The Math of Baseball

Will Cranford

Introduction:

The aim of this presentation was to introduce the sport of baseball, explain how the game is measured, and to introduce more advanced analytics. In addition, I explored the limitations of statistics to value players and how statistics are affected by outside factors such as teammate performance and stadium conditions.

Baseball:

Baseball is considered “America’s Pastime”, and the most popular professional baseball league, Major League Baseball, has been around since 1869. The game is played on a diamond shaped outdoor field, where two teams compete in an untimed game. Each team starts nine players. The team that begins on defense sends out all nine players to the field, while the team that begins on offense sends out one batter at a time. The batter’s goal is to help his team score runs. The pitcher, a player on defense, throws the ball to the catcher, another defender. The batter attempts to advance around the four bases by either hitting the pitched ball into the field or by drawing a walk (when the pitcher throws four inaccurate pitches before throwing three accurate ones). The batter either makes an out (the defense catches the batted ball or throws the ball to first base before the batter gets there) or reaches a base. The batting team then sends up the next hitter in the order. The batting team keeps batting until they make three outs. At that point, the batting team and the defending teams switch, so the team that was batting is now out on the field, and vice versa. Each team gets to bat nine times, and the team with the most number of runs (a team scores a run when a batter makes it all the way around the four bases) wins the game. Each

Major League Baseball team plays 162 games. In order to reach the playoffs, a team usually must win at least 90 games.

The Development of Analytics in Baseball:

Baseball began in the 19th century as a game played by generally working-class men who used the sport to earn some money to supplement their incomes. However, as the game grew in popularity, the amount of money in the game grew, and the need for record keeping developed. Individual player statistics have been reliably recorded since 1876. Baseball is a sport that produces many statistics. Every time a batter comes up to bat, a statistic is recorded, whether that be a single, home run, strikeout, or a number of other possible outcomes. There are roughly 80 at bats every game, and the statistics are compiled at the end of every game for each batter, pitcher and fielder. These individual game statistics are added to a player's season statistics, and at the end of every season, a player's season statistics go into the record books. These season statistics allow teams to evaluate players and for the media to vote on awards.

The nature of baseball itself lends itself to statistics. Along with the large amount of data produced, the fact that individual contributions can be isolated is quite helpful for statistical analysis. There is a phrase that baseball is "an individual sport disguised as a team sport". Each time a batter comes to the plate and gets a hit, analysts can be confident that the batter himself was responsible for the result, rather than a teammate. Not every sport works this way. Soccer is an excellent example of a sport in which its difficult to isolate a player's contribution. For instance, let's say a player scores a goal. How much credit should go the player that scored, and how much credit should go to the player that passed the ball to the player that scored? What about the teammates that played good defense and allowed the team to possess the ball? Baseball statistics are much more straightforward with regards to determining an individual's

contribution. However, there is some teammate interdependence. When a pitcher gives up a hit to the batter, analysts have to consider how much of the blame should be assigned to the pitcher, and how much should be assigned to the defense, which failed to catch the ball.

Even though baseball is a sport that lends itself to statistical analysis, this analysis did not become common among major league baseball teams until the 1980's. One reason for this late development is that until the rise of computers, massive amounts of data were difficult to analyze by hand. Spreadsheet software made such analysis a lot easier to perform, and simulation software also allowed for a more in-depth look at data. Another reason is that many in baseball fought the use of statistics, as they (often athletes and managers that were former players) deemed that math had no place in baseball. However, as baseball teams began to run more and more like businesses, the teams that were anti-analytics were left behind. Information is power, and the teams that had more information on which players were valuable often won the most games. The field known as "sabermetrics", the study of baseball statistics, took off in the late 1990's, and was made popular in the 2011 movie Moneyball. This big budget movie detailed the rise of the Oakland Athletics, a team that did not have much money to spend on players, but through the use of statistics, they tried to find undervalued players, players that could help the team win even though they cost little money. This team made the playoffs quite often in the 2000's and cemented the importance of analytics in baseball. Now, all 30 teams have large analytics departments, often hiring elite college graduates that have studied math, physics, economics or computer science.

Player Evaluation:

The currency of baseball is wins. The teams that win more games have a greater chance of making the playoffs. The teams that have a greater chance of making the playoffs sell more tickets, attract more television viewers, and make more money. So the goal of baseball analytics as it pertains to player evaluation is to determine which players will cause your team to win the most games. There are two types of players: Position players, who hit and play on defense, and pitchers, who just pitch. In this paper, I will mainly focus on the evaluation of position players.

I mentioned that the currency of baseball is wins. Indirectly, the currency of baseball is runs, as to win the game, you have to score more runs than your opponent. When a position player comes up to hit, his goal is to produce runs for his team. To measure just how many runs a player produces, the run expectancy matrix can be used.

Base Runners				2010-2015		
1B	2B	3B		0 outs	1 outs	2 outs
__	__	__		0.481	0.254	0.098
1B	__	__		0.859	0.509	0.224
__	2B	__		1.100	0.664	0.319
1B	2B	__		1.437	0.884	0.429
__	__	3B		1.350	0.950	0.353
1B	__	3B		1.784	1.130	0.478
__	2B	3B		1.964	1.376	0.580
1B	2B	3B		2.292	1.541	0.752

The run expectancy matrix has 24 entries, one corresponding to each base-out state. A base-out state is the situation in which there are a certain number of men on the bases and a certain number of outs. The number in the matrix is the number of runs a team is expected to score over the rest of the inning. These expected values were calculated using the average runs scored in these situations using the dataset from the years 2010-2015. An example of an interpretation of the matrix: The lower right entry tells us that with runners on first, second and third base and two outs, the average team scores .752 runs. To calculate a batter's impact on the team's success, you subtract the run value in the preceding state from the run value in the state that followed the hitters at bat, and then add the number of runs that scored during the at-bat. To find the total number of runs produced over the course of the season, you sum up this change in run expectation from each at bat. The best hitters add about 70 runs per season with their bat. Baserunning ability can be measured in much the same way.

However, this calculation of value has some drawbacks. Using the average is fine for an estimate, but it contains many assumptions, which need to be assessed in order to have a more accurate estimate. Players that hit for good teams see more opportunities with men on base, where they have a greater chance to increase the team's expected runs scored. So great offensive teams often have their player's statistics overinflated. Another bias is that certain teams play in more offensive friendly environments. For instance, the Colorado Rockies play in Coors field, a stadium at great elevation. When at higher altitudes, the air is thinner and the ball travels further, which results in higher-scoring games. So certain teams look better than other teams on offense by this measure. To eliminate teammate bias, it is assumed that players have no control over which situations they excel in and which situations they perform poorly in. In other words, there

is no such thing as a “clutch” ability, in which players elevate their performance in the most important situations. To eliminate environmental impacts, each stadium is assigned a park factor, a constant which scales production to league average rates. This causes players that play in smaller parks or high altitudes to not have an advantage in these statistics.

Along with batting and baserunning, a player’s fielding ability is measured. Fielding ability analysis is less of an exact science. There are fewer statistics, and the ones that do exist measure a player’s ability to not make mistakes rather than their ability to catch more balls. This dearth of statistics is changing. Recently, Major League Baseball installed motion tracking cameras in all 30 stadiums, which allows teams to the position and velocity of all players on the field and the ball. This allows for teams to calculate how many runs a defensive player saves. These defensive runs are then added to batting runs and baserunning runs to calculate a player’s total value.

Reflection and Self Evaluation:

I was pleased with how this presentation went. At the beginning of the class, I asked how many students were familiar with the rules of baseball, and only two students raised their hands, a number lower than what I was expecting. Having taken this into account, I made sure that the rules of the game were clear at every point. I know how easy it is to get lost in the math of something you know nothing about. After my first presentation, in which my explanation of Dijkstra’s algorithm did not come out as clear as I would have liked, I made sure that my explanation of the run expectancy matrix was clear, and I felt like the class understood the details. The class’s comments were largely positive, though some questioned my choice of video for the explanation of baseball.

Sources:

www.fangraphs.com/tht/how-to-measure-a-players-value-part-2/

http://tangotiger.net/wiki_archive/Linear_Weights_System.html

https://www.youtube.com/watch?v=skOsApsF0jQ

https://www.fangraphs.com/statss.aspx?playerid=945&position=OF

http://tangotiger.net/re24.html

https://www.fangraphs.com/library/principles/linear-weights/