database that contained a structured format of every play that has happened since 1950. For each season, Retrosheet has a separate file for each team that contains all of the home games played that season. Each file is an ASCII text file of a series of records that details the events of each game.

The first step I took to create the database was I download the zip files for every decade since 1950. After unzipping the folders, I had a few different types of files. I had a file for every year and team in either an EVA or EVN format which is Retrosheet's special format for an ASCII text file. These files contained all the information about each team's home games for the season including all of the play-by-play data. There was also a file for every year and team in a ROS format which contained all the roster information for the team that season most importantly containing the player id and their first and last name. Finally, there was a file for each year that contained the team information which had the abbreviation for each team as well as their location and name.

Converting the Unstructured Data into a Structured Format

In order to make the data into a format that is combatable with a schema, I used a program from the Chadwick software project which converts the text from the play-by-play data into a csv file. Retrosheet provides a few different command line programs to produce a box score, a structured format of the play-by-play data, or a summary of the general game information. The Chadwick program is also a command line program which produces a structured format of the play-by-play data, but it does it for an entire event file and not just for a single game. I made a window batch file which called the program for every event file to convert them into structured csv files.

Creating a Database

I used MySQL to create a database to hold of the play-by-play data. I called this database retrosheet and defined a few different schemas. First, I made one for events, one for games, and one for subs. The event schema held all of the play-by-play data. The game schema held of the

generic information for each game and the sub schema held the information for any player substitutions that were made in each game. Next, I uploaded all of the event csv files into these schemas. Then I created a separate year id attribute for the event and game schema. Each tuple had a game id which identifies the home team, data, and number of game (either 0 if a single game or 1 or 2 if it was a double header). I was able to extract the year from this game id to create the year id attribute for each tuple. The addition of a year id made it easier for plays to be searched for by year.

After all of the play-by-play data was uploaded which had approximately 11 million tuples, I created two more schemas. One for all of the player information and one for all of the team information. I then uploaded to text files to fill these schemas. For the player schema I made the player id a primary key to eliminate any duplicates since players play for different teams meaning their information was in more than one file. Once this data was uploaded, I had all of the data I needed in my database.

Data Analysis Steps

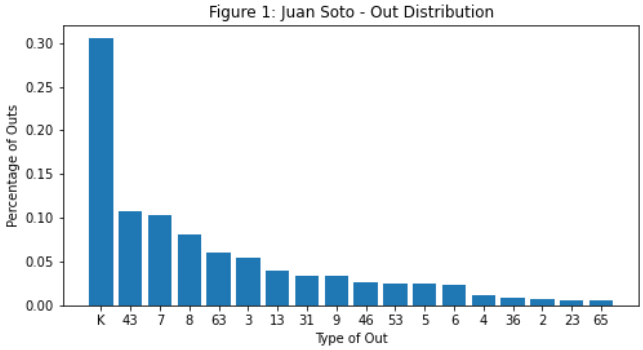
Once I completed my database, I started working on analyzing the data to answer my question about Juan Soto's out distribution. The first step I took was to perform a query that gave every at bat that Juan Soto has ever had in the MLB. The event table only has the batter id, so I performed a join with the player table to be able to search for Juan Soto. I exported the results from the query into a csv file.

I took the csv file from the query and used Python to clean the data and produce a graph. In Python, I used the NumPy, Pandas, and Matplotlib packages. First, I uploaded the csv file that was produced from the query in MySQL into a Pandas dataframe. The event text attribute is a string that provides an extremely detailed explanation of the play that occurred. I simplified this text by taking the first section before the forward slash. I also eliminated any text that followed a special character since this was also providing more detail, and I was just looking for the generic description of the play. Next, I created another Pandas dataframe that only contained tuples that resulted in an out produced by the Juan Soto. If the string in the event text contained a character that corresponded to something other than out, it was not included in the dataframe. Once I had a dataframe with all of the outs that Juan Soto has every made I calculated the percentage for each type of out. Using these percentages, I created a bar graph with the type of outs and the percentage for each. I did not graph any outs that occurred less than 0.5% of the time.

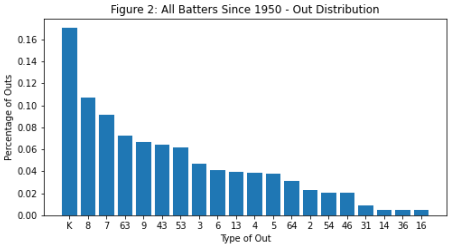
After doing this process for Juan Soto's data, I decided to investigate other conditions to determine if they provided similar results. I looked at every out made since 1950, every out made since 1950 for just right-handed batters and just left-handed batters, every out made since Juan Soto made since his MLB debut in 2018, and every out made since 2018 for just right-handed batter and just left-handed batters. To look at the other conditions, I made the steps I did for Juan Soto's data generalized. In MySQL, I performed the query and exported the result to a csv file. Then in Python, I had one function that requires the csv file from MySQL and a string for the name of the graph and it produces the bar graph. It calls two other functions I wrote, one to clean

the event text attribute and another to create the dataframe with only outs. Finally, it uses Matplotlib to create the bar graph.

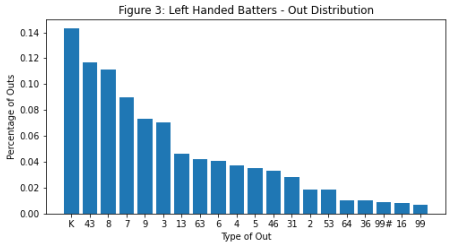
Results

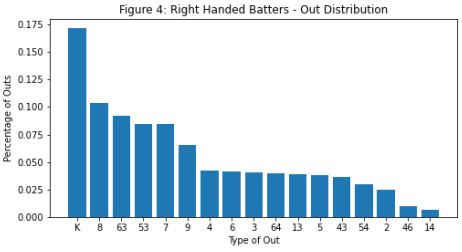


According to Figure 1, it can be seen that after strikeouts Juan Soto's most common out was a groundout to second base. This confirms the suspicions that I had about Soto constantly grounding out to second base. While it occurred about 11% of the time which is not nearly as much as the 31% of the time strikeouts it is still a significant amount more than most of his other outs produced. Close to a groundout to second base was flying out to left field which accounted for 10% of his outs. Other than flyout to center field all of the other outs that Soto produced accounted for less than 5% of his outs.

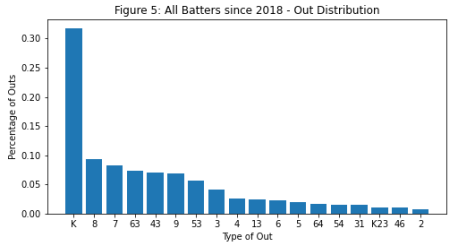


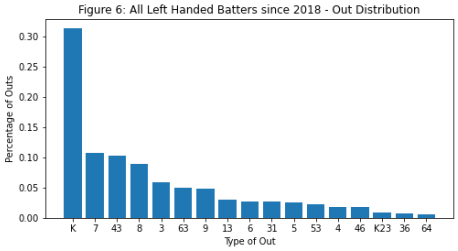
Next, I compared Soto's result to every out that was made by every player since 1950. Figure 2 shows that the distribution was different than Soto's distribution. In both cases, strikeouts were the most common, but after that the distributions greatly differed. Since Soto is a left-handed batter, I decided to split the outs made into two groups. One with right-handed batters and one with left-handed batters to see if there was a difference. As seen in Figure 3, left-handed batters out distribution resembled a similar distribution to Juan Soto, but there were a few differences. Strikeouts were the most common for both and groundouts to second base were also the next frequent. After that, the outs varied slightly. For only left-handed batters, the next most frequent out was a flyout to center field whereas Juan Soto's was a flyout to left field. Another significant difference was Soto flyout to right field about 4% where all left-handed batter flyout to right field about 7% of the time. Looking at Figure 4, right-handed batters more closely resembled the distribution for all players. This is most likely due to the fact that there have been more right-handed batters which means the skew the total distribution more towards their distribution. An interesting observation is that right-handed batters only grounded out to second base about 3.5% of the time which is significantly lower than the 11.5% for left-handed batters. This shows that while I noticed Juan Soto grounded out to second base frequently these is likely to due to him being a left-handed batter and where they tend to hit the ball on the field.

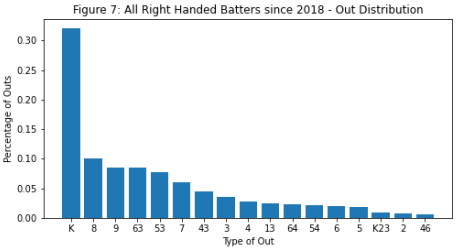




Another comparison I made with Soto's distribution was with every out made by every player since 2018 which is when Juan Soto made his major league debut. Figure 5 shows the distribution for outs since 2018 and is different from Soto's distribution after strikeouts being the most common. I split the outs into right and left-handed batters as well. Figure 6 shows that the distribution for left-handed batters which was similar to Soto's distribution. The main difference was Soto grounded out to second slightly more frequently than flying out to left field, but this was the opposite for all left-handed batters. The distribution of left-handed batters since 2018 closer resembles Soto's distribution as compared to the distribution of left-handed batters since 1950. According to Figure 7, the right-handed batter distribution was different from Soto's distribution and followed the similar observations of all right-handed batters since 1950.







An observation I made was the percent of strikeouts that accounted for outs nearly doubles when looking at all outs since 1950 versus all outs since 2018 going from 17% to 32%. Also, for all outs since 1950 the right-handed batters struck out 17% of the time whereas left-handed batters struck out 14% of the time. This gap decreased when looking at from 2018 where both right and left-handed batters strikeout for approximately 32% of their outs. This shows the changes in how the game is being played especially in recent years.

Conclusion

Juan Soto was grounded out to second base in 11% of his outs. While this may not seem like it occurred that often in the game of baseball where a 30% success rate in hits (equivalent to a .300 batting average) is elite, the other 70% of the time the batter is getting out. This means that 10% of those outs are a significant portion since there are many different ways to get out.

While the data confirmed my suspicions, it also showed that Juan Soto was not unique in this occurrence. The out distribution for all left-handed batters was similar to Soto's and the distribution for right-handed batters was not. This shows the difference between right and left-handed batters and where they are more likely to hit the ball. Left-handed batters are more likely to pull the ball meaning hit it towards the right-side of the field which results in more ground balls to first or second base. However, right-handed batters tend to hit more ground balls to the left-side of the field. This explains the difference in out distribution since right and left-handed batters are hitting more balls to different parts of the fields resulting in different outs being made more or less often.

Future Work

To further the research done in this project, creating a user interface to perform searches and produce graphs in one step would be beneficial. In this project, I used MySQL to perform the searches in the database and then exported these results to Python where I performed few different functions that I wrote to produce bar graphs to represent the out distribution. A user interface would allow individuals to find answers to their own questions about different players or different years or any other conditions. The interface provides a solution to the problem I had where there is not easily accessible data on out distribution even though there is data out there to be used to find the answer.

While I used this data for purely an observational purpose, it can be helpful for MLB teams. If a player is likely to hit the ball to a certain place, a fielder can be more prepared to position themselves to make the play. This has been done by teams with data collected from Statcast which is tracking technology that collects and analyzes large amounts of baseball data. From this data teams have been able to determine where a player is likely to hit the ball, so they shift their defenders into a better position to make an out. However, the shift is being banned for the upcoming 2023 season so other strategies will have to be taken. This data can provide valuable insight to help improve the game of baseball.

References

"Juan Soto - Spray Charts - Batted Ball." *FanGraphs Baseball*, FanGraphs, https://www.fangraphs.com/players/juan-soto/20123/spray-charts?position=OF&type=battedball.

Walsh, John. "The Advantage of Batting Left-Handed." *The Hardball Times*, FanGraphs, 15 Nov. 2007, https://tht.fangraphs.com/the-advantage-of-batting-left-handed/.