Week 5-7 Report - Hypothesis Testing

By: Jai Arora, Robert Banas, Bailey Harrison, Simran Mallik

Data Cleaning

Our team first utilized the MLB website to find the datasets that we were going to be using to gather information on the XBH, BABIP, SO%, BB/K stats. The first step consisted of utilizing the filters on the MLB website to separate our data into five separate datasets (data set a, data set b, data set c, data set d, data set e), which represented players within the Post-All-Star Break of 2018, Pre-All-Star Break of 2019, Post-All-Star Break of 2019, the first 30 games of 2020, and the final 30 games of 2020 respectively.

From there, we went on to copy and paste these filtered datasets into google sheets. We manually cleaned the datasets to only include the names of the players, their team, XBH, BABIP, SO%, and BB/K. Once we had condensed the datasets within google sheets, we transferred the datasets into R so that we could use code within R to randomize our datasets to sample out 15 players within each dataset.

Some of the players that were picked out randomly originally using R were not in both the post-All-Star break and pre-All-Star break datasets, which is why we had to manually go and pick players randomly in replacement of these players that were randomly chosen but not present in both datasets.

Once we had our random players chosen for all the datasets, we began to import the data into google sheets. We split up the 15 players as a team and manually went on to find the statistics of each player on the MLB website to import them into google sheets. For the most part, it was just a matter of finding the values for the XBH, BABIP, SO%, BB/K statistics for each player and plugging them into google sheets; however, for the 2020 season, in particular, we had to manually calculate these statistics.

The reason that we had to manually calculate the values for the XBH, BABIP, SO%, BB/K of the players within the 2020 season is because information on the first 30 games of the 2020 season and the last 30 games of the 2020 season was not available on the MLB website. Therefore, we had to improvise, and we decided to divide up the 2020 season into two portions. The months of July and August served as the first half of the season, and the month of September served as the second half of the season.

The way we manually calculated the XBH is we summed up the doubles (2B), triples (3B), and home run (HR) statistics. For the BB/K statistic, we added up the total walks and divided it by the total strikeouts. For the SO% statistic, we added up the total strikeouts and divided it by the total at-bats. Lastly, for the BABIP statistic, we subtracted the number of home runs from the number of hits and then divided the result by
(at-bats - strikeouts - home runs + sacrifice flies).

Finally, we created additional datasets named Result 1 and Result 2 that combined the post-All-Star break of 2018 and pre all-star break of 2019 as well as the post-All-Star break of 2019 and the first 30 games of 2020 respectively. These datasets held the numerical differences between the values for the stats between the two seasons/datasets for each individual player as well as the average difference for each statistic. We eventually averaged the Result 1 and Result 2 statistic to use in our 2021 Pre All Star break prediction.

Statistically Significant Correlations

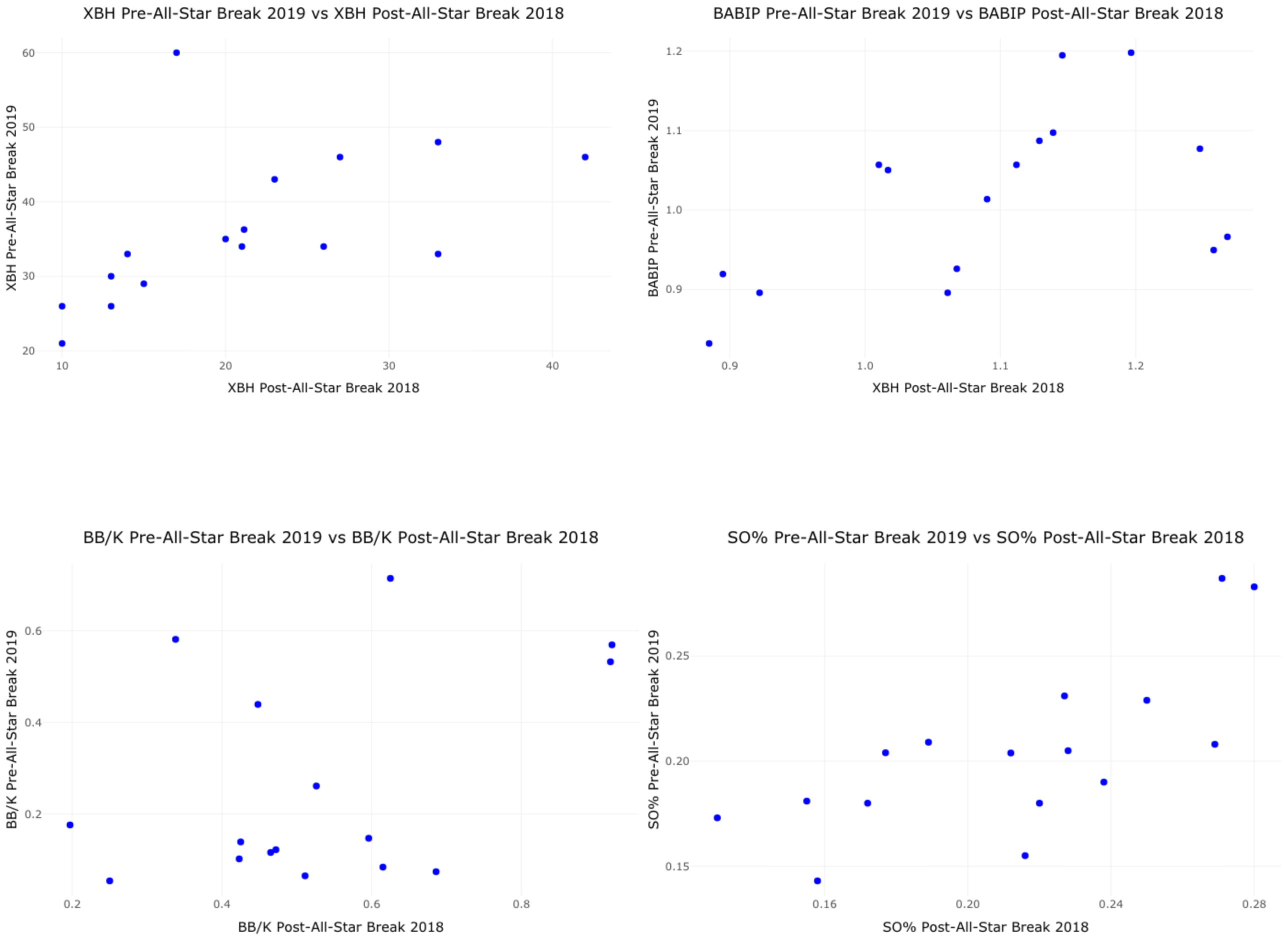
Our team identified if there was a statistically significant correlation between player performance post-All-Star break in one season and player performance pre-All-Star break in the next season. This was conducted by checking if there was a statistically significant correlation (a p-value of less than .05) of the average score of each metric for the different seasons using correlation tests in R.

Between the post-All-Star break of 2018 and the pre-All star break of 2019, there was a statistically significant correlation between the average XBH, BABIP/League Average score, and SO% for the two seasons. There was no statistically significant correlation between the average BB/K for the two seasons.

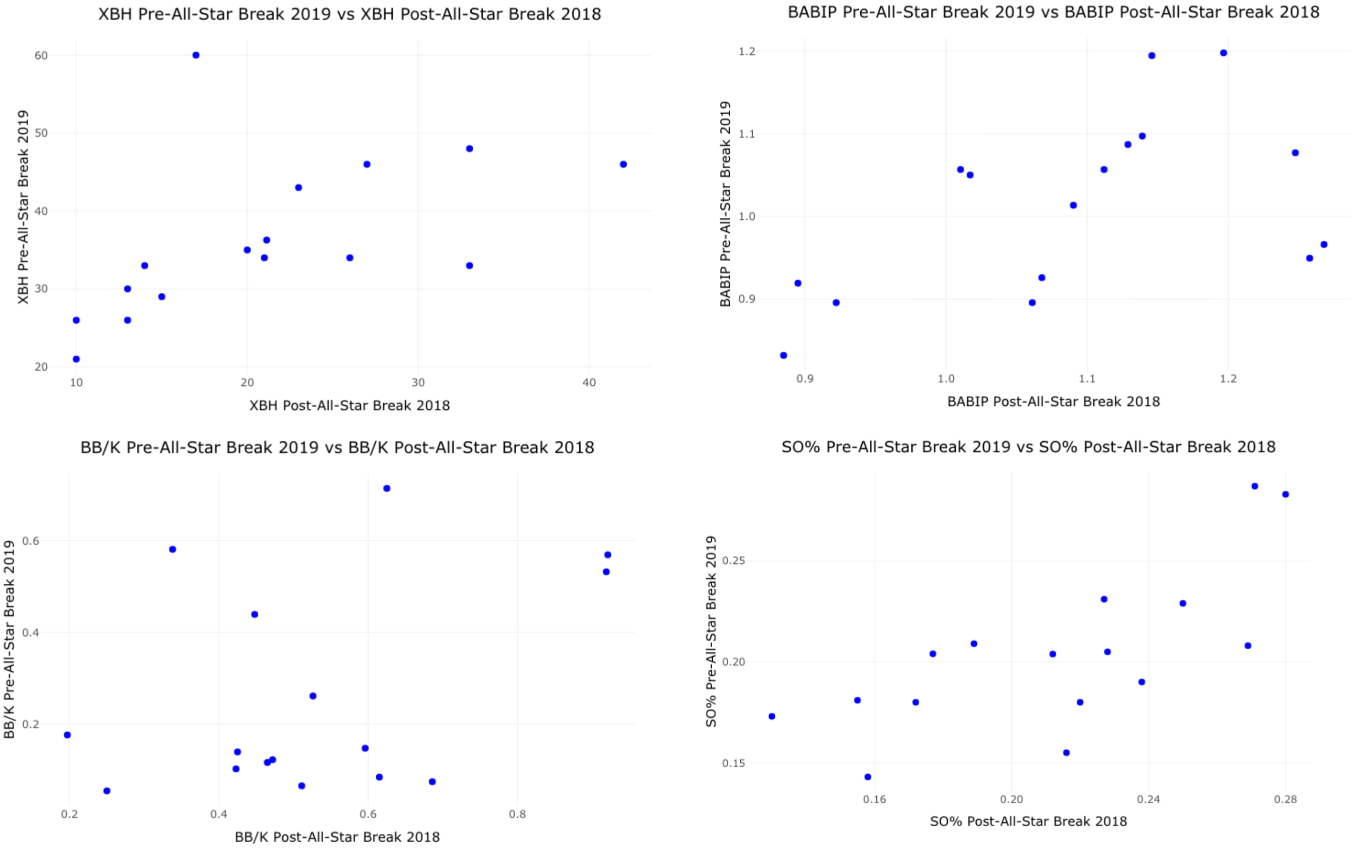
Interestingly, for the post-All-Star break of 2019 and the first 30 games of 2020, there was a statistically significant correlation between the average SO% and BB/K for the two seasons. There was no statistically significant correlation between the average XBH and BABIP/League Average for the two seasons.

A statistically significant correlation indicates a high possibility of using one variable to predict the other variable. If there is a statistically significant correlation between a metric from one year and the same metric for the following year, it’s likely that this metric is reliable in terms of predicting future player performance in that metric. (It’s important to distinguish that high correlation in one metric may not imply high prediction power for a player’s overall future performance since one single metric is not predictive of the holistic player ability/performance.) The only metric that had a statistically significant correlation over every season between 2018 and 2020 was SO%, indicating that this metric may be reliable in terms of predicting future player performance for that metric. However, because of the limited data which spans only 3 years and includes the first and last 30 games of 2020 rather than the usual number of games for the normal length of an entire season, it is possible that our correlation tests could be flawed. With more data, it may be more likely that different metrics experience a statistically significant season-to-season correlation, implying a greater player performance prediction power for these metrics.

Comparing Metrics in 2018 Post ASB Versus 2019 Season Pre ASB



Comparing Metrics in 2019 Post ASB Versus 2020 Season Pre ASB



XBH Pre-All-Star Break 2019 vs XBH Post-All-Star Break 2018

BABIP Pre-All-Star Break 2019 vs BABIP Post-All-Star Break 2018

BB/K Pre-All-Star Break 2019 vs BB/K Post-All-Star Break 2018

SO% Pre-All-Star Break 2019 vs SO% Post-All-Star Break 2018

Problems Encountered

Randomizing Players

One issue we ran into was finding players that qualified for datasets in different years. When randomizing, multiple times we found that players in our post-All-Star game dataset did not show up in the next season's pre-All-Star game data. This was because they did not qualify for the 3.1 plate appearances required per game played to achieve the status of playing a "full" season. We had to manually select certain players in the datasets to get all 15 players qualified.

Normalizing the BABIP Metric

As discussed in our week six meeting, we wanted to find a way to normalize the BABIP metric since its formula incorporates counting statistics which can lead to issues in smaller sample size correlation tests. Therefore, we normalized the BABIP of each player to the league average BABIP in that year (.295 in 2018 - .298 in 2019 - .292 in 2020). This gave us a ratio of what percent a player's BABIP is better or worse than the average of 1. For example, Josh Bell's ratio in dataset A was 1.01, meaning his BABIP was 1.01-1= .01 or 1% higher than the league average. Looking at dataset B, his BABIP ratio was 1.057, showing that his BABIP was 5.7% higher than the league average in 2019. Taking the difference between these numbers, we see his BABIP differential from pre-All-Star break 2019 and post-All-Star break 2018 was around 4.7%. Doing this same methodology, accounting for different league average BABIPs in each year, and taking the average for all players between dataset A and B, and then between C and D, and then evenly weighting these two numbers, the overall average differential was around -6.57% from post all star breaks to the next season's pre all star break. Scaling this to a historical league average BABIP of .300, our dataset estimates that a player's BABIP can expect to drop around 19.7 basis points from post-All-Star break to next year's pre-All-Star break. We used this value in our 2021 pre all star break predictions. It was relatively similar to our original, non-normalized BABIP estimation change of 20.9 basis points. Still, we see our normalization process decreased variation in BABIP results by around 1.33%.

0 values in BB/K Statistics

Another issue that we ran into was the fact that two of our predicted BB/K values were negative which is statistically impossible. Our analysis predicted that the average difference in BB/K values from post-All-Star break to pre-All-Star break in the next season drop by around .1016. However, two players with poor plate discipline that had horrible BB/K numbers in the last ~30 games of 2020 were Javier Baez and Raimel Tapia. Baez had a pitiful ratio of .036 and Tapia had a value of .095. Therefore, our predictions for these players in 2021 technically have a floor of 0. It is unrealistic to think that they will never walk in 2021, so instead, we suggest that they maintain around the same poor ratios that they displayed in 2020.

XBH Prediction

For the reasons discussed in our week six meeting, unlike the other statistics which took the average of result one and result two, only the result one statistic was used to predict a change in the XBH metric for the first half of the 2021 season. This is because of the number of games played and differential in at-bats due to the shortened 2020 COVID season.

Why Our Predicted Results May Be Different

1. **Vladimir Guerrero Jr.-** As a 22-year-old top prospect with an 80 graded hit tool, one can be really excited about his potential to breakout next season. Getting his feet wet in 2020 paired with allegedly losing 40 pounds of weight this offseason makes it very possible that Vlad exceeds our expectations from our model. Specifically, his plate discipline (and therefore his BB/K) could improve as he gets more accustomed to MLB pitching, and his power numbers (and XBH statistic) could outperform our model with his more lean muscle body.
2. **Javier Baez-** The former MVP runner-up is in a contract year; therefore, we think he could either produce well and use this as motivation, or let it get in his head if the Cubs cannot figure out an extension with him. Additionally, Baez complained how he struggled in 2020 because he was not allowed to view game film of himself in his previous at-bats. It will be interesting to see if the MLB changes this rule for 2021 or if Baez can adapt and improve his BB/K and SO% numbers if they do not.
3. **Jesus Aguilar-** Jesus Aguilar had a strong comeback season in 2020 after a not so good 2019 season. He almost beat his XBHs from 2019 in just 51 games in 2020. The only issue with projecting Aguilar's 2021 season is the amount of playing time he will receive. He and Garrett Cooper split time between first base and designated hitter in 2020. Now in 2021, there is no designated hitter for the National League so his at-bats may decrease. From our model we suggest, he will get around 24 XBHs, but we think it can be higher, it just depends if he is the everyday starter.
4. **A.J. Pollock-** Coming off a World Series win with the Los Angeles Dodgers A.J. Pollock should have another good season. Now that Joc Pederson is gone Pollock should get more at-bats. Last season Pollock had a career high in slugging percentage, but most of his XBHs were home runs. From our model we predict him to have 26 XBHs before the All Star break. Potentially more if he can hit more than base hits and home runs. Also, Pollock doesn’t walk a lot so our model suggests he will have a low BB/K.
5. **Didi Gregorius-** Playing all 60 games for the Philadelphia Phillies in 2020, Didi Gregorius had himself a solid shortened season, compared to his disappointing 2019 season. Ideally, if Didi Gregorius duplicates his numbers from 2020 over a full 162 game season, he has the potential to be a real offensive threat in the middle to the bottom of the lineup for the Phillies. Our model predicts about 26 XBHs for Didi and we believe he can do that.
6. **Mark Canha-** As a relatively late bloomer, it will be interesting to see how Canha performs in his age 32 season. Notably, with former Oakland Athletics Left Fielder Robbie Grossman signing with the Tigers, Oakland’s manager Bob Melvin said that Canha will play a bigger role with the team in 2021, indicating that his numbers, specifically XBH could improve with more at-bats.
7. **Kole Calhoun**- As we mentioned in a previous project, Kole Calhoun is one of the most underrated players in the game right now. That being said, he is coming off minor offseason knee surgery, and could miss the start of the season. With no spring training at-bats, he may take a while to get back up to speed during the 2021 season.

8. **Willy Adames-** With the Rays having top prospect Wander Franco, who is a SS like Adames, scratching at the door of being ready to play in the MLB, this is a big season for Willy. It will be interesting to see if he can prove his worth and play at an elite level to keep Franco in the minor leagues this season. However, being the insane talent that Franco is, it would not surprise me if Wander stays in the minors for about ~1.5 months till May so the Rays can save a year of his service time, and then call him up to take Adames' place. Therefore, Adames' numbers could drop: specifically his XBH statistic in our dataset.
9. **Cavan Biggio-** Heading into his age 25 season, it'll be interesting to see how Biggio performs in his first full season. One article predicts that Biggio could see a drop in his home runs (XBHs and BABIP metrics) since the MLB is changing the ball so fly balls travel less far. Biggio had the 8th lowest distance of home runs in 2020, so some believe some of his previous home runs will wind up falling short on the warning track in 2021.
10. **D.J. LeMahieu-** Riding into the season after signing a new 6 year, $90 million contract, LeMahieu may be feeling a little relaxed with that type of financial security. Now, as a 32-year-old, after consistently stellar seasons, will D.J. be the same type of player, or will he finally slip out of his prime? It will be interesting to see how he performs.
11. **Raimel Tapia-** Raimel Tapia received the majority of the playing time in 2020 for the Colorado Rockies after Ian Desmond opted out of the season because of COVID. Tapia had a fairly good 2020 season, but he strikeouts a lot compared to his walks. From our model we predict his BB/K to be 0, but that is most likely wrong because he should get at least one walk. With little power, Tapia should be a good source of runs, batting average, and steals.
12. **Cesar Hernandez-** Having a better than career season, Cesar Hernandez is looking to produce that again for the Cleveland Indians. Not having the greatest power, Hernandez should remain a solid contact hitter. From our model, it suggests his BABIP will decrease, but still be above the league's average. Also, it predicts his BB/K will decrease a lot.
13. **Joey Votto-** Coming into the 2021 season as a 37-year-old, this year may be the start of Votto's twilight years in Cincinnati. Still, in his Hall of Fame career, Votto has had arguably one of the best batting eye's of all time. Therefore, we think our model prediction that his BB/K and SO% will worsen is probably false. However, his power may start declining with old age, so his XBH stat could underperform.
14. **Austin Riley-** With 2021 being Austin Riley's only third season in the league, the young third baseman still has a lot to learn for the game. His stats being slightly low can cause some doubt, but he has improved in his first two seasons. Last year Riley's strikeout rate dropped from 36.4% to 23.8% while he improved his walk rate from 5.4% to 7.8%. Our model suggests his strikeout rate will slightly increase by 4%, but that is still lower than his rookie season.
15. **J.D. Davis-** Breaking out during the 2019 season, J.D. Davis had a disappointing 2020 shortened season. Playing 56 out of 60 games his numbers were below his career averages, but his OBP was above average. This can be due to his high amount of walks he had last season. In 2019 Davis played 140 games and had only 38 walks, but in 2020

he played 56 games and had 31 walks. If he continues increasing the number of walks he draws for the 2021 season his BB/K can still be high and increase.