W200 Project 2

Team Members:

Sean Norris, Shivani Sharma, Alexander To

Git Repository:

Name: Project2_Norris_Sharma_To

Link: https://github.prod.oc.2u.com/UCB-INFO-PYTHON/Project2_Norris_Sharma_To

Primary Dataset:

The data used in this assignment are taken, primarily from the website: https://baseballsavant.mlb.com/. This dataset contains the latest batted ball and pitch f/x data from major league baseball (MLB). These data have been supplemented with data from https://baseballreference.com and https://fangraphs.com. We have pulled the regular season data from 2015 to 2020. These data are contained in corresponding folders within the "Data" folder of our repository. More details of the MLB seasons analysed in this project are provided in the Appendix.

Introduction and Research Questions

In 2015, Major League Baseball (MLB) installed state-of-the-art tracking technology in all 30 ballparks. This technology enables tracking of the flight of the ball and the movement of players throughout the game. This tracking allows analysts, broadcasters and others unprecedented insight into what is actually happening on the field. For instance, we now know the location, speed, and type (e.g., fastball, curveball, etc.) of every pitch in every game. Similarly, for hitters we know the angle, direction and speed of every ball hit. The implementation of this technology marks the dawn of a new era in baseball known as the "statcast" era.

For this report we will examine a few macro trends from these data in answering open questions and reexamining conventional wisdom currently present in the MLB.

Research Questions:

- What are the most effective pitches (e.g., hardest to hit well) in baseball during the statcast era?
- What can we learn about how elevation influences pitching and hitting in the statcast era?
- What can the data we've collected during the statcast era, tell us about whether there are players who can consistently demonstrate an ability to hit the ball harder in the clutch than in non-clutch situations?

Within each section we will include an explanation of the question, how we intend to operationalize our analyses, data visualizations and an explanation of our findings.

Question 1 - Effective Pitches

Over the last five years, evidence of a change has emerged in the approach of pitchers to the most central conflict in baseball, their battle with hitters.

In August, Tom Verducci of Sports Illustrated reported fastball usage by pitchers in 2020 dipped below 50% for the first time this decade. Verducci ascribes this to the success hitters are having (as measured through traditional statistics) against the fastball. We examine this claim further in the Appendix.

Anecdotal evidence has also emerged. Trevor Bauer, the 2020 NL Cy Young Award winner, discussed this on his YouTube channel in June, as he was breaking down an at-bat he pitched against multiple-MVP winner Mike Trout, "This was back when sinkers [i.e., 2-Seam Fastballs] were okay in the league before they were, just, the worst pitch in baseball."

This prompted us to ask the question, "What's the most effective pitch (e.g., hardest to hit well) in baseball in 2020, and has that changed during the statcast era?"

The following are some example fields we will use to evaluate our question:

- **pitch_name** we will be using this field to distinguish one pitch from another (e.g., fastball versus curveball)
- **launch_speed_angle** a combined metric from statcast that estimates how "well" a ball was hit (e.g., combination of exit velocity and launch angle) and is described in more detail below
- **release_speed** an estimation of how fast a pitch is traveling in miles per hour (mph) as measured coming out of the pitcher's hand
- **game_type** indicates whether the game was played during pre-season, exhibition or post-season, for the purposes of this analysis we will only study regular season games
- **year** indicates the year in which the game was played, for this analysis we will examine all years for which we have data (e.g., 2015 through 2020)
- **type** indicates whether the ball was put in play, called a strike or called a ball by the umpire, for this analysis we will only consider strikes and balls put in play and exclude balls, assuming they were unhittable
- **pfx_z** vertical movement, measured in inches, of the pitch from 40 feet away as compared to a theoretical pitch thrown at the same speed with no movement

Data Cleaning

After reviewing how often each pitch was used and performing secondary research, we've re-coded some of these data to improve our analyses and make them more closely resemble reality.

- *D/M/Y to Year.* We changed the date time format on our dataset to only reflect the year of the games, as we are not concerned with specific months or days.
- *2-Seam Fastball → Sinkers*. There has been controversy over whether these two pitches are actually different from one another. Additionally, it appears that in 2020, MLBAM decided they are the same and have grouped them because no 2-Seam Fastballs were recorded in 2020. For ease of comparison between seasons we re-coded all "2-Seam Fastballs" as "Sinkers."

- *Fastball → 4-Seam Fastball.* Early in this period a few pitches with the designation “Fastball” were recorded, the 4-Seam Fastball is the most common fastball and these are most likely to be a misclassified pitches of that type
- *Dropping Intentional Balls and Pitch Outs.* We’ve scrubbed these from our dataset because these are not attempts by the pitcher to get the batter out. Further, intentional balls do not exist after 2017 because that was when MLB allowed teams to intentionally walk batters without having to actually through the pitches
- *Dropping Screwballs and Unknown.* We’ve scrubbed these pitches from our analysis due to small and inconsistent sample sizes

Sanity Check

Before we continue it will be important that we examine whether these data seem plausible. In general, we should find that pitches recorded as “4-Seam Fastballs” are thrown with more velocity and less movement than those recorded as “Curveballs.” In the figure below we see the vertical movement (i.e., Y-Axis) and speed (e.g., X-Axis) of Curveballs plotted against Fastballs. There are many more Fastballs (890,455) in this dataset than Curveballs (188,027). Below we see that Fastballs are generally faster and drop less. This suggests that while there may be some noise in these classifications (e.g., 90+ mph Curveballs) in general the classifications hold. This was confirmed by looking at the average speeds for each pitch type (discussed further below), and seeing that they followed the order understood by most baseball fans (e.g., 4-Seam Fastball is highest and Eephus is lowest). Additional noise may be explained by some pitches being poorly executed (e.g., unintentional movement and speed).

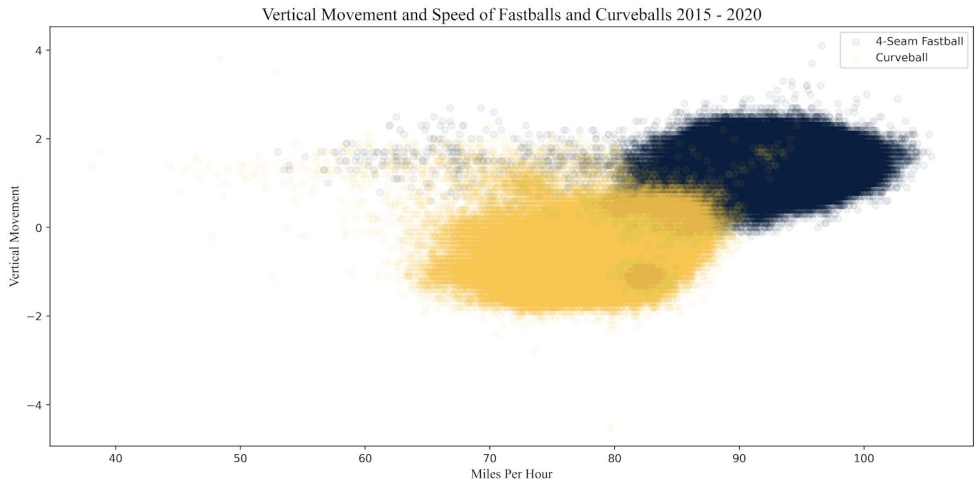


Fig 1. Vertical Movement vs Speed of Fastballs and Curveballs, 2015-2020.

Proportion of Balls "Barreled" By Pitch Type

Tom Tango, an author, blogger and analyst working for Baseball Advanced Media created a statistic called "barrel." A barrel is a batted ball event with an exit velocity of at least 98 mph and a launch angle between 26-30 degrees. Major leaguers hit between .520 and .667 on balls that met this criteria in 2020. For every mph over 98, the range of launch angle bounds expands, according to mlb.com. This is because as the exit velocity increases there is a wider range of launch angles for which balls hit will most likely land safely for a hit.

In the figure below we see the percentage of strikes thrown that have been “barreled.” As you can see 4-Seam Fastballs (2.05%) are barreled more than other pitches. They are followed by Sinkers (1.88%) and Forkballs (1.77%). On the lower end you have Curveballs (1.27%), Knuckle Curves (1.25), and the rarely thrown Eephus (1%).

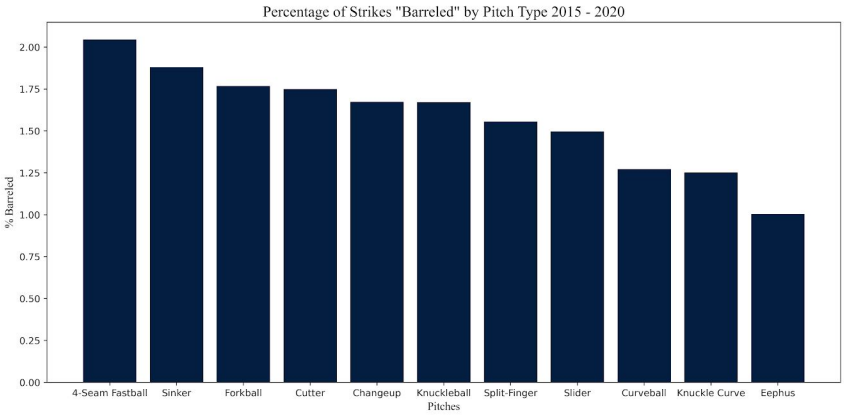


Fig 2. Percentage of strikes barreled by pitch type, 2015-2020

In fact, it seems this phenomenon of 4-Seam Fastballs being barreled more than other pitches has gotten worse, for pitchers, over the last five years. In 2015, 1.78% of 4-Seam Fastballs were barreled whereas 2.44% were barreled in 2020. The same phenomenon exists for Sinkers as well as they’ve gone from 1.58% to 2.01% barreled from 2015 to 2020, respectively.

The authors note that the observed percentage differences between each pitch type are very small, which naturally raises the question of how accurately we are able to resolve these differences, and the error in the figures presented. The relevant measurement error in this case relates to the error in statcast’s algorithm to correctly identify the pitch type - i.e., how often it incorrectly identifies a pitch by type. This data however is not available within the dataset, and could therefore not be added to the plot.

Conclusion

There is evidence in these data that hitters are getting better at barreling pitches overall (i.e., 1.5% to 2.0% from 2015 to 2020, respectively) and that any gains made by pitchers are done so on the margins by throwing more pitches that are slower and move more than the 4-Seam Fastball. In fact, the top four pitches in mph from 2015 - 2020: 4-Seam Fastball (93.3), Sinker (92.3), Cutter (88.5), and Forkball (86.3) nearly match the top four pitches by “Strikes Barreled,” with only Cutter and Forkball switching places on the “Strikes Barreled” list. A table with these values is provided in the Appendix.

This analysis suggests it’s the Curveballs, Knuckle Curves and Eephuses, the last of which are normally only thrown by position players who are subbed in during lopsided games, are the most effective pitches in baseball. While a more rigorous statistical analysis may help confirm these hypotheses our analysis suggests you may be seeing your favorite team’s first baseman taking the mound more often in 2021.

Question 2 - Elevation

The effect of altitude batted ball events has been known roughly since 1990 when Yale physics professor Robert K. Adair suggested a 400-foot drive at sea level world travel 430 feet in Denver and 450 feet in Mexico City due to their relatively higher altitudes in his book, “The Physics of Baseball.” Adair suggests

balls fly further because the air is less dense at higher altitudes, he also mentioned the same effect would cause pitches to move less, making them easier to hit. This, however, was a theoretical examination of this effect and was left largely unexamined for many years afterwards.

Later, Keith Woolner wrote an article in 2006's "Baseball Between the Numbers," where he tried to quantify this effect using outcome statistics (e.g., hits, runs ,etc.) by comparing the home and away results for a single team to determine a park's effect on that specific statistic. Meaning, he would subtract the Colorado Rockies team batting average on the road from their batting average at home to determine Coors Field's (their home park) effect on batting average.

However, now that we have statcast data we can examine the effect of the altitude on actual batted balls. Hence our question, "What can we learn about how elevation influences pitching and hitting in the statcast era?"

To study this question we will be analyzing statcast data from the same period as the Pitch Effectiveness questions (e.g., Regular Seasons from 2015 - 2020). The following are some example fields we will use to evaluate our question:

- **home_team** is a recording of who was the home team at the time of the batted ball event, this will be used as a proxy for park
- **elevation** of the park from sea level in feet
- **hit_distance_sc** is an estimate produced by statcast for how far each batted ball went
- **launch_speed** is an estimate produced by statcast for how quickly in mph, the ball has come off the bat
- **launch_angle** is an estimate produced by statcast for the angle at which the ball has been hit in degrees

We will also be looking at how elevation affects the movement of different pitches. In order to do this we will be looking at the vertical and horizontal distances the ball moves. Capturing the full effect of elevation on movement is a little difficult due to gravity and multiple factors in play, but the following variables will be added to help us approximate the effect:

- **pitch_type** refers to the type of pitch thrown (e.g., fastball, curveball, etc.)
- **pfx_x** is an estimation of the "break" (e.g., "tail" or "cut") of the pitch in the horizontal in feet from the catcher's perspective, from 40 feet away (e.g., the earliest a hitter can perceive movement)
- **pfx_z** is an estimation of "break" (e.g., "rise" or "sink) of the pitch in the vertical direction, in feet from the catcher's perspective, from 40 feet away (e.g., the earliest a hitter can perceive movement)

For further investigation of pfx_x and pfx_z please find The Anatomy of a Pitch link on the references page.

Data Cleaning

In order to accurately estimate the relationship between elevation and hit distance, we made the following additional modifications to our data in order to improve our analysis.

- *D/M/Y to Year.* We changed the date time format on our dataset to only reflect the year of the games, as we are not concerned with specific months or days.
- *Dropping Unknown Values*. We have scrubbed this batting data from our analysis as to not skew/confuse our sample.
- *Filtering for Launch Speeds and Launch Angles.* We have chosen to only look at launch speeds of above 98 MPH as to make sure that we are filtering out balls that are hit below an average speed. We are only looking at data that has a launch angle between 26 and 30 degrees, as a way of comparing similarly batted balls.

Sanity Checks

- **Reasonable hit distances:** Using '.describe()' we were able to confirm the hit distances for our specified ranges of launch speeds and angle were floats (e.g., precise) and reasonable (e.g., non-negative)
- **Elevation data:** There was not much sanity checking that needed to be done here, we just made sure all the elevations were positive and accurate
- **Pitching data (pfx_x, pfx_z):** Using .describe() on pitching variables of interest we were able to confirm they had the correct data types (e.g., float) and dropped any unknown values
- **Plotting home_team vs avg hit distances:** Before loading in our elevation data, we wanted to see if Coors field had the highest hit distance average, as we knew from research it was the stadium at the highest elevation. Below is an initial plot of the hit distance average versus home team (which is a proxy for stadium location). From the figure, which can be found in the Appendix, we confirmed Coors field (i.e., COL) had the highest average hit distance

Hit Distances and Elevation

To further examine this effect we modeled the relationship between average hit distance and elevation for each team/park. To do this we used a scatter-plot and a linear regression of our two variables. Based on Adair's book, it would follow that the stadiums with the highest elevations would have the highest average hit distances for a set range of launch angles and a set launch speed. Therefore, we expect to see a positive correlation between average hit distance and elevation. For reference, Colorado's stadium has the highest elevation and Philadelphia has the lowest. A table is provided in the Appendix along with other exploratory analyses. Below is our linear regression describing the relationship between Elevation and Hit Distance.

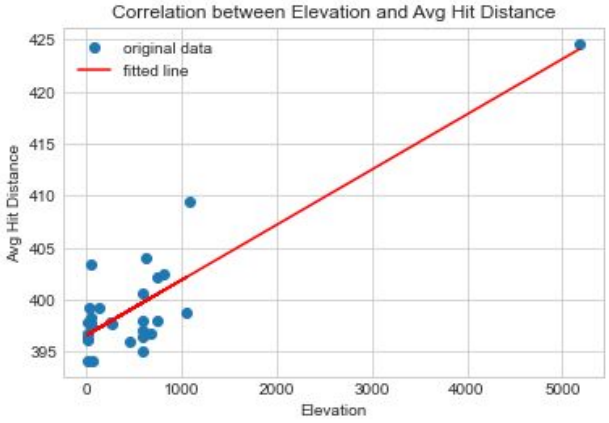


Fig 3. Scatter plot of average hit distance vs elevation.

Our scatter plot displays a linear relationship between our variables of interest. We found a slope of 0.0053, an intercept of 396.5836, and an R^2 of 0.7580. Our high positive R^2 value indicates that our linear model explains about 75% of the data. In the Appendix we've provided the scatter plot without Colorado which shows a weaker relationship. However, we've also provided an explanation of the physics involved in this phenomenon which explains why the effect is more dramatic at higher altitudes.

Pitch Movement and Elevation

To study the effect of elevation on pitch movement, we focused on the horizontal and vertical movement of the ball at different stadiums, as defined in "The Anatomy of a Pitch" article cited in our references.

To illustrate these differences we chose to highlight the stadium with the highest elevation (i.e., Colorado) and the one with the lowest elevation, (i.e., Philadelphia). We expect that at higher altitudes the ball will move less. This would theoretically, make them less effective, contributing to the effects observed by Woolner.

The figure below illustrates the horizontal and vertical movement of Curveballs pitched in Colorado (in purple) and Philadelphia (in red). The Curveballs in Colorado have a smaller area and therefore have less horizontal and vertical movement. A similar view for Sliders and 4-Seam Fastballs is provided in the Appendix. We note that the effect seems less pronounced for 4-Seam Fastballs. Further analysis is required to confirm these differences.

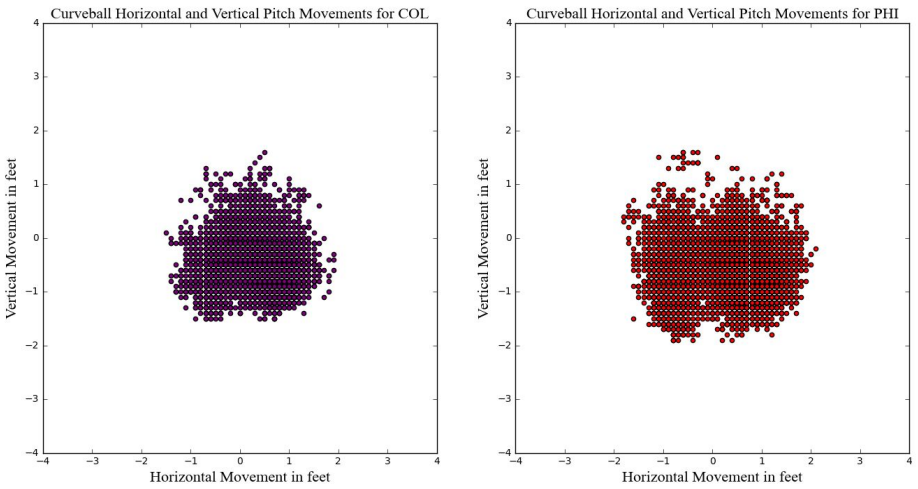


Fig 4. Scatterplot of vertical vs horizontal movement for curveballs in Colorado (left) vs Philadelphia (right).

Conclusion

Overall, Adair's theoretical examination seems to be supported by the statcast data. The data shows that elevation has a strong linear relationship with hit distance, on average. Though the effect might not seem quite as dramatic for parks that don't have significant differences in altitude, elevation does make a difference. For balls hit within the same range of angles and speed, on average we can expect that they will generally travel farther in stadiums with higher elevation. Though capturing movement of pitches is a little more difficult, overall we can see that elevation has an effect on pitches and their effectiveness as well. Higher elevations cause pitched balls to have less horizontal and vertical movement, making them easier for batters to hit. Teams looking to construct or move stadiums should keep this in mind!

Question 3 - Clutch Hitting

The existence of “clutch hitting,” meaning, the ability of a hitter to get hits when his team needed it most, was a long held piece of conventional wisdom in baseball history. It seems plausible right? After all, some of the game’s best players of all time were known for their clutch hitting: Derek Jeter, David Ortiz, Roberto Clemente and Joe DiMaggio.

This idea inspired many to ask if the best clutch hitters are simply the best hitters and their ability to deliver in the clutch is imperceptible from their ability to deliver at any point during a game or season. In 1977’s Baseball Research Journal Dick Cramer asked, “Do clutch hitters exist?” in one of the seminal pieces of advanced baseball research. Others, including Nate Silver, suggested any observed differences in “clutch” hitting are due to small sample sizes.

As a result, this became the new conventional wisdom, that clutch hitting doesn’t exist. However, these analyses relied on outcome data as it was all that was available. The problem with outcome data are many factors (e.g., quality of pitching faced, quality of defenders faced, park, etc.) contribute to whether a batted ball lands safely as a hit or not. Our hypothesis is all a hitter can really control is how well they hit a given pitch in a clutch situation.

For this reason we ask, “During the statcast era, are there players who have been able to consistently demonstrate an ability to hit the ball better in the clutch than in non-clutch situations?” Only the statcast data provides us with an ability to isolate a hitter’s effort completely. In this context, hitting a ball better would imply the ball is hit with an exit velocity and launch angle needed for home runs and extra base hits.

As before we will study regular season statcast data from 2015 - 2020. The following are some the fields we will use to evaluate our question:

- **home_score** from statcast, records the pre-pitch home score
- **away_score** from statcast, the pre-pitch away score
- **inning** from statcast, the pre-pitch inning number
- **hit_distance_sc** an estimate produced by statcast for how far each batted ball went
- **launch_speed** an estimate produced by statcast for how quickly in mph, the ball has come off the bat
- **launch_speed_angle** a combined metric from statcast that estimates how “well” a ball was hit (e.g., on the barrel)
- **xBA_from_launch_angle** a combined metric which measures the likelihood that a batted ball will become a hit, based on the outcome of similar balls with similar hit characteristics.

Specifically, we will be comparing a player’s performance across these metrics in the 7th inning or later, when the batting team is ahead one run or tying, or in the event that the batting team is trailing, that the runs to tie the match or lead are ‘on deck.’ To illustrate this definition, imagine the batting team is trailing 3 runs, in the 7th inning, and there is a runner on third base. This is defined as a clutch situation, because the batting team can score 2 runs from the play, setting up the next batter to equalise if they hit a home run. The logic for this was implemented as a filter to sort clutch and non-clutch events (rows - each row in our database is a game event) which translates to:

$$number\ of\ loaded\ bases\ +\ 2\ >=\ trailing\ score\ difference. \quad \text{Eq. 1}$$

Note, according to our definition, the trailing score difference in Eq. 1 is always negative. As an additional point, we filtered the sample set to only include players who have recorded over 200 batting events in clutch situations during our sample period, in order to reduce the likelihood of anomaly and 'one-off' performances events affecting our results.

Data Cleaning

The following columns have been cleaned:

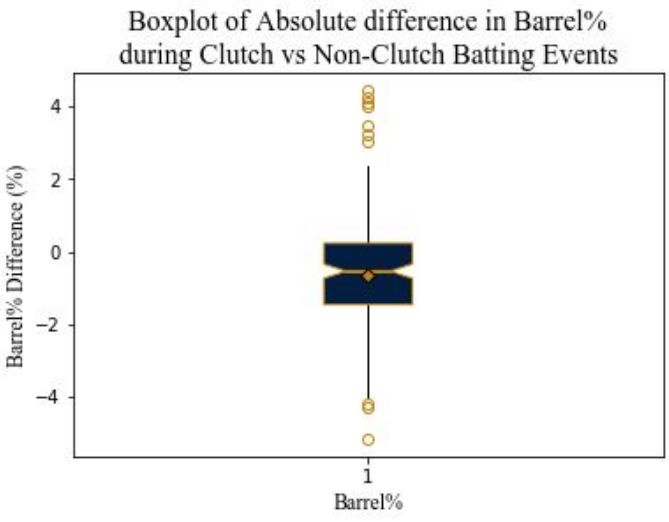
- *game_date*: Contains the game date and was renamed to *year* and converted into a string.
- *Game_type*: Regular season games only.
- *launch_speed_angle*: All rows with empty values in this column have been removed, since these relate to non-batter events (i.e. the batter did not make contact with the ball).
- *on_1b, on_2b* and *on_3b*: Originally these columns gave the MLB playerID of the player on that base, or NaN if the base wasn't loaded prior to that event occurring. So that we could calculate the number of loaded bases, the NaN values were converted to zeroes, and the playerIDs converted to ones.

Sanity checks

- Base status: We confirmed, by checking the unique values within each column, there were only ones and zeros in the columns *on_1b*, *on_2b* and *on_3b*. Similarly, we found the number of rows in each column are equal, and given that the whole data frame has 3856615 rows, we can see that all rows are accounted for.
- The launch speed angle contains a numeric value of between 1 to 6 inclusive and empty for non-batting events. We confirmed this was the case.

Barrel%

For our main analysis of clutch hitting we will use the same barrel percentage we used in the effective pitches analysis but consider it from the hitter's perspective. Specifically, we will compare barrel percentages for hitters between clutch and non-clutch situations (as described above). We then subtract the percentage of "barrelled" balls in non-clutch situations from the percentage of "barreled" balls in clutch situations. In this way, a positive value shows the batter tends to make better contact more often, on a percentage basis, during clutch situations. After filtering for a minimum of 200 events, we were left with a sample size of 177. The boxplot shown here illustrates the distribution



Absolute difference in player Barrel% in clutch vs non-clutch situations

of this difference for this sample. A bar chart showing each individual player's score is shown in Appendix C. Also included in Appendix C are other examinations of player performances in the clutch.

For roughly two-thirds (67%) of these players, their percentage of barreled balls in clutch situations decreases, suggesting that most players' batting performance reduces in clutch situations. The average decrease for all players with more than 200 batted clutch events was -0.63% absolute. Finally, the largest positive difference was calculated for Avisail Garcia, with a barrelled percentage 4.4% higher in clutch situations relative to his non-clutch performance, which is roughly 4 more balls in 100, whereas the largest negative difference was calculated for Cory Seager, whose barrelled percentage dropped 5.15%, or roughly 5 less barrelled balls in 100 during clutch situations. While some of this may be explained by Seager facing left-handed specialists or better pitching because he is an overall better player, it is likely that the opposing team has their best relief pitchers pitching in these situations because they are so critical. It is also interesting to note that Seager was awarded the 2020 World Series Most Valuable Player (MVP) award, an award given for performing best on the game's biggest stage.

The standard deviation of the clutch vs non-clutch barrel percentage difference for all players was 1.62 %, and is a compound metric of no physical significance, but is useful to assist in sorting the players, to identify those who exhibit clutch vs non-clutch differences of magnitude greater than similar players in their category. A similar standard deviation is calculated throughout this section for other batting metrics simply to assist in this categorization effort.

From our analysis, we would estimate that there is a roughly two in three chance that the average player in a clutch situation will experience a decrease in batting performance, and we have previously calculated the average extent batting performance would deteriorate for each statistic. But how should this affect a team manager or coaches decision making?

Conclusion

Our analysis suggests some players do tend to perform consistently better in clutch situations whereas others tend to perform consistently worse, even though, overall, most players tend to perform slightly worse in clutch situations. However, the magnitude of average performance change, both positive and negative, is not so large that a high performing player in good form - even with a poor response in the clutch situation - is now likely on the same level as a mid-low range hitter with a better response under pressure. There are extreme cases where a player's response in the clutch - either positively or negatively - may be enough to push them into a level seen in their best/worst seasons.

While we do have evidence clutch hitting exists for a handful of extreme cases, these variances need to be put into context by comparing interseason performance, overall performance and the player's current form. Meaning, even if one of your better players is likely to underperform, relative to themselves, he won't likely underperform to the extent that it makes sense to send out an overall worse player who copes better under pressure.

References

- The Fastball Is Disappearing. What Does It Mean for MLB's Future?
- The Anatomy of a Pitch: Doing Physics with PITCHf/x Data
- "Baseball Between the Numbers." California Bookwatch, Aug. 2006
- Adair, Robert Kemp. The Physics of Baseball. 3rd ed., Updated, And expanded., Perennial, 2002
- Do Clutch Hitter Exist?

Appendix

Statcast Seasons

Year	Season Dates	No.Games	No.Teams
2020	July 23 – October 27	60	30
2019	March 20 – October 30	162	30
2018	March 29 – October 28	162	30
2017	April 2 – November 1	162	30
2016	April 3 – November 2	162	30
2015	April 5 – November 1	162	30

Appendix A - Pitch Effectiveness

Table A1 Pitch Speed by Type (2015 - 2016)

Pitches Names	Average Pitch Speed (mph)
4-Seam Fastball	93.3
Sinker	92.3
Cutter	88.5
Forkball	86.4
Split-Finger	85.0
Slider	84.6
Changeup	84.1
Knuckle Curve	80.8
Curveball	78.2
Knuckleball	76.2
Eephus	67.3

Pitch Mix

It seems that Verducci is, at least, directionally correct in his assessment that pitchers are altering their approach. The pitch mix (e.g., how often each pitch has thrown) has changed over the past five years. In the figure below we can see that 4-Seam Fastballs and Sinkers (seen in blue) have dropped in usage from 2015 - 2020, while Sliders, Curveballs, Changeups, Curveballs and Cutters (seen in gold) have been increasing their share. Knuckle Curves, Split Fingers, Knuckleballs, Forkballs and Eephuses (seen in gray) have all gone down but were very lightly used to begin with.

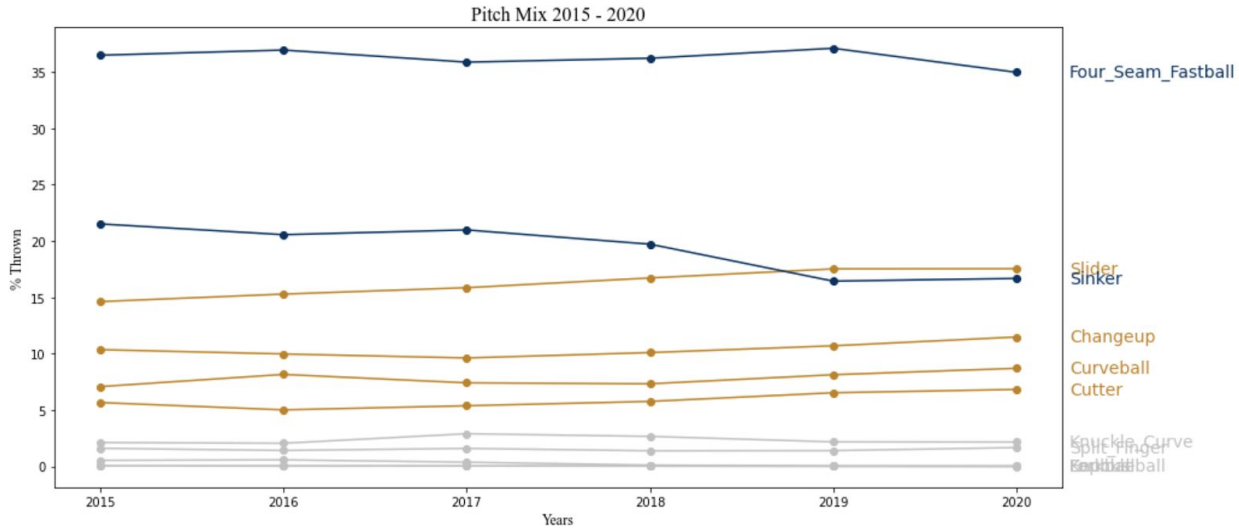


Fig A1. Percentage of pitches thrown by pitch type, 2015-2020.

Appendix B - Elevation

Hit Distances by Ballpark

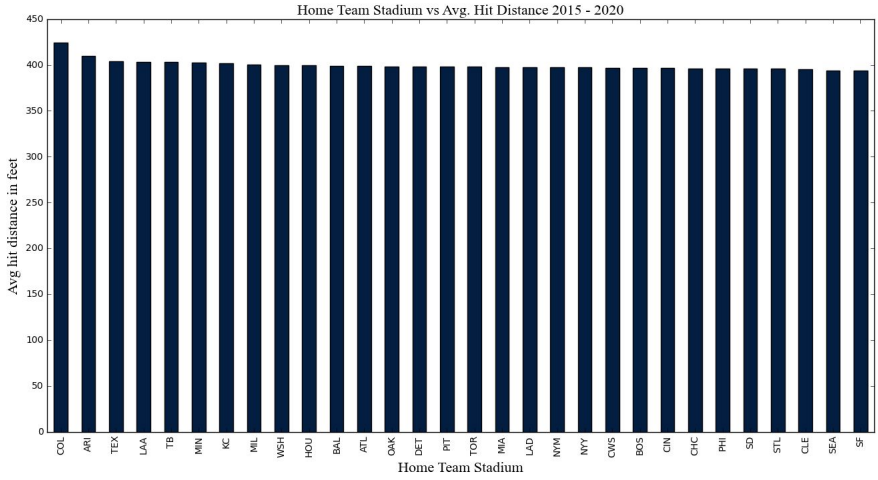


Fig B1. Average hit distance for each stadium, 2015-2020

Hit Distances and Elevation

This figure above provides distributions of hit distances for each home team/stadium. The notched box plots show that the medians do increase overall with elevation.

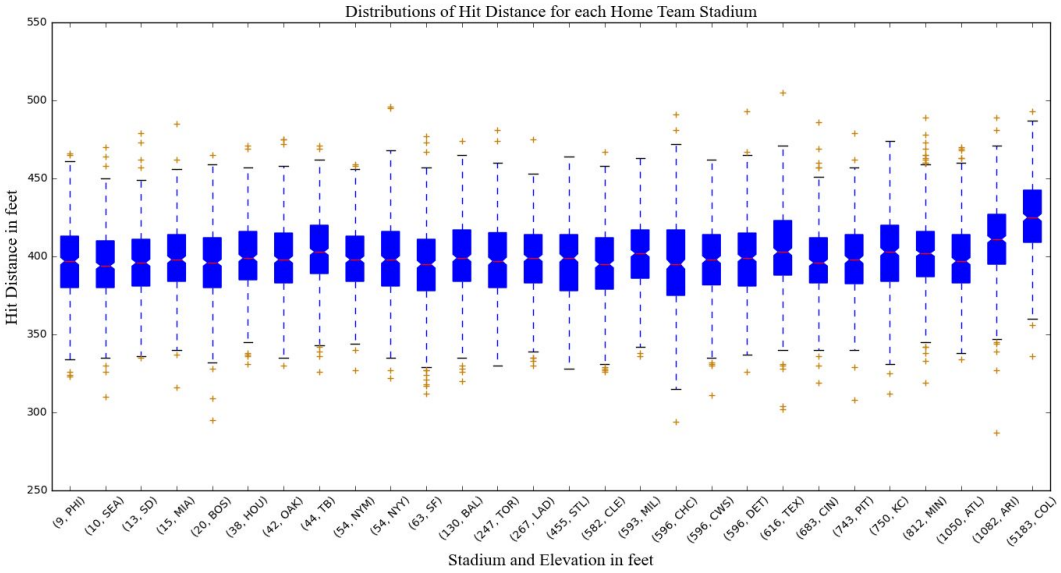


Fig B2. Boxplot of hit distance for each stadium, including stadium altitude.

Ballpark Elevations

team	elevation_in_feet
COL	5183
ARI	1082
ATL	1050
MIN	812
KC	750
PIT	743
CIN	683
TEX	616
CWS	596
CHC	596
DET	596
MIL	593
CLE	582
STL	455
LAD	267
TOR	247
ANA	160
BAL	130
SF	63
NYY	54
NYM	54
TB	44
OAK	42
HOU	38
WAS	25
BOS	20
MIA	15
SD	13
SEA	10
PHI	9

Hit Distance and Elevations Without Colorado

Below, you will see our scatter plot without Colorado:

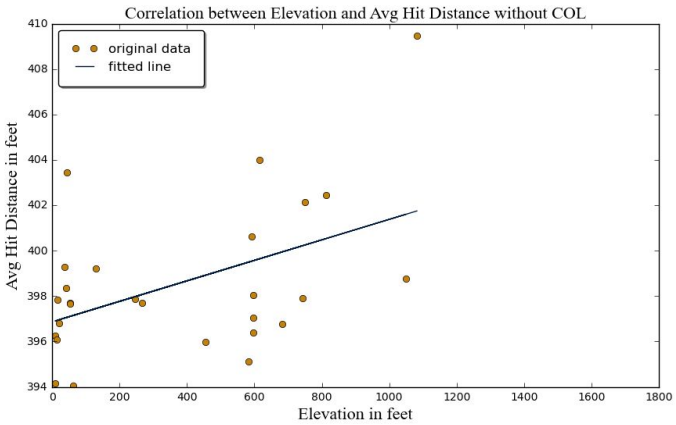


Fig B3. Scatter plot of average hit distance vs elevation, not including Colorado.

Without Col, our scatter plot has a slope of 0.004515, and intercept of intercept: 396.86, and an R-squared of 0.2231. The relationship between stadiums within 0-1000ft show more scatter, indicating that the relationship between hit distance and elevation is not as strong at these lower altitudes. In the following section, we will discuss the physical reasons as to why this is the case.

Air Density Versus Hit Distance

The drag resistance F_D on a particle in motion within fluid is given by the formula:

$$F_D = \frac{\rho v^2 C_D A}{2} \qquad \text{Eq. B1}$$

Where v is the particle velocity, A is the effective area, C_D is the drag coefficient, and ρ is the fluid density. In our analysis where a ball is travelling through the air, air is the effective fluid and therefore the resistive force on the baseball is directly proportional to the air density. The air density is a function of altitude as the density of air tends to decrease with altitude. This relationship is shown in Fig B1 below.

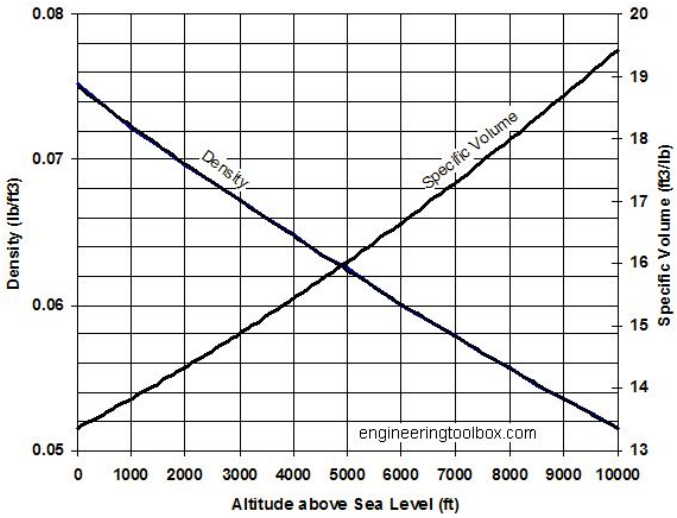


Fig B4: Relationship between altitude (ft) and air density(lb/cubic feet). source:https://www.engineeringtoolbox.com

Fig. B1 shows that the air at 5000ft is 0.012 lb/cubic feet less dense than the air at sea level. Therefore, a baseball in flight at 5000ft has roughly 16% less drag force exerted on it during flight compared to that at sea level. In contrast, a ball hit at 1000 ft travels through air which is roughly 0.03 lb/cubic feet less dense than that at sea level, which is 4% less dense than at sea level. Since there is a direct linear relationship with the altitude and the drag force on the baseball, we would expect that due to the relatively smaller difference in air density in the 0-1000ft altitude range, that the elevation will not have as strong influence on the hit distance as when the ball is hit at 5000 ft, where there is a four fold decrease in drag resistance due to the altitude dependent air density. This explains the poorer R^2 value of 0.2231 in the linear regression calculated for balls hit within the 0-1000 ft range, and the improved R^2 value of 0.7580, when we include the 5000 ft datapoint. By effectively increasing the altitude range, we are able to test a broader range of data points, and lessen the extent of noise (from for instance humidity, wind direction and speed) on the data at lower altitude and hence see the true dependency.

Pitch Movement: 4-Seam Fastballs and Sliders

We studied 4-Seam Fastballs which we know from our Pitch Effectiveness research tend to move less vertically than Curveballs. In the figure below this effect seems less pronounced than it did for Curveballs. Sliders show a similar pattern to what we observed with Curveballs.

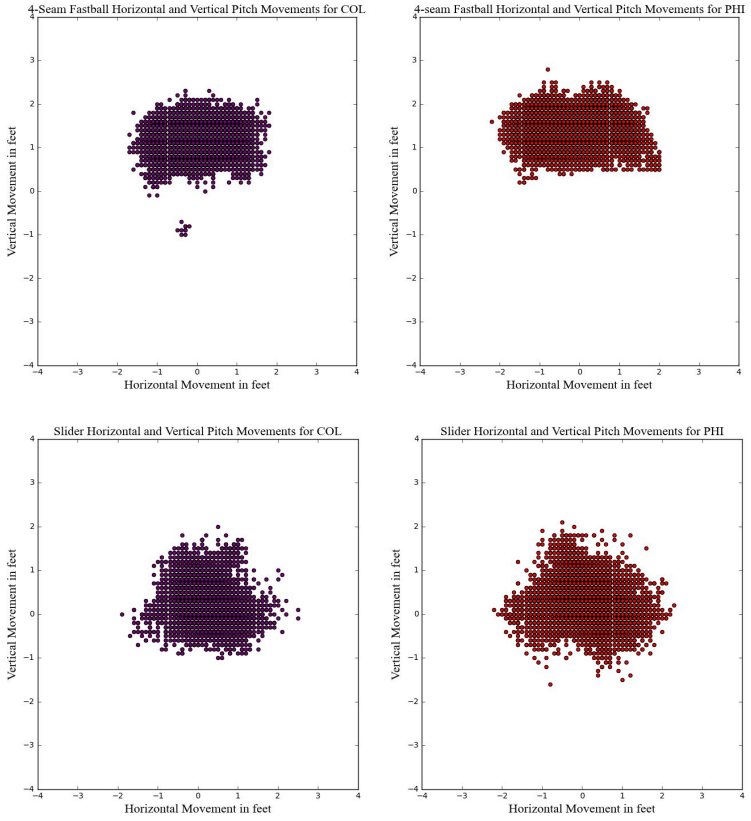


Fig B5. Scatter plot of vertical vs horizontal motion for pitches thrown in colorado (left) and philadelphia (right), by pitch type, including fastballs (top) and sliders (bottom)

Appendix C - Clutch Hitting

Base Status Check

We can check this by taking these columns and looking at the unique values and number of entries:

Function	on_1b	on_2b	on_3b
unique	[0.0, 1.0]	[1.0, 0.0]	[0.0, 1.0]
count	3856615	3856615	3856615

Launch Speed Angle Check

Dataframe	Function	Output
clutch_df	unique	[3.0, 2.0, 6.0, 5.0, 4.0, 1.0]
	count	569660
non_clutch_df	unique	[3.0, 2.0, 6.0, 5.0, 4.0, 1.0]
	count	96827

Estimated Batting Average from Launch Angle:

An estimated batting average can be calculated based on the launch angle that the player hits the ball. It can be thought of as a summary statistic, which presents the outcome from a combination of characteristics. In comparison to real world outcomes, it can be used to compare whether a player is under or over-performing. We looked at the relative difference in individual estimated batting averages for players in clutch and non-clutch situations, and this data is shown in Figure C1 below.

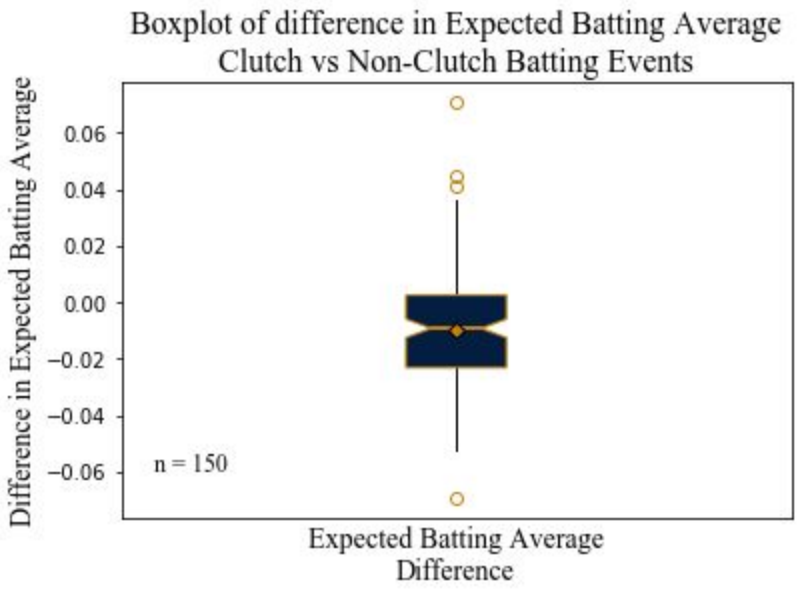


Fig C.1 Difference in player expected batting average from speed angle in Clutch vs. Non-clutch batting events.

The majority of players (70%) will experience a decrease in their estimated batting average by on average -0.01 pts and the standard deviation of the difference between clutch and non-clutch variance for all players is 0.02 pts.

Hit Distance:

A similar analysis was performed for the hit distance. In this analysis, the average hit distance during clutch events was calculated, and the average hit distance during non-clutch events was subtracted from that value by player, for players with more than 200 recorded clutch batting events. The results of the analysis are shown in Fig C2 below.

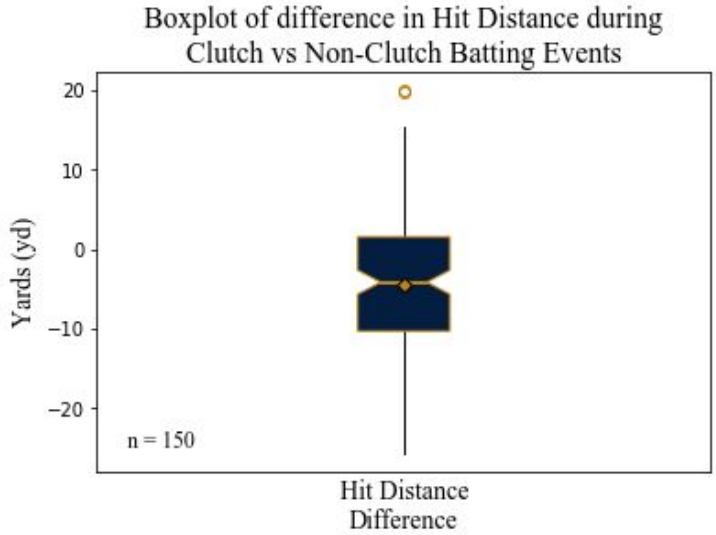


Fig C2 Boxplot of the difference in average hit distance by player during clutch vs non-clutch situations.

Fig C2 shows again roughly two-thirds (68%) of players hit on average 4.6 yards shorter during clutch situations, suggesting that for most players, their performance tends to decrease in high-pressure situations. The standard deviation of all differences was 9.5 yards, and the maximum increase and decrease in hit distance was +19.9 yards and -25.6 yards for Brandon Belt and Justin Turner respectively.

Launch Speed:

To gauge whether players hit harder during clutch situations, we averaged the launch speed of players in clutch and non-clutch situations, and subtracted the latter from the former, for players with over 200 recorded clutch batting events. In this way, an increase in the batting speed will be shown as a positive value, and the trend for all players meeting the minimum batting event criteria is shown in Fig C3 below.

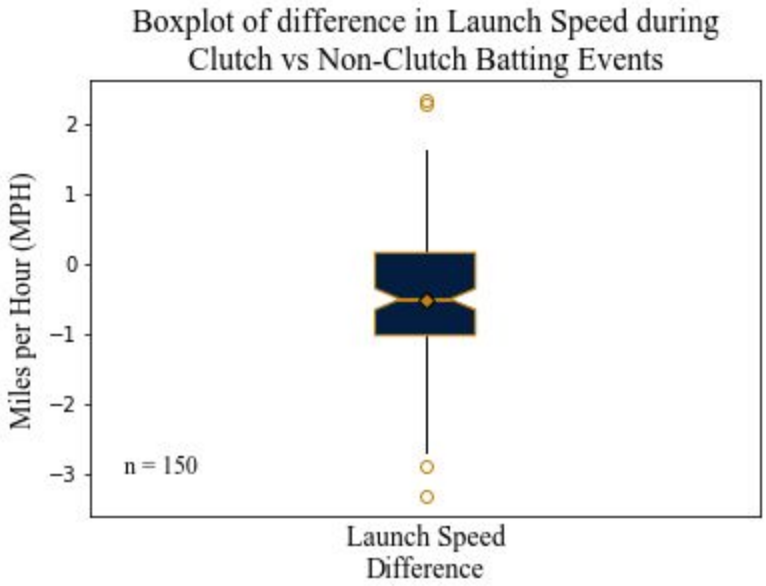


Fig C3 Difference in launch speed by player during clutch and non-clutch batting events.

Again, most players tend to hit the ball slower in clutch situations by on average 0.5 miles per hour than during non-clutch situations. The standard deviation in this case was 1 mile per hour, and the maximum increase and decrease in launch speed was 2.3 and -3.3 miles per hour for Victor Martinez and Georger Springer respectively.

Clutch Hitting Vignette

If we return to our original statistic on Barrel %, let's take the two players on the extremes, Avisail Garcia and Corey Seager. Imagine that you are the coach of a fictional major league team and both Garcia and Seager are sitting on your bench. It's the first game of 2021 and you are down in the last inning, with loaded bases such that the tying runs are in hand and it's the last play. Let's also assume that it is the beginning of 2021 and the 2020 player stats are the most current indicator you have of current player form as reference, and therefore we will be using 2020 statistics as the reference for player performance. You can only send either Seager or Garcia out, who should you send out to bat?

On the one hand, you have Garcia, who in 2020 ranked no. 341 by Barrel % in 2020 with a Barrel % of 3.8 %. On the other hand, you have Seager, 2 x All star and 2020 World Series MVP who ranked 24th by Barrel % in 2020 with a Barrel % of 16.1 %. However, our analysis has shown that Garcia's Barrel % increases by almost 4.5 % in clutch situations, whereas Seager's Barrel % tends to drop by more than 5 % on average in clutch events. If we were to boost Garcia's Barrel % by 4.5 % and reduce Seagers's by 5 % to roughly compensate for the clutch effect, Seager would still have the theoretically better Barrel %, meaning he is still more likely to hit the ball on the barrel of the bat, even though he doesn't respond as well in clutch situations as Garcia.

We can apply this analysis to the batting average statistic, which is displayed fully in this Appendix. Garcia's and Seager's 2020 expected batting average in regular season games was 0.238 and 0.307 respectively. Our analysis on the expected batting average from speed angle in the preceding section showed that Garcia's expected batting average is likely to increase by 0.022, whereas Seager's is likely to decrease by 0.044. Although closer this time, again this would not be enough to bring Garcia's clutch-compensated expected batting average into a similar level as Seager's in 2020. Again, Seager would be the wiser choice, even though we have chosen the two extreme players from the perspective of clutch reponse.

An interesting perspective is brought into this analysis when performance over time metrics are shown, which depict the magnitude of performance variance season to season. Table C1 shows the average annual Barrel %, average expected batting average from speed angle (xBA from Speed Angle) for regular season games, and the full year batting averages for Garcia and Seager respectively. Note empty rows indicate missing data in the dataset. A summary of the average and standard deviation of these values across all years is shown in Table C2 below.

Table C1 Annual Batting Statistics for Garcia and Seager

Name	year	Barrel%	xBA from Speed Angle	Batting Average
Avisail Garcia	2015	4.963	0.356	0.257
	2016	8.108	0.363	-
	2017	8.845	0.368	0.330
	2018	11.583	0.339	-
	2019	11.717	0.359	0.282
	2020	3.788	0.317	0.238
Corey Seager	2015	10.256	0.400	-
	2016	8.350	0.390	0.308
	2017	8.537	0.399	0.295
	2018	8.333	0.360	-
	2019	7.342	0.325	0.272
	2020	16.092	0.396	0.307

Now it is possible to contrast the typical variance in performance Garcia and Seager exhibit season to season against the variance in performance we may expect based on i) the typical variance of all players

in clutch situations, and ii) the typical variance of these individual players (i.e. Garcia and Seager) in clutch situations. Note here we are assuming that the players we analysed in clutch situations exhibit on aggregate clutch responses which are representative of all players. Performing statistical analysis to calculate the degree to which this sample is representative of all batters is outside the scope of this investigation. Table C1 and C2 show that the variance season to season for Seager and Garcia players (3.29 and 3.22%) respectively is greater than the variance we would expect to observe in the average players from being in a clutch situation (-0.63%). We have also calculated that the median season-to-season standard deviation in Barrel % for all players between 2015-2020 is 2.66%. Therefore, if we didn't have data on Garcia and Seager, we would argue that from the typical season to season variance of all MLB players, we would expect that the form of the player will be a stronger factor than their performance in clutch situations. However, since we do have data from Seager and Garcia, and because we know that these are example players who are at either extreme of clutch response, it is interesting to note that their variance in Barrel % and expected batting average, is actually greater than the standard deviation of their typical annual variation in these statistics.

Table C2. Annual Batting Statistics for Garcia and Seager

Name / Statistic	**Barrel%**		
	Interseason st.dev.	**Clutch Difference**	**2020 Barrel%**
Garcia	3.29%	+4.40%	3.8%
Seager	3.22%	-5.15%	16.1%

Thus going back to our hypothetical, the coach should send out Corey Seager to try and rescue the match, he's the better player on paper and happens to be in form. Even though he is likely to underperform, based on his 2020 form, he won't likely underperform to the extent that it makes sense to send out a player who copes better under pressure.

Fig C.4 Bar Chart of Barrel% Difference during clutch vs non-clutch positions, by player.