Power and Defensive Position in Professional Baseball

Abstract

In this paper, I analyzed data from the 2018 MLB season to see to what extent a player's primary defensive position is related to that player's power as a hitter. The metric used to measure power in this analysis was expected slugging percentage (xSLG). The purpose of this study was to see if differences in player types, such as size and athletic abilities which generally determine a player's position, are at all related to differences in power by players at different positions in Major League Baseball. The idea for this paper stemmed from my passion for baseball and my strong desire to pursue a career in the field of baseball analytics.

Research Question:

Do certain positions in baseball have more power than others, on average?

In modern Major League Baseball, different types of players play different positions on the field because each position requires a unique type of athleticism. For example, center fielders tend to be the fastest player on the field because they command the outfield and have to cover so much ground. Additionally, first basemen tend to be tall but not necessarily very fast because they have to be able to stretch to catch any ball that is thrown to them, but they do not have to move around very much on the field compared to other players. The purpose of this study is to see if those trends on defense translate to the offensive side of the game in any way, even though a player's offensive performance has essentially no effect on his defensive skills, and vise versa. Since a player's overall skill as a hitter is nearly impossible to quantify (as of now), I specifically looked at power in this study. In baseball, power is a specific type of offensive production that typically leads to a lot of extra-base hits, it is a very valuable tool to have as a hitter. This study will allow me to see which positions have better power-hitters than others and which positions are similar in the power category, in general.

Data Collection:

The data for this study was collected from baseballsavant.mlb.com, a website that tracks not only basic player statistics, but also advanced statcast metrics. The subjects of this study were Major League Baseball players who had a minimum of 250 plate appearances in the 2018 season (the most recent complete season). I used the inclusion criteria of at least 250 plate appearances because it removes players from the dataset that do not play at least somewhat regularly. The players that do not play regularly could potentially distort the data because they tend to only enter the game in later innings as a pinch hitter and do not play defense nearly as much as the starters who usually play offense and defense for the entire game. The inclusion criteria narrowed the dataset down to 313 players.

Variables:

The explanatory variable in this study was primary position according to baseball savant (including DH), a categorical variable. The response variable in this study was expected slugging percentage (xSLG), a quantitative variable. Ordinary slugging percentage is essentially a measure of the number of bases that a player records per at bat, on average. It is calculated using the following formula: SLG= [singles+(doublesx2)+(triplesx3)+(homerunsx4)]/AB, or SLG=(total bases/AB). It is a better measure of power than batting average because it gives more weight to doubles, triples, and homeruns than it does to singles. Expected slugging percentage, however, is more indicative of a player's skill than regular slugging percentage because it focuses only on the quality of contact and takes defense out of consideration. In calculating xSLG, each batted ball is given a single, double, triple, or homerun probability based on its exit velocity and launch angle as compared to those of other batted balls in the Statcast database. These probabilities are then plugged into the same aforementioned slugging percentage formula to calculate xSLG.

Explore Data:

- *Summary Statistics (Calculated using R):*

Position:	**1B**	**2B**	**3B**	**C**	**CF**	**DH**	**LF**	**RF**	**SS**
Mean:	0.448	0.392	0.421	0.394	0.398	0.490	0.435	0.427	0.396
SD:	0.0533	0.0464	0.0593	0.0473	0.0749	0.0655	0.062	0.0668	0.0554

According to the summary statistics and boxplots (boxplots are shown in the appendix, Figure 1), the positions with the highest and lowest average expected slugging percentages are DH and 2B, respectively. The boxplots make it easy to see that there are at least some visual differences in xSLG for different positions. The variability shown by the differences in means or the centers of the boxplots represents between groups variability. This type of variability is essentially explained by the differences in position among the players in the dataset. This is the type of variability we are primarily interested in for the purposes of this study. An F-test will help determine whether these differences are statistically significant or simply due to random variation.

Additionally, the variability for each individual position is relatively equal. The variability shown by the standard deviations and the sizes of the boxplots represents the within-group variability. This type of variability is unexplained variability that is simply due to individual players in each position group not having the same xSLG. According to the summary statistics, the highest standard deviation is less than 0.03 greater than the smallest one. The boxplots also show that each position has a nearly equal spread. The fact that the within-group variability is nearly equal for each position will make it easier to analyze the data and form conclusions.

F-Test:

Before an F-Test is carried out with this data, we must ensure that all of the validity conditions are satisfied. (1) The samples are independent from each other: one players hitting generally has no effect on another player's hitting, so this condition is satisfied. (2) Sample standard deviations are roughly equal: the largest sample standard deviation is less than twice the smallest standard deviation, so this condition is satisfied. (3) Sample size is large: the overall sample size is 313 players and there are at least 20 in each group, so this condition is satisfied. All validity conditions for the theory based F-test are satisfied, so we can continue with the analysis.

Hypotheses for Assessing Statistical Significance Between Multiple Groups:

- Null Hypothesis: no association between position and xSLG, all group means are the same.
- Alternative Hypothesis: there is an association between position and xSLG, at least one group mean is different.

ANOVA F-Test Output from R: (full output shown in appendix, Figure 2)

F-statistic: 7.2205 **p-value**: 9.031e-09

The small p-value indicates that it is very unlikely that the differences in xSLG between position groups would occurr by chance alone. We can reject the null hypothesis and conclude that there is some association between position and xSLG.

Pairwise T- tests:

Since the F-test has confirmed that an association exists, we can now look at differences between each position and see which positions differ and by how much. Performing pairwise T-tests using R allows us to do this. The output from performing pairwise T-tests with the data using R is shown in the appendix (Figure 3).

The p-values in the output can be interpreted as the probability of getting the group means that we got from the data assuming that those two positions in fact have the same amount of power. If the p-value is large, it means that it is very plausible that the two positions actually have the same power and the difference is just due to chance. Likewise if the p-value is small, it means that it is very unlikely that the difference occurred just by chance. The pairwise T-tests allow us to look at differences between each and every position to see if the differences are significant.

Inferences/Interpretations from Pairwise T-tests:

The first interesting thing that I noticed is that positions that "go together" tend to not be significantly different in terms of power. What I mean by that is that there are certain sets of positions on the baseball field that players tend to play if they play multiple positions. This is because these positions require similar skill sets for players to excel at them. Common examples are third base and shortstop, left field and right field, and second base and shortstop. According to the output, all three of these pairs of positions have high p-values and are not significantly different from each other. This is interesting because it shows that positions that require similar defensive skill sets also have similar offensive production, on average.

Another interesting value was the one between center field and shortstop. These two typically are home to the most athletic players on the field, and are arguably the most important defensive positions. It is very rare to find players that play both centerfield and shortstop, but few would be able to argue that both these positions do not require exceptional speed and agility. The p-value for these positions is high, meaning that they have similar power, on average.

The designated hitter position also gave some interesting values. Recall from the aforementioned summary data that the designated hitter position had the highest mean xSLG. According to the pairwise T-tests, the designated hitter position is significantly different from every other position. This allows us to conclude that the designated hitter position is significantly more powerful than every other position, on average. This is interesting because designated hitters tend to be strong players who are not very fast or defensively skilled, they are typically the best power-hitter in the lineup. This trend has been confirmed with the pairwise T-tests.

It is important to mention that carrying out multiple t-tests increases the risk of type-I error. Type I error is when we reject the null hypothesis, when in reality the null hypothesis is true. So it is possible that some of the significant differences from the pairwise t-tests could be false positives.

Conclusions

The analysis led to the conclusion that there is an association between position and power in Major League Baseball, and that certain positions that are similar defensively also have similar levels of offensive power, on average. The study also tells us that there is sufficient evidence to conclude that the designated hitter position is the most powerful position, on average.

Limitations

It would be difficult to generalize the results of this study to all levels of baseball across all time. Baseball is a game that is constantly changing, so the trends analyzed in this study could be drastically different 20 years from now or 20 years ago. However, it is very reasonable to say that the results can be generalized to all of Major League Baseball within a few seasons of this study. For example, the results of this study could be used to predict trends for next season. It would be interesting to perform a similar analysis across multiple seasons to see how/if the trends change across different eras of baseball.

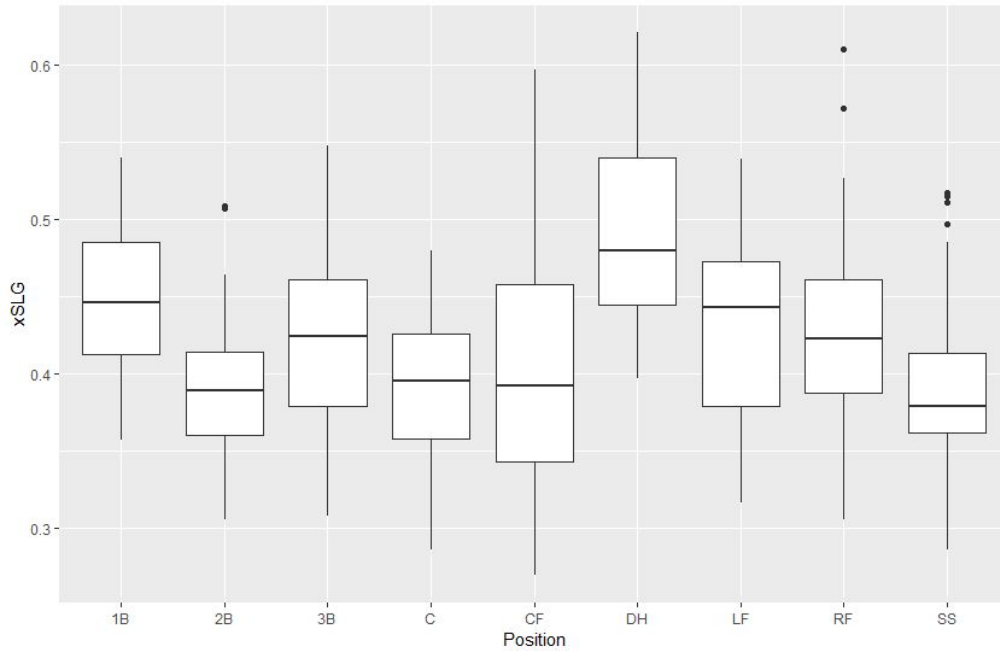
Another limitation of this study is that is assumes that xSLG is an accurate measure of how good of a power hitter a player is. There is not a metric that exists today that will perfectly measure a hitter's power, but xSLG is one of the best ones available. If I were to further investigate this research question, I would carry out the analysis using other statistics (such as homeruns or hard-hit rate), and see if the trends remain the same.

Lastly, the R-squared value for the model is only 0.1597. This means that about 16% of the variability in xSLG is explained by the primary position of a player. This is a very low R-squared value, but obviously it does not mean that there is a lack of statistical significance in the model, only that the model does not have very much practical significance. If I were to do further investigation on this study, I would include additional quantitative player attributes to predict a hitter's power. Including additional quantitative explanatory variables would likely increase the R-squared value and the practical significance of this study by a significant amount.

Appendix

(Figure 1)

Parallel Boxplots (Created using R):



(Figure 2)

(The p-values that are statistically significant at $\alpha = 0.05$ *have been highlighted in red.)*

```
> pairwise.t.test(est_slg, pos, p.adj="none")
```

	1B	2B	3B	C	CF	DH	LF	RF
2B	3.0e-05	-	-	-	-	-	-	-
3B	0.04592	0.03102	-	-	-	-	-	-
C	0.00012	0.86350	0.05631	-	-	-	-	-
CF	0.00040	0.66117	0.10549	0.79530	-	-	-	-
DH	0.02505	1.7e-07	0.00024	5.5e-07	1.7e-06	-	-	-
LF	0.34483	0.00161	0.30701	0.00401	0.00987	0.00368	-	-
RF	0.11393	0.00670	0.63229	0.01502	0.03352	0.00068	0.56078	-
SS	0.00012	0.78940	0.06289	0.93000	0.85867	5.7e-07	0.00439	0.01663

(Figure 3)

```
> model1 <- lm(data=expected_stats, est_slg~pos)

> anova(model1)

Analysis of Variance Table

Response: est_slg

           Df  Sum Sq   Mean Sq F value    Pr(>F)

pos         8 0.20076 0.0250951  7.2205 9.031e-09

Residuals 304 1.05657 0.0034755
```

References

MLB Statcast Glossary. (n.d.). Retrieved May 2, 2019, from http://m.mlb.com/glossary/statcast

Statcast Leaderboards. (n.d.). Retrieved May 2, 2019, from
https://baseballsavant.mlb.com/expected_statistics?type=batter&year=2018&position=&
eam=&min=250
2018 Statcast Expected Statistics Leaderboard