Can Stolen Base Decisions be Efficiently Simplified?

Exploring the decision-making process for runners and catchers as they attempt to gain an advantage in one of baseball's most underrated tactics.

GitHub repo URL: https://github.com/yangjus/smt-data-challenge-2023

Abstract

Objective: Stolen bases have been a vital component of baseball since the game's inception, but there are numerous strategic decisions that go into the success or failure of the catcher's throw attempt. We wish to determine which statistics best indicate the success or failure of a catcher's throw as well as draw conclusions on how the data can be used to predict outcomes in the future.

Methods: Given a large dataset containing information regarding ball position, player position, and game events, it was vital to reduce our data to that involving stolen bases. We filtered out most of the data, leaving just the play sequences in which there was a catcher attempting to throw out a runner going to second base. Then, we could begin calculating and comparing the pop time and throw speed of the catcher as well as the runner's leadoff distance.

Results: Before we could determine if our statistics were indicative of one another, we needed to ensure these statistics were important to stolen base success. Our grouped boxplots determined that high throw speed and low pop time lead to advantages for catchers in stolen base situations. Additionally, larger leadoffs lead to more success for runners. We also created a scatter plot to determine how large of a leadoff a runner needs to increase his probability of successfully stealing second base. However, the data is inconclusive, so we determined it would be more beneficial to analyze the catcher's impact on the play. So, we created a scatter plot displaying throw speed and pop time to find the "sweet spot" that catchers can target. From this, we determined catchers should average around 0.75 seconds for their pop time and throw the ball between 30 and 35 meters per second.

Conclusions: From this experiment, we were able to determine how baseball tracking and statistics can lead to a team changing approaches over time. We determined that pop time and throw speed were vital to the success of a catcher but must be related to one another in an efficient manner. Similar trends can be found across the diamond, which have led and will continue to lead to the evolution of baseball as a sport.

1. Introduction

Throughout the constant action of baseball, the concept of the stolen base is one that is overlooked. The ability to move a runner into scoring position or to open the door to a sacrifice fly RBI is essential to the final outcome of a game. While it may seem simple, stolen bases have significant analytical promise, as base coaches are constantly evaluating runner speed, pop times, and the catcher's throw speed to determine when it is appropriate to attempt to steal a base.

As athletic training intensifies and is more scientifically proven, we have seen a clear uptick in stolen bases. Young speedsters like Ronald Acuna Jr., Esteury Ruiz, and Bobby Witt Jr. have taken over the game of baseball, bringing new value to having speed. Teams have responded to this by adding players to their roster not because of their talent at the plate or on defense but because of their usability as a pinch runner in tough situations. The importance of analyzing stolen bases is ever more crucial with the recent change in MLB rules that allowed the rate of stolen base attempts to soar.[1]

Through this project, we hope to find out what factors play a role in the success of a stolen base and what thresholds need to be met. This can be visually demonstrated to help inform analytics departments of "sweet spots" that allow for more success in the stolen base interaction. Since stolen base attempts can be simulated on runner and catcher-based velocities, there is a strong case to be made that stolen base analytics are among the most important for coaches. This is especially true when teams find themselves in a close game late, where one misstep can make all the difference.

2. Methodology

When first evaluating stolen bases, it is important to look at the key variable factors that are in play. The ones that stuck out were catcher release time, catcher throw speed, and the magnitude of the runner's lead. This was found by evaluating the runner's position relative to first base when the pitcher released the ball. Since some of these statistics affect each other, we thought it would be insightful to visualize them against one another as well.

Identifying the Game Event Sequence

Since attempted steals of second base are much more apparent in the game of baseball, along with other considerations[2], we chose to focus on those events. We first had to filter the game event data in order to achieve this. The sequence of game event codes[3] that correspond to a stolen base attempt were:

- (1, 1): pitcher pitches ball
- (2, 2): catcher catches the pitch
- (2, 3): catcher throws the ball
- (4/6, 2): fielder receives the ball (4 is second baseman, 6 is shortstop)
- (0, 5): end of play

In order to filter the data based on this sequence of game event codes, we first combined all *game_events* tables together and then grouped them by the *game_id* and *play_id* columns since each stolen base scenario will be a different play. By filtering these groups with the *player_position* and *event_code* columns, we can identify which plays have the (2, 3), (4/6, 2) combination. We also ensured that groups containing these particular combinations only have one instance and that the (2, 3) and (4/6, 2) rows are consecutive.

Verifying Stolen Base Attempts

Next, we grouped the stolen base attempts based on success. We combined the game event data with the plays where the sequence of game event codes indicated a stolen base attempt. This data displayed the catcher, infielder, and first base runner's coordinate position at every play, which gave insight into the result of the attempt. After an initial sorting of the game event data, there were many cases where a stolen base attempt was not made at all. One example was when there

was no change among the batter and baserunners before and after the play. We had to omit all unclear cases in order to protect the integrity of the data.

After removing all unclear cases, we evaluated the remaining cases where the first play was at the top of the inning and the next was at the bottom, or vice versa. We decided to include these cases because it is more than likely that the last play before the change of sides involved a stolen base. These scenarios are what made up our stolen bases dataset used for analysis. We were able to later join these instances of stolen base attempt successes and failures to our other cleaned tables for analysis.

Although we were able to narrow down the event code to a collection of stolen base attempts, there are some important edge cases that may have caused some stolen base instances to be missing from our analysis. As the catcher attempts to throw out a runner at second base, the ball may bounce on the ground before it reaches the fielder covering the base. This scenario would be identified by a (255, 16) event code in between the catcher's throw and the fielder's collection. These sequences were filtered out and thus not included in our collection of stolen base attempts.

Similarly, we do not have any instances where the ball bounced before reaching the catcher. This includes wild pitches and passed balls, which we particularly chose to omit. These scenarios result in a significantly higher success rate for runners, which could skew the results. Additionally, catcher pop time varies significantly when a bouncing ball is in play, and in the event of a catcher releasing the ball from the backstop, his throw velocity could not be compared to a standard sequence as the distance from second base would be greater.

Catcher's Throw Velocity

Part of our analysis involved the ball's velocity as it leaves the catcher during stolen base plays. To do this, we took our data frames that contained all stolen base sequences and joined the corresponding *player_pos* table to these rows, doing a left join based on the *game_str*, *play_id*, and *timestamp* columns to get the coordinates of the catcher and the infielder at these exact event occurrences. We focused on the catcher's coordinates and timestamp right when he throws the ball and the corresponding infielder's coordinates and timestamp right when he catches the ball

from the catcher. Then, using the euclidean distance formula[4] , we calculated the distance and time elapsed between the catcher's release to the infielder's catch. Finally the resulting velocity (m/s) is calculated after converting to the appropriate units and dividing distance over time. We removed resulting rows with velocities less than 15 m/s because we found that almost all are edge cases (e.g. case where ball bounces from the catcher's hand) and can be excluded since it won't affect our analysis.

First Base Runner Leadoff Distance

To determine the "sweet spot" of the first base runner's leadoff distance, we first found the distance between the first base runner and the first base plate when the catcher catches the ball for a stolen base attempt. Our strategy is to use the *player_pos* table to get the coordinates of the first base runner and the timestamp where the catcher catches the ball right after the pitch. Then we used the Euclidean distance formula[4] based on the first base plate's position of (62.76, 63.64) at that timestamp.

Catcher's Pop Time

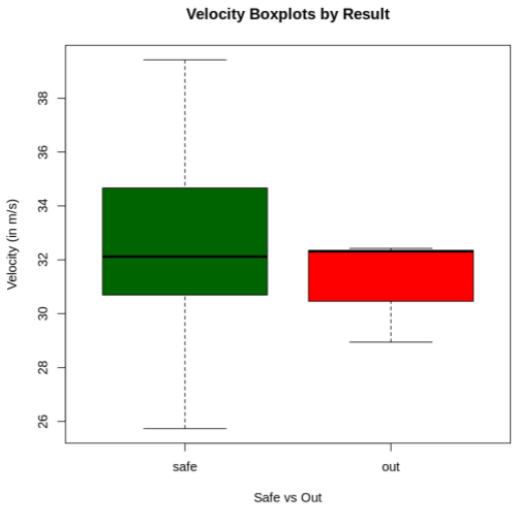
In order to evaluate the catcher's pop time, we used the timestamps found at the *player_position* and *event_code* coordinates at (2,2) and (2,3) in order to find the time it took for the catcher to catch the pitch and release the ball towards second base. We did this by calculating the difference between the two event timestamps and joining the success of the steal attempt with the pop time and throw velocity to develop our plot (see Figure 3).

3. Findings

When we first began to do analysis on our data, we wanted to ensure that the statistics we were targeting were appropriate and significant contributors to the effect of a stolen base attempt. So, we created side-by-side boxplots displaying throw velocity, pop time, and leadoff distance for safe and outplay results. We also calculated the mean of each statistic for each result to gain an additional understanding of how these statistics impact the play. The results can be found below in Figure 1.

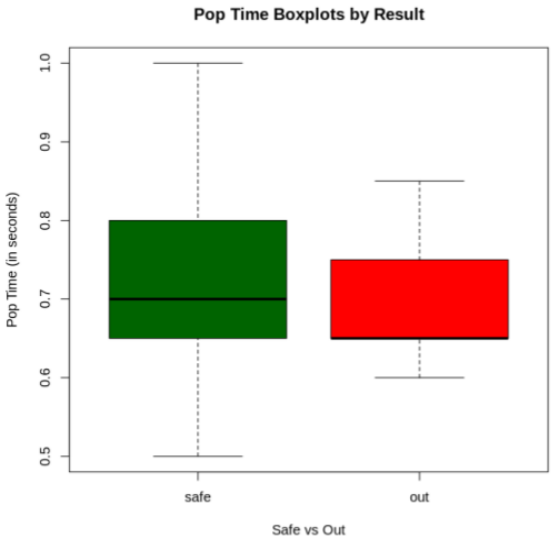
Figure 1

Throw Velocity

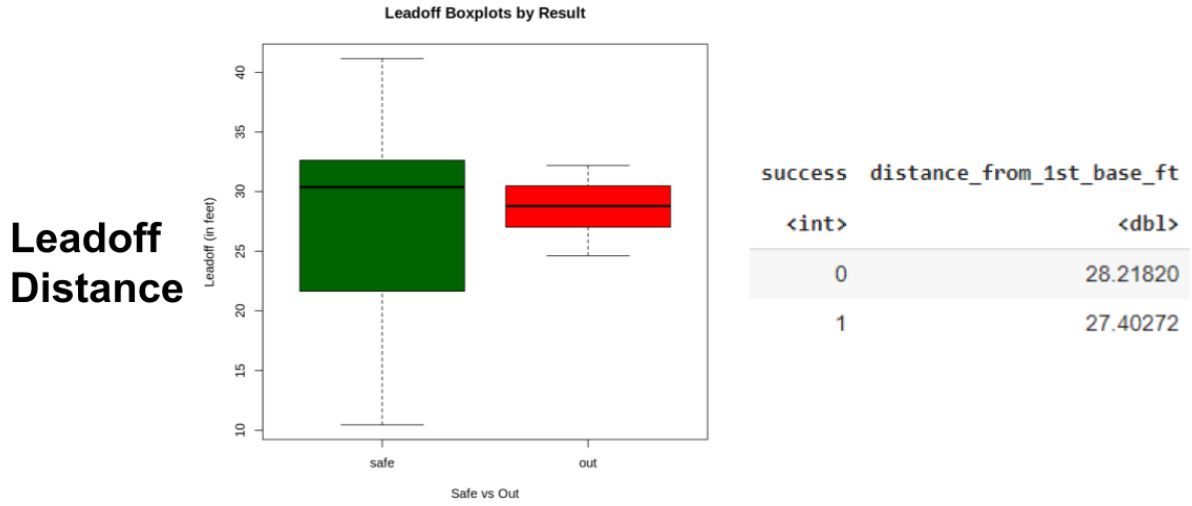


success	velocity_m_per_s
<int>	<dbl>
0	32.62926
1	31.89337

Pop Time



success	pop_time
<int>	<dbl>
0	0.799717
1	0.695000



success	distance_from_1st_base_ft
<int>	<dbl>
0	28.21820
1	27.40272

Figure 1. Grouped box plots and means for throw velocity, pop time, and leadoff distance by play result. All success instances of 0 result in a safe runner while 1 results in an out

These graphs are very important as they confirm our baseball intuition with statistical validation. There is a greater median and mean throw speed for plays ending with an out. There are also smaller median and mean pop times for outs. Finally, there is a larger median and mean leadoff distance for runners who are safe rather than out. This validation allows us to move forward with our analysis and use these statistics to better understand how catchers and baserunners can improve their stolen base tactics.

Offensive Side

To delve more into whether a runner had more of an impact on the outcome of a stolen base attempt compared to a pitcher, we wanted to see if there was any relationship between a first base runner's leadoff distance and stolen base attempt outcomes. To determine if there is a pattern for leadoff distance (the independent variable) versus stolen base attempt outcome (the dependent variable), we grouped the data points by successful steals and created a scatterplot shown in Figure 2.

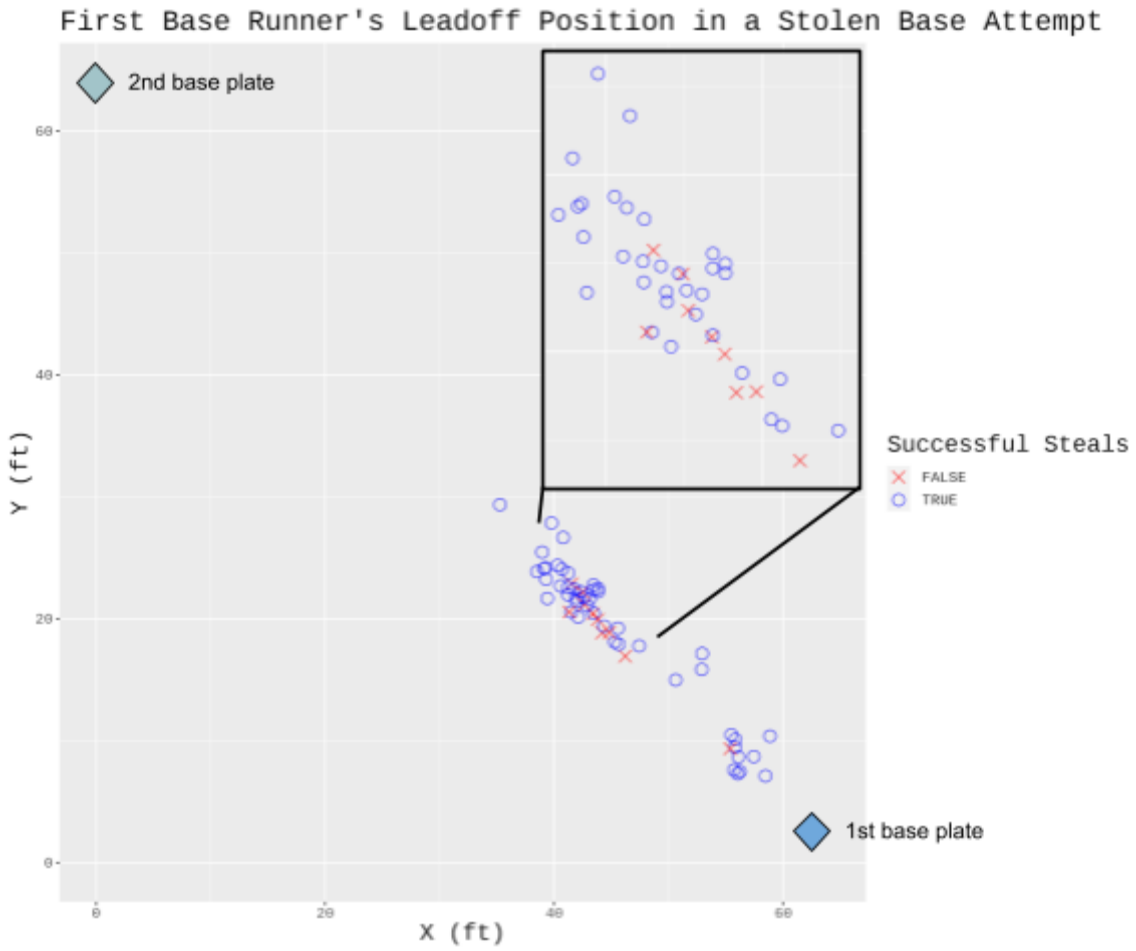


Figure 2. Visualization of first base runner's leadoff positions using a scatter plot with a zoomed in section of data

From our analysis, the runner's leadoff position's impact on the outcome of a stolen base attempt was inconclusive. Although there is a main cluster of data points labeled as successful steals around position (43, 22) in Figure 2, we cannot conclude that it is the "sweet spot" for achieving successful steals since there is also a cluster of data points labeled as unsuccessful steals in the same relative position. Since there is no isolated cluster of successful attempts, this means that we can focus more on the catcher's impact on the outcome of a stolen base attempt.

Defensive Side

In order to explore the catcher's impact on the outcome of a stolen base attempt, we decided to evaluate how pop time and throw velocity impact each other and if that affects the outcome of the play. We were able to generate a scatter plot to find where the "sweet spot" of catcher pop

time and throw velocity lies. Success is viewed in the eyes of the catcher, where false is a stolen base and true is a throwout. The results are shown below in Figure 3.

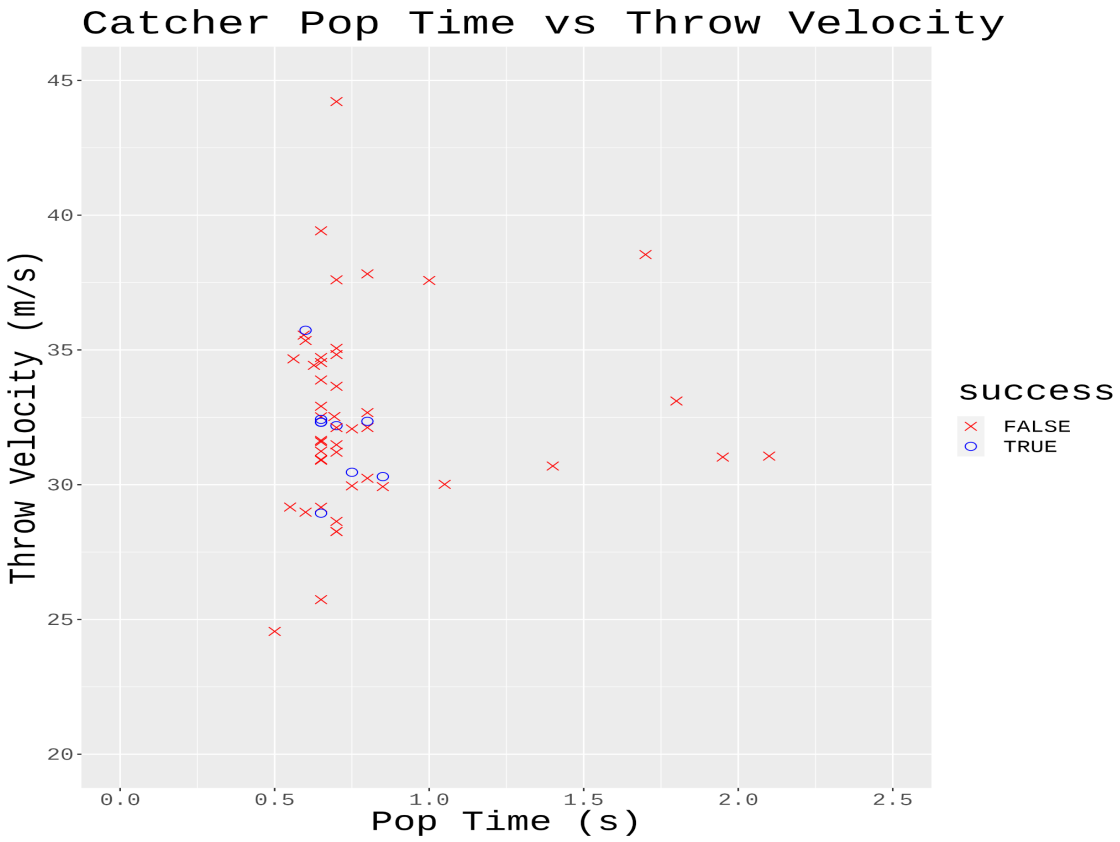


Figure 3. Catcher Pop Time vs Throw Velocity

We can see here that all of the throw outs occur in one specific region around .75 pop time and 30-35 m/s throw velocity. This is important to note that high throw speeds and quick pop times alone are yielding unsuccessful results.

4. Conclusion

From this project we were able to really see how the game and stat tracking has evolved over the past 120 years. Baseball analytics fanatic Bill James sees the hobby as an “ever-expanding line of numerical analysis” [5], which holds a lot of truth as new statistical measures are developed all the time. In addition, the preciseness of those stats are more specific as radar and timing technology rapidly develop. While the interaction between a base stealer and a catcher seems so simple, the complexity has grown significantly over the years. Using the data from the early 1900s, we determined that the efficient management of pop time and throw speed was the most important factor in a catcher’s success. However, baseball evolves over time as these trends are noticed and corrected.

5. Future Projects

There is a lot more that can be done in this sector of baseball analytics since this is just an initial review of historical data. As stolen base attempts become more prominent[1], it is important for teams and players to analyze more ways to increase their successful stolen base attempts on offense and increase their rate of catching steals on defense. Some ways this research can be further developed and used include:

- Positioning of the fielder relative to second base
- How pitch location affects the catcher’s pop time/throw speed matrix
- How pitch speed affects the catcher’s pop time/throw speed matrix

6. Appendix

1. Dayn Perry Apr 26. (2023, April 26). *MLB rule changes: "Year of the stolen base" is already threatening to rewrite the record books*. CBSSports.com. https://www.cbssports.com/mlb/news/mlb-rule-changes-year-of-the-stolen-base-is-already-threatening-to-rewrite-the-record-books/
2. If we included attempts to steal third base, then other factors such as whether the hitter is right-handed or left-handed would come into play.
3. In the format (player_position, event_code) based on the provided SMT Data Challenge glossary.
4. Euclidean distance = $\sqrt{(y_2 - y_1)^2 + (x_2 - x_1)^2}$
5. Bechtold, T. (2021, April 8). *State of Analytics: How the movement has forever changed baseball – for better or worse*. Stats Perform. https://www.statsperform.com/resource/state-of-analytics-how-the-movement-has-forever-changed-baseball-for-better-or-worse/