The Role of Time and Distance in First Base Pickoffs

https://github.com/addisonkline/smt-data-challenge

Abstract

Contrary to popular belief, it appears that the success rate of first base pickoffs is not affected by the distance of the runner from first base. It is likely that a more prominent factor in how often these pickoff attempts are converted into outs is the time it takes for the pickoff to occur. However, this unexpected conclusion makes more sense when considered further. The reality is that drawing comparisons between the distances of two runners from first base usually amounts to mere inches of difference. The mean distance of a runner from first base on a successful pickoff attempt is 10.77 feet, and the mean distance on an unsuccessful pickoff attempt to first base is 11.59 feet. These measures were calculated by using the 97 full professional games worth of data provided by SMT. And as one can see, the difference between these two means is small at just over nine inches; that's about one step. It is reasonable to say that first base runners generally take similar leads, so a runner's lead usually doesn't separate him much in terms of how often he'll get picked off. As such, it would logically follow that another factor must matter more such as the aforementioned time the pickoff takes or perhaps the speed of the runner to get back. With the given data, there was no simple method for analyzing a runner's speed, but analyzing the timing of each pickoff play was possible. And through our analysis, it was indeed shown to be a factor in how often first base pickoffs succeed. The mean time for a successful first base pickoff is 1.7 seconds, while the mean time for unsuccessful attempts is 2.5 seconds. Already, these surface level findings concerning time and runner distance contradict the common viewpoint that the larger a runner's lead is, the more likely he is to get picked off. It is an interesting phenomenon to discover: if these findings were applied to baseball, they could change how aggressive runners are in taking their leads and possibly result in more stolen bases.

1. Introduction

It's a cold October night in St. Louis during game 4 of the 2013 World Series. It's the bottom of the ninth, two outs, a runner on first, and the Cardinals trailing 4-2 to the Boston Red Sox. Their dominant closer Koji Uehara is facing legendary Cardinals slugger Carlos Beltran. The stage is set for a dramatic ending if Beltran helps lead the Cardinals to an improbable come from behind victory. Uehara delivers a ball and a strike; it's an even count. The crowd is on their feet, eagerly awaiting the outcome. But then, the unexpected happens: Uehara steps off the mound and fires the ball to first baseman Mike Napoli, who swiftly tags out runner Kolten Wong. Uehara and the Red Sox erupt into cheers, while Wong slams his helmet on the ground in disgust. Cardinals fans are left speechless as to what they have just witnessed, and with this shocking end to the game, Boston ties the series, which they eventually go on to win.

Looking back, this play ended up being a real turning point in the World Series, and it begs the question: would Wong have been safe with a shorter lead? Well, to effectively answer the question, one has to consider many factors associated with what makes pickoffs successful. As such, by conducting a thorough analysis of the data provided, this paper aims to provide evidence for the assertion that the runner's lead rarely matters in how successful a first base pickoff is, and how it is the duration of the pickoff that has a greater impact.

2. Data

We decided to focus our analysis on what we call valid pickoff throws to first base, which are pickoff throws to first base that satisfy two conditions: 1) the first baseman successfully catches the pitcher's pickoff throw, and 2) no more throws are made after the first baseman catches the pickoff throw. By limiting our analysis to this group of pickoffs, we eliminated two possible outcomes for a pickoff throw: namely, unsuccessful catches by the first baseman, and pickoff throws that resulted in rundowns, both of which would potentially skew the data and make later analysis significantly more difficult.

Four categories of data were provided to us, namely those in the folders “game_events,” “game_info,” “player_pos,” and “ball_pos,” with the first three being particularly useful for our project. The “game_info” data included play-by-play information on whether or not bases were occupied, helping us determine whether the baserunner was safe or out following a pickoff throw. The “game_events” data included information on the events occurring on the field, with event code 6 referring to a pickoff throw. The “player_pos” data included the exact position of both the baserunner and first baseman at intervals of 10 milliseconds, which we used to calculate distances to first base for both players.

One limitation of the data was the lack of information on pitcher handedness, a factor which plays an important role in making a pickoff throw to first base. We would have also benefited from more information on the runner beyond just their position on the field, such as orientation and extension, as those also play an important role in how fast the runner can return to first base. The provided files also lacked data on basic game information such as count and game score, both of which play a role in determining when a pickoff is more likely to take place. Though information on the specific players involved in any given pickoff was available, player-specific data is not (at least not without extensive computation), meaning potentially useful metrics like average player sprint speed were effectively unobtainable.

3. Methodology

The first overall step of our analysis consisted of compiling, transforming, and filtering the relevant data into one single CSV file to be analyzed separately, the code for which can be seen in “scraper.py” *(see appendix 1)*. The first step of scraping was reading every single CSV file provided and compiling them into four single data frames: “ball_pos,” “game_events,” “game_info,” and “player_pos,” which would drastically simplify the process of data filtering later on. Our next step was to make each individual play completely unique in the overall dataset, which would drastically simplify future steps in our analysis *(see appendix 2)*. We therefore created a “game_play_id” field in each data frame that is concatenated the “game_str” and “play_id” fields (e.g. play number 17 in game "1901_05_TeamLI_TeamA3" would have a “game_play_id” value of "1901_05_TeamLI_TeamA3_17").

Once all “game_play_id” values were created, we filtered out all the valid pickoff plays to first base, as defined in the “Data” section above *(see appendix 3)*. Once all valid pickoffs were found, our next step was to filter them down to plays that would be usable in our analysis (i.e. plays for which no necessary data is missing). To do this, we first filtered out all pickoff plays by “game_play_id” that, for whatever reason, did not exist in the “game_info” data frame. We then filtered out all pickoff plays that had missing first baseman data in the “player_pos” data frame. The result of this filtering was the list of pickoff plays used in our analysis.

Our next step was finding the following values for each pickoff: whether or not it was successful, the time it took to complete the pickoff, the runner’s maximum distance from first base during the play, the runner’s X and Y coordinates at their farthest from first base, the first baseman’s maximum distance from first during the play, and the first baseman’s X and Y coordinates at their farthest. To determine the success of a pickoff, we checked the “game_info” data frame: if the play immediately following the pickoff had no runner on first base, then the pickoff was considered successful *(see appendix 4)*. The pickoff time was calculated simply by subtracting the pickoff's first timestamp from the pickoff's final timestamp. In order to calculate distances from the bag, we used trigonometry to determine the X and Y coordinates of first base to be (63.6396, 63.6396) *(see appendix 5)*.

After determining these coordinates, we calculated the straight-line distance between the runner and first base at each timestamp for every pickoff *(see appendix 6)*. We then isolated the maximum distance to record that value and the X and Y coordinates of the runner at that distance. We then repeated this process for the first baseman to get the maximum distance of the first baseman to the first base bag as well as their coordinates at that position.

Once this data was calculated for every pickoff in the filtered list, we compiled the list of all filtered pickoffs and associated data into a single file called "pickoff_data.csv." Data analysis was then conducted in separate scripts, such as "Data Analysis.Rmd" and "analysis-2.py", which is detailed in the "Discussion" section below.

4. Discussion

We first analyzed the size of a runner's lead from first base. To simplify matters, we created two separate data frames: one containing the maximum runner distances for successful pickoffs, and one containing these distances for unsuccessful pickoffs. Additionally, to get a better idea of what is considered a large or small runner distance, we found the 0th, 25th, 50th, 75th, and 100th percentiles for runner distances on all pickoff plays. Using these percentiles, we discovered whether the amount of successful and unsuccessful pickoffs vary by quartile. The bar graph below visually displays our results.

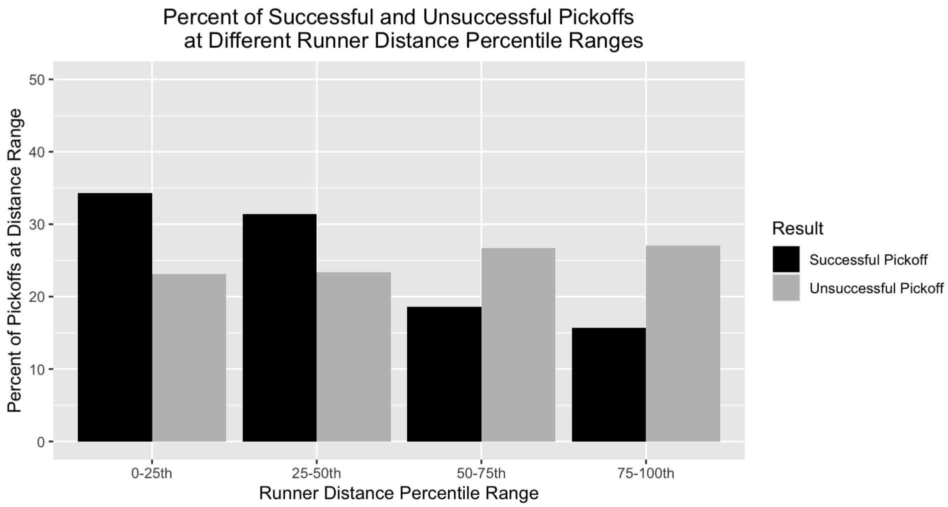


Fig. 2: The proportion of successful and unsuccessful pickoff throws for the four runner distance quartiles. Interestingly, pickoff attempts are more likely to be successful the closer the runner is to the first base bag.

According to the graph, most successful pickoffs actually occurred when the runner had a below- average lead (0-50th percentile). And in fact, 64.3% of successful pickoffs happened when runners took below-average leads. This is the opposite of what one would expect given the common assumption that bigger leads result in more pickoffs. Moreover, 53.3% of unsuccessful pickoffs occurred when runners took larger leads (50-100th percentile), also directly contradicting this assumption. Therefore, this graph gives us proof that leads may not matter as much in how frequently a pickoff succeeds. That said, the only way to truly corroborate this claim is to find the pickoff success rates at these quartiles. This is precisely what we did next as the graph below shows.

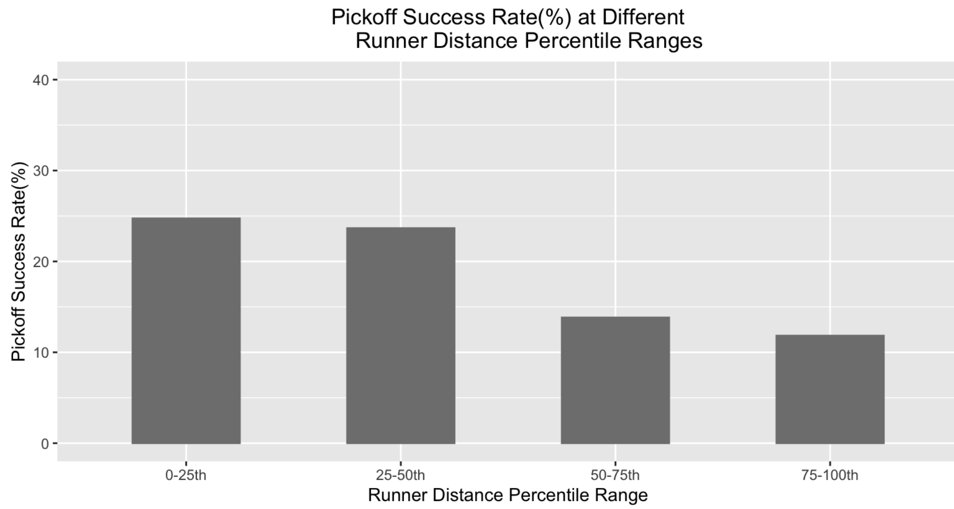


Fig. 3: Pickoff success rate by runner distance quartile. Counterintuitively, runners farther from the bag have a lower chance of being picked off at first base.

Looking at this graph, we can find it to be another piece of supporting information. Pickoffs are surprisingly more successful at the lower quartiles. At the bottom quartiles, pickoffs had success rates of 25.8% and 22.6%, respectively. In contrast, pickoffs only had a 14% success rate at the third quartile and an 11.8% success rate at the fourth quartile. This is clear evidence of pickoff success not being correlated with increased lead length. However, to make sure our results were not due to luck, we ran a statistical test using a binomial distribution (*see appendix 7*). The results from our test determined that there was less than a 15% chance of observing the specific number of successful pickoffs for each quartile. As such, it is improbable these results are due to chance. Thus, Cardinals fans can breathe a sigh of relief: it is unlikely that Kolten Wong was thrown out because of his lead.

Still, if it wasn't Wong's lead that got him thrown out, what did? The answer we came up with is that it was Uehara's pickoff speed. Some pitchers are simply faster with their pickoffs than others. However, we wanted to test how plausible this theory was, so we started by creating two data frames containing the time in seconds each successful and unsuccessful pickoff took.

Then, we made two histograms that show whether the frequency of successful and unsuccessful pickoffs vary by duration.

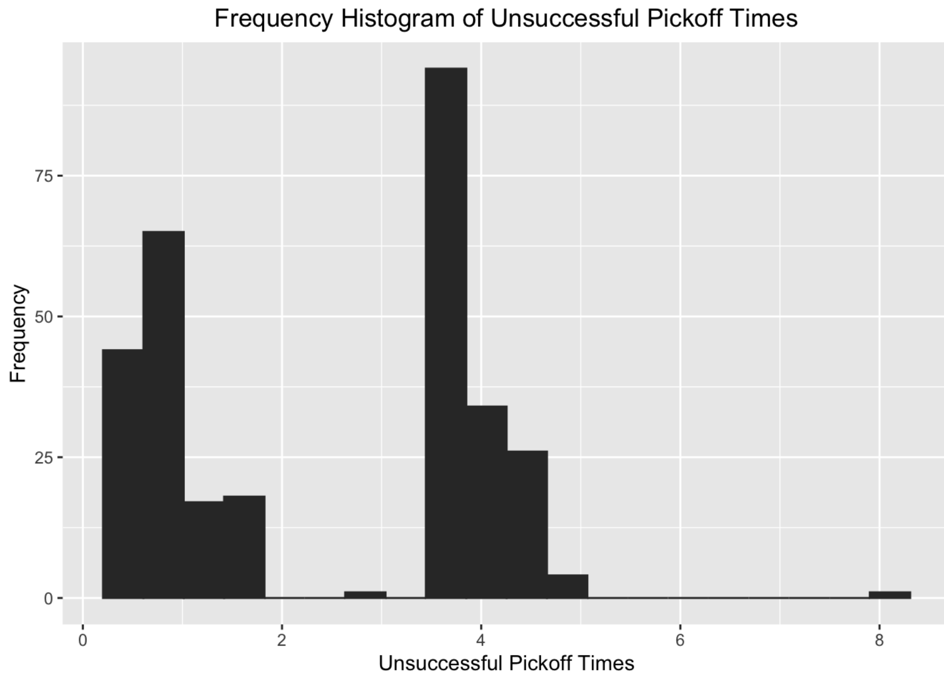


Fig. 4: All unsuccessful pickoffs we analyzed plotted by the time they took from start to finish. There are two pronounced clusters: one around 0.5-1 seconds and another around 3.5-4.5 seconds.

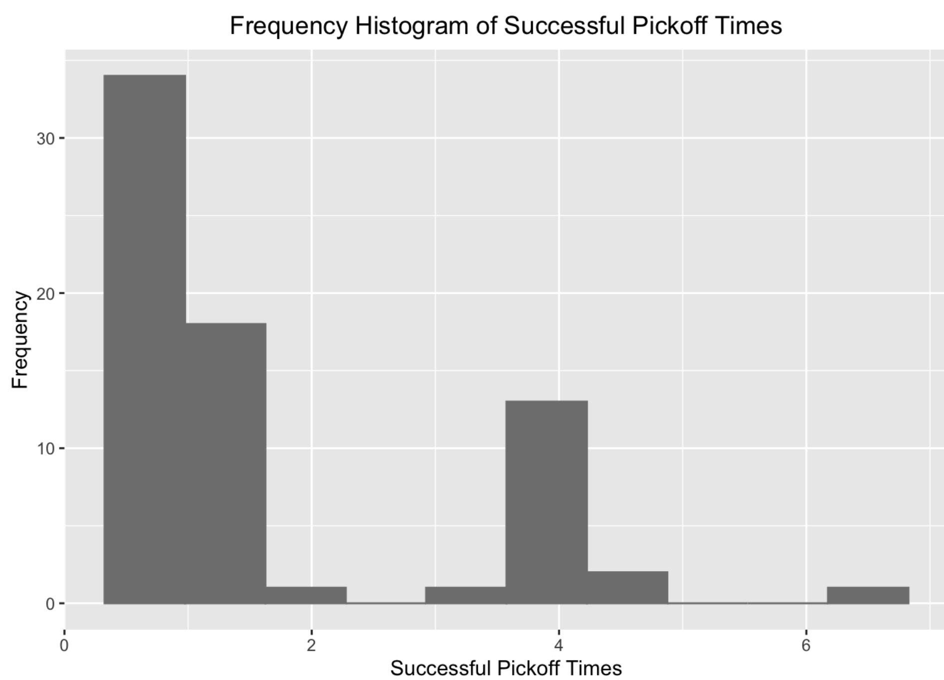


Fig. 5: All successful pickoffs we analyzed plotted by the time they took from start to end in seconds. The data mostly clusters around the 0-2 second range, with another smaller cluster in the 3-5 second range.

As we can see, there is a significant difference in pickoff frequency when the factor is time. 74.3% of successful pickoffs take under two seconds compared to only 47.4% of unsuccessful pickoffs. These proportions are not even close; there is a definite difference in the typical duration of a successful versus an unsuccessful pickoff. This conclusion is also backed by the aforementioned mean durations of successful and unsuccessful pickoffs at 1.7 seconds and 2.5 seconds, respectively. As such, it is evident that time is the more prominent factor in pickoff success. Consequently, the moral of the story is this: runners generally shouldn't worry about how long their lead is, but rather how quick the pitcher's pickoff move is.

Our analysis yielded unexpected results, but these results could fundamentally change baserunning in baseball. Many people believe the length of a runner's lead determines how likely he is to get picked off. As we have seen though, this does not seem to be the case. We carefully considered the effects of lead length on the success rates of pickoffs by using distance quartiles, determined our results were not because of chance, and demonstrated the more significant impact time has on pickoffs. Armed with this knowledge, teams with fast runners can encourage them to take larger leads. This will likely result in more stolen base opportunities and help teams get more runners into scoring position, a critical aspect of winning baseball games.

5. Improvements

There are also some potentially crucial factors in the success of any pickoff throw that we did not include in this project but could evaluate in a follow up analysis. Two examples are pitcher position and ball position, both of which are obviously important components of a pickoff. The latter is particularly useful, because when coupled with pickoff time, the pickoff throw's speed can be calculated.

If we could follow this project up with additional resources, there are a number of possible improvements we can make. For example, player-specific statistics on baserunning data like average sprint speed could potentially provide insight into why odds of being picked off counterintuitively decrease with distance from the bag, as it is possible (if not likely) that faster runners take bigger leads off first on average. Another potential improvement that could be made is the addition of more splits, such as pickoff odds by count or game score.

6. Appendix

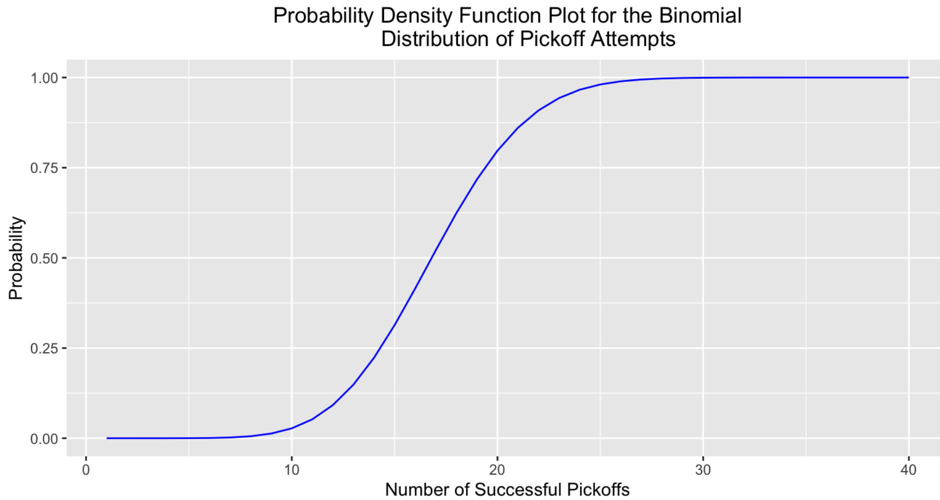
1) As per the project submission instructions, the original SMT data folder is no longer in the GitHub repository. This means that if run as is, “scraper.py” would not work since it references a directory that does not exist. However, scripts used for the analysis portion of this project use the “pickoff_data.csv” file, so they can still be run without any changes.
2) After the four data frames were filled, it became clear that the “play_id” field provided in the data (or the “play_per_game” field in the case of “game_info”) was fundamentally limited in the sense that in the sense it was exclusive to each game (e.g. the first play of every game would always have a “play_id” of 1). This would have made finding a given play in another data frame cumbersome, as a minimum of two identifying variables (“play_id” and “game_str”) would need to be checked to verify that two plays referenced in separate data frames are in fact the same play.
3) To find all plays that fit our definition, we checked the “game_events” data frame for instances of three consecutive rows with “event_code” and “player_position” values matching the following:

player_position	event_code
1	6
3	2
0	5

In the figure above, the first row represents the pitcher initiating a pickoff throw. The second row represents the first baseman successfully catching the throw, and the third row represents the end of the play.

4) In a vacuum, a pickoff play followed by no runner on first base could mean that the runner advanced. However, a pickoff where this happens would almost never end upon the initial throw to the first baseman–it is much more likely the batter enters a rundown or that the first baseman misses the throw entirely. We therefore operated under the assumption that for the set of valid pickoff throws to first base used in our analysis, no batter on first base following the pickoff meant the runner was out.
5) This follows from the fact that first base is exactly 90 feet from home plate and at a 45-degree angle when looking directly from home to the pitcher's mound. Since home plate is located at (0, 0) in the coordinate system provided, we can use trigonometry to find the X coordinate of first base to be 90*sin(45deg) = 90/sqrt(2) ~= 63.6396. The 45-degree angle makes finding the Y coordinate easy: Y = 90*cos(45deg) = 90*sin(45deg) ~= 63.6396.
6) For two objects at respective coordinates (a, b) and (c, d), the straight-line distance (more formally known as the Euclidean distance) between them is equal to the following: sqrt((c - a)^2 + (d - b)^2).
7) Pickoffs lend themselves nicely to a binomial distribution, a probability distribution that describes the number of successes in a fixed number of trials with only two outcomes: success or failure. For our purposes, it tells us how likely it was to get the exact amount of successful pickoffs or less for each quartile range, given the mean pickoff success rate of 18.72%. To run this test, we split the original dataset of 374 pickoffs into four groups of 93 (two pickoffs were left out since 374 is not divisible by 4). Each group represented a quartile range. We ran the test for all groups and created a line graph that visualizes the binomial distribution's probability density function, which is the equation used to

evaluate the probability of a certain number of successes or less occurring based on the usual rate of success, regardless of any external factors.



According to this plot, the odds of observing the specific number of successful pickoffs for each quartile was below 15%. For the top and bottom ranges (0-25th and 75-100th), it was even below 10%. Therefore, we can reject the claim that our results were due to luck at the 0.1 significance level for these ranges. We cannot do so for the middle ranges, but we can still be relatively confident seeing as those numbers of pickoffs being successful had a smaller than 15% chance of occurring.

7. References/Acknowledgements

1) The game-ending pickoff in Game 2 of the 2013 World Series can be found on YouTube here: https://www.youtube.com/watch?v=eQnHJwIabBU. Per the project submission instructions, this video file has been uploaded to the GitHub repository under the name "Uehara picks off Wong for final out.mp4."