MLB Attendance, Home Field Advantage, and Batting Performance

Ethan Rubenstein

Introduction

In baseball, the home field advantage is a widely known notion that a team benefits from playing at their own stadium. The advantage of playing at the home field is often attributed to the idea that players are local to the area, play there for half of the season, and are therefore more comfortable with their own housing and familiar routines than staying in hotels located in less familiar places. Additionally, the home crowd is another source that is often cited in defining the home field advantage, providing confidence boosts for players and inducing nerves in the visiting team. While the home field advantage factors of familiarity and proximity to home do not generally change from game to game, crowd size can vary significantly. More crowded games may create more pressure for the visiting team and boost confidence for the home team, which can potentially manifest in player performance. The effects of the home field advantage in crowded games may materialize in particularly stressful situations, such as batting with the bases loaded and two outs, a high leverage situation in which the result of the at-bat can have a marked impact on the outcome of the game.

In 2015, Jones showed that there is only a minor home field advantage in terms of win rate for MLB games . This paper will seek to establish whether stadium attendance consistently contributes to the home-field advantage on the player-level, rather than the game level. Run expectancy based on 24 base-out states (RE24), the

average number of runs that a team would expect to score in the current inning given a particular game state (runners on base, outs) will be used to determine player performance. RE24 was chosen for two primary reasons: first, it is contextual, depending on the current game state. As a result, one would expect that a home-team batter would add more to the RE24 on average after an at-bat if there is a player-level home field advantage, indicating that home team at-bats result in a more optimal outcome more frequently than away team at-bats. Secondly, RE24 accounts for discipline at the plate and pitching performance. If a home field advantage does exist for players, it may affect pitchers as well - an away team pitcher would be expected to throw worse pitches, resulting in more hits, hit by pitch (HBP), and walks. Inversely, a home team pitcher would be expected to have a boost in performance from the home team advantage, decreasing the number of hits, HBP, and unintentional walks.

The research hypotheses for this paper are as follows:

Hypothesis 1: There is a significant difference in mean RE24 between home and away teams across different stadium attendance levels.

Hypothesis 2: In high leverage situations, home teams have a significantly higher average RE24 than away teams in high-attendance games.

If the relationship between stadium attendance and RE24 is significant between the home team and the away team, stadium attendance may become an important predictor in determining the end result of an inning in play. Additionally, stadium

attendance would be of interest to oddsmakers who offer live wagers on MLB at-bats. Lastly, MLB team management may gain an additional incentive to draw higher attendance at their home games.

Data Collection

Data was collected from three sources. First, Retrosheet's MLB 2022 game log dataset provided attendance for each game in the 2022 regular season. Stadium attendance, however, is limited by the capacity of the stadium itself; the smallest ballpark by capacity, Tropicana Field, seats just 25,000 attendees ("Rays eliminate upper-deck seating, reduce capacity to 25,000", 2019), while Dodger stadium seats more than double that amount at 56,000. In order to normalize attendance figures across different stadiums, the attendance proportion for each game was calculated as the recorded attendance divided by the stadium capacity. Calculation of the attendance proportion necessitated the use of a second dataset containing stadium information, sourced from StadiumDude, a website containing information about NHL, MLB, and NFL stadiums. Stadium capacity was mapped to each game in the Retrosheet dataset and was subsequently used to calculate the attendance proportion variable located in that same dataset. Out of the 2,430 games in the 2022 regular season, 2,403 had readily available attendance figures. The list of excluded games includes two games played at non-MLB ballparks (Reds - Cubs at Field of Dreams (Footer et al., 2022) and Red Sox - Orioles at Bowman Field (Silver, 2022)) because they do not present a clear home and away team for analysis in this paper.

An ordinal variable named *attendance group* was additionally created in the Retrosheet dataset to eventually compare batting performance across different levels of attendance. This variable has four levels — low, med-low, med-high, and high — which represent observations below the first quartile, between the first and second quartiles, between the second and third quartiles, and above the third quartile in attendance proportion respectively.

Doubleheaders, or games between two teams scheduled for the same day, presented a challenge in this project. The Retrosheet dataset contained a variable *number of game* denoting what kind of game it was (normal single game, first game in a double header, etc), but there were several instances of doubleheaders that were missing attendance data for game one, but not game two. Retrosheet recorded attendance for 69 game twos (accounting for all doubleheaders played in 2022 ("MLB Doubleheaders - 2022", 2022)) of doubleheaders as compared to just 45 game ones. As a result, I opted to remove all doubleheaders from the dataset. Empirically, this decision is supported by the fact that for doubleheaders that have both game one and game two attendance available, attendance can vary significantly. For example, the Rockies - Tigers doubleheader on April 23, 2022 recorded an attendance proportion of 0.91 for game one and 0.70 for game two, despite the Tigers winning the first game 13-0 at home. This decision was also justified by the fact that the at-bat dataset, described in the following paragraph, does not discriminate between game one and game two of doubleheaders, so attendance data cannot be accurately mapped with a one-to-one relationship (i.e. an at-bat from game one might be mapped with the attendance data from game two).

After the addition of attendance proportion to the Retrosheet dataset, R code modified from Lecture 3 was used to scrape 2022 regular season pitch-by-pitch data from the *baseballr* package, and then calculate the run expectancy based on the 24 base-out states (RE24) for each plate appearance. A new variable based on teams that played and on the date that they played, *gameid*, was added to both the at-bat and Retrosheet datasets to act as a key for ease of merging the attendance data into the at-bat dataset (e.g. San Diego playing in Philadelphia on May 5 would have a *gameid* of 2022-05-19 PHI SD). This first required some of the *home* and *away* team abbreviations in the Retrosheet dataset to change and match those in the at-bat dataset.

Finally, all 9th inning at-bats were removed from the dataset. During the final inning (barring extra innings), teams begin incorporating new strategies that maximize win probability rather than runs in the final inning of the game. As a result, the 9th inning is excluded from the RE24 matrix, so 9th inning run expectancy should not be calculated using the RE24 averages for innings 1-8. After merging the attendance data into the at-bat dataset and removing doubleheaders, games with missing attendance data, and games played at non-MLB ballparks, and 9th inning at-bats, the final number of 2022 regular season at-bats included in this analysis stands at 155,890.

Descriptive Statistics

RE24 and Home/Away Team

Fig. 3.1 shows a box plot of RE24 per inning (the away team always bats in the top of the inning, while the home team always bats in the bottom). From the plot, the median and spread of RE24 look identical between the home and away team. Both means are around zero, which is the league average; the home team had an average RE24 of -0.0012, while the away team had an average of -0.0083. The histogram (Fig 3.2) shows that RE24 is somewhat normal, with the most frequent plate appearances being slightly below zero (i.e., an out without high stakes) for both the home and away team.

Figure 3.1

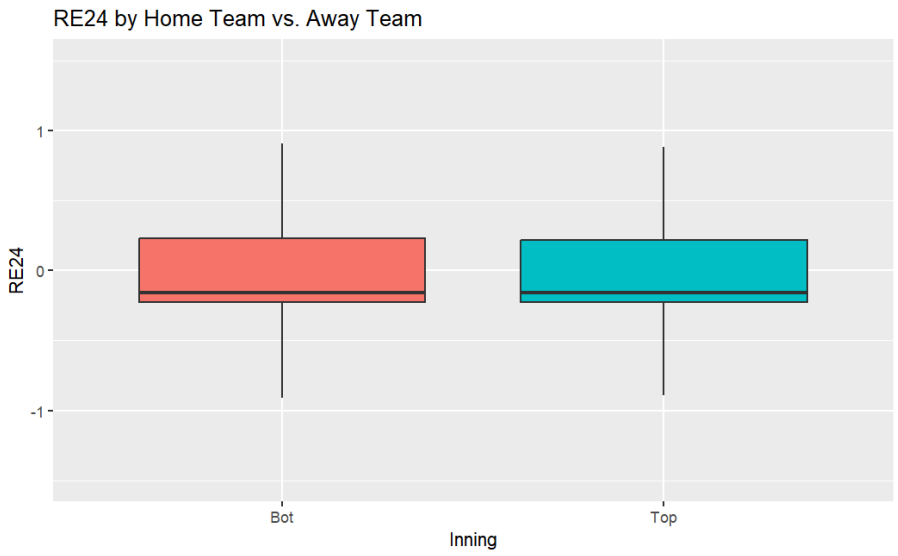
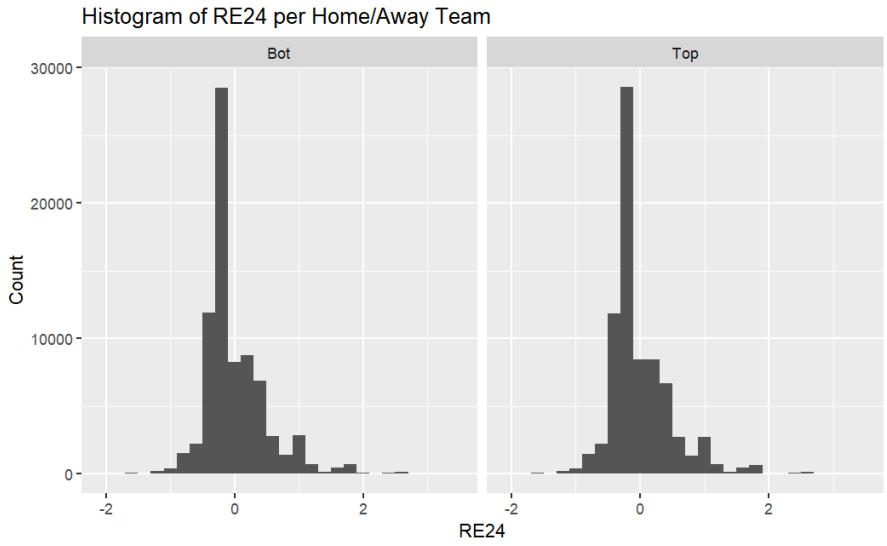


Figure 3.2



RE24 and Home/Away Team

Fig. 3.3 shows a box plot of RE24 per attendance level. Similarly to Fig 3.1, the median and distribution of RE24 looks identical between all levels of attendance. The mean for *high*, *low*, *med-high*, and *med-low* were -0.0090, -0.0060, -0.0022, and -0.0021 respectively. Fig 3.4 also shows that the distributions of RE24 per attendance level are nearly the same, and somewhat normal, centered around zero.

Figure 3.3

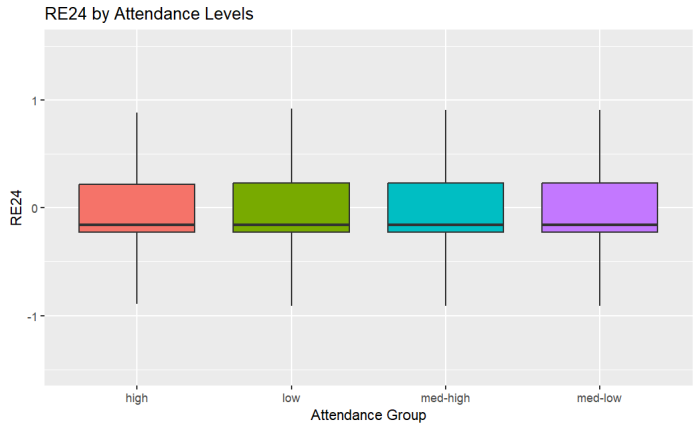
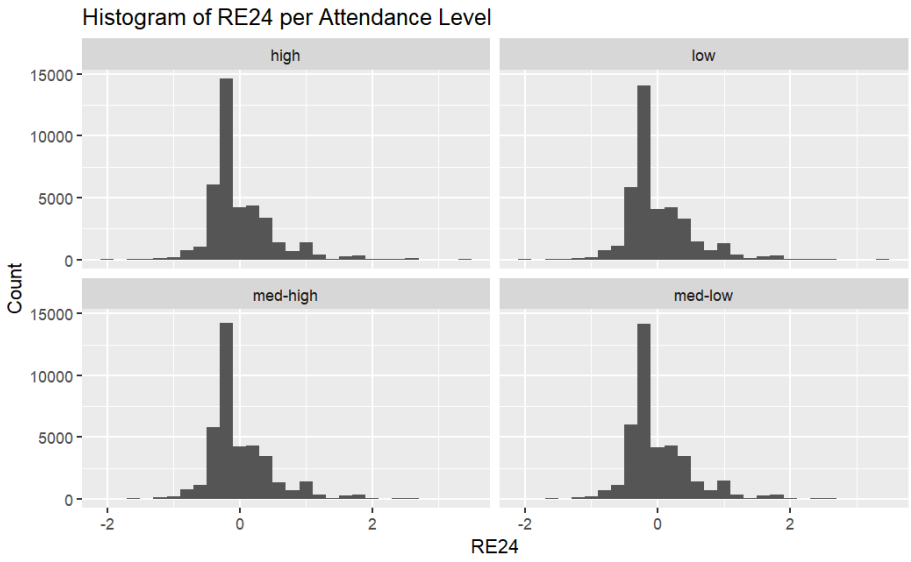


Figure 3.4



RE24 and Home/Away Team per Attendance Level

Fig 3.5 shows a box plot of RE24 between attendance groups and home and away teams. Like the previous two figures, the spread and median are nearly identical between all groups, except for the away team in highly-attended games, which has a lesser upper quartile and shorter tails than the other seven groups that indicates more “average” plate appearances than normal (RE24 is less extreme). The mean per group is shown in Fig 3.7. Teams playing at home with a medium-high stadium attendance had the highest mean RE24 (0.0045), followed by teams playing at home with the highest attendance (0.0030), meaning that these teams were expected to add *slightly* more than the average number of runs per plate appearance. Fig.3.6 shows that the distribution of RE24 between all levels of attendance and home/away teams is nearly identical and centered around zero.

Figure 3.5

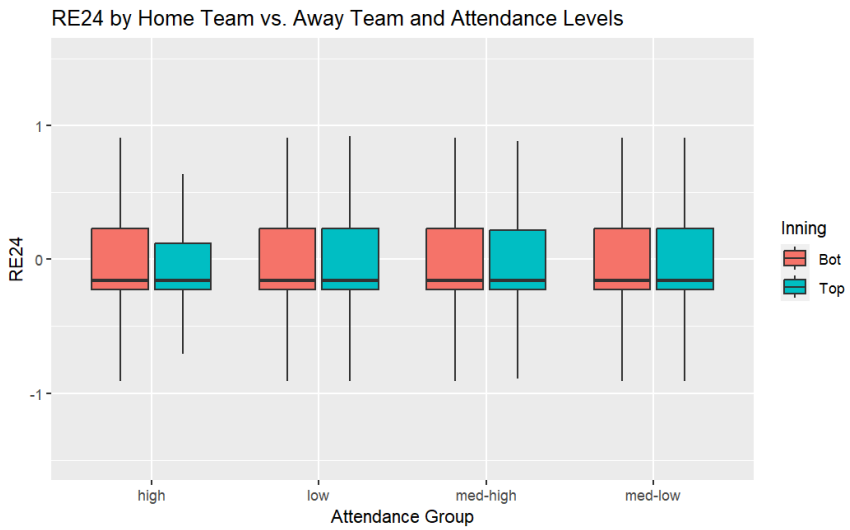


Figure 3.6

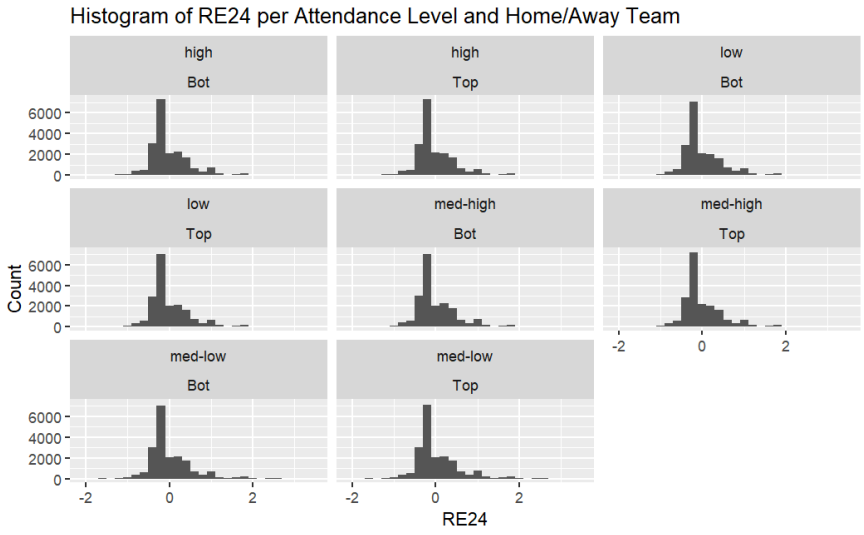


Figure 3.7

Attendance Group	Home/Away Team	Mean RE24
high	home	0.0033104984
low	home	-0.0064700464
med-high	home	0.0061201899
med-low	home	-0.0019316655
high	away	-0.0204951982
low	away	-0.0035951460
med-high	away	-0.0091444288
med-low	away	-0.0009829135

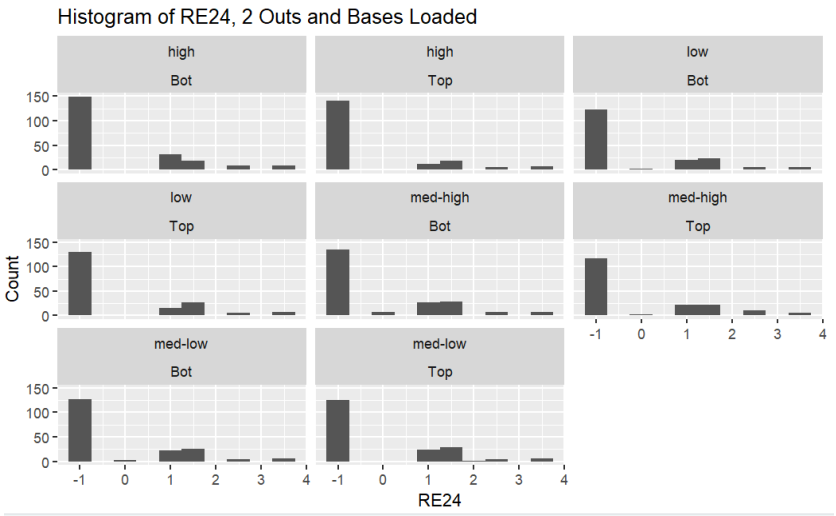
RE24 per Home/Away Team and Attendance Level in High Leverage Situations

RE24 was measured across home/away teams and attendance groups for two high leverage situations: bases loaded with 2 outs, and 8th inning at-bats where the score differential is one or zero (tied game).

Fig 3.8 shows the histogram for the 2 outs, bases loaded scenario. Note that the sample size is only 1,536, which is broken down even further into the eight attendance level and home/away team groups. However, the distributions appear to be similar across each attendance level and home/away team. The distributions make sense — RE24 is at its

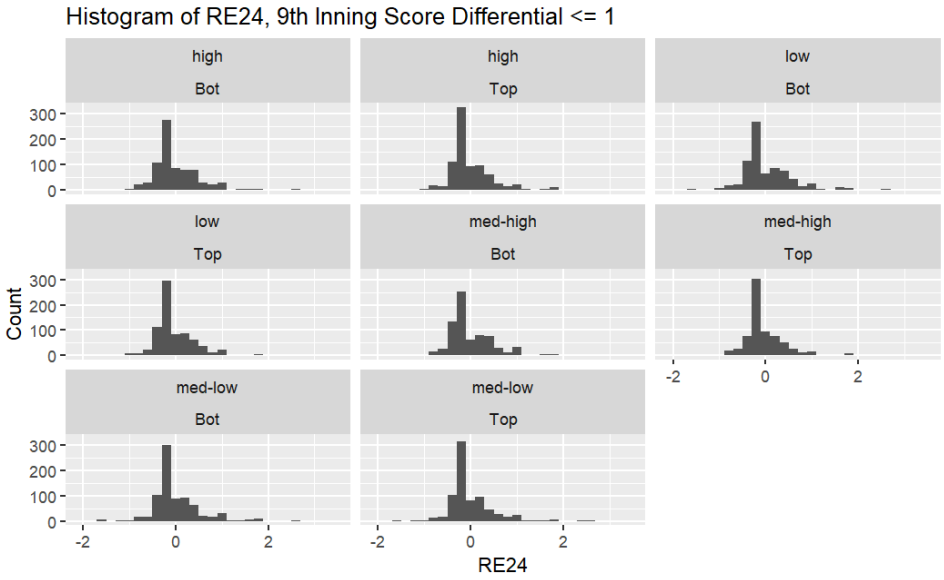
maximum when a grand slam is hit, hence RE24 between three and four, though most plate appearances will result in an out and bring the RE24 back down to zero (necessitating that the RE24 be negative for that plate appearance).

Figure 3.8



Lastly, Fig 3.9 shows a histogram of RE24 across home/away teams and attendance levels in the 8th inning when the score is tied or a team is trailing by one run. The distribution appears to be slightly right-skewed, with most plate appearances ending in a negative RE24 (out), but with a wide range of positive RE24 plate appearances. The larger right tail can be explained by multi-run home runs, which add more to RE24 than the worst-case scenario can typically subtract. There does not appear to be a difference between the distributions across either variable. The sample size for this scenario is 6,145.

Figure 3.9



Inferential Statistics

Hypothesis 1, a significant difference exists in the means of RE24 between home and away teams across different stadium attendance levels, was tested using a two-sample Wilcoxon test across each level of attendance. Wilcoxon tests were used after Shapiro tests and Anderson-Darling tests for the normality of the overall sample and individual groups were found to be significant. The result of the test Wilcoxon was only significant for games classified as high (adj. $p < 0.0035$, V = 197,438,950) and was not significant for medium-high (adj. $p = 0.1052$, V = 192,146,673), low (adj. $p = 0.8970$, V = 183,839,017) and medium-low (adj. $p = 0.8970$, V = 191,514,305) games. The test shows that there is not a significant difference in mean RE24 between the home team

and the away team in games that had low, medium-low, or medium-high attendance, but there was a significant difference in highly attended games.

After the two-sample Wilcoxon test, one-sample Wilcoxon tests were run to determine (1) whether the median RE24 for home teams in highly attended games was significantly greater than zero, and (2) whether the median RE24 for away teams in highly attended games was significantly less than zero. The results of the test for home team median RE24 greater than zero was not significant (p = 1, V = 86,744,835), while the results of the test for away team median RE24 less than zero was significant (p < $2.2 * 10^{-16}$, V = 77,100,149).

The second hypothesis, home teams in high-attendance games have a significantly higher RE24 than away teams in high leverage situations, was also tested with two-sample Wilcoxon tests. The results for the bases loaded, two outs scenario were significant (p = 0.042, V = 21,520.5), while the results for the 8th inning score differential less than 1 scenario were not significant (p = 0.316, V = 321,483).

The results of these tests indicate that the home field advantage does not uniformy manifest in improved RE24 for home teams. However, it does show that away teams in high-attendance games have a worse average RE24 than the league average of zero. The two high leverage situations outlined in this paper did not support a shared conclusion. RE24 for home teams in high-attendance games was significantly higher than the away team with the bases loaded and two outs, but this was not true for 8th inning at-bats with a difference in score of one or zero.

Discussion

The impact that stadium attendance has on player performance as exhibited in this paper is consistent with prior work on the relationship between attendance and MLB game outcomes. Bukenya et al. analyzed the run difference between home and away teams from the 2015-19 seasons and the 2020 season in which COVID-19 prevented fans from attending games, finding no relationship with attendance (Zimmer et al., 2021). However, Groetzinger et al. found that home teams were 4% more likely to win and could be expected to hit one more home run than away teams during the 1996-2005 seasons, indicating that a home field advantage does exist, even if it is not directly correlated with attendance (Smith et al., 2009).

The lack of a clear relationship between RE24 and stadium attendance for both home and away teams may be explained by a number of factors. First, better teams tend to perform well both on the road and at home, while worse teams perform worse than average. This means that better teams neutralize the opposing team's home field advantage and worse teams lose their home field advantage, regardless of stadium attendance. Additionally, some teams may just be more popular than others and consistently draw in more fans regardless of talent level, while other teams do not have a large, loyal fanbase to attend their games, even when performing well. Home field advantage may not depend on player performance at all — the minor benefit added for teams playing at home may be partially caused by more umpire bias, which even at a low level can change the game outcome given the relatively low sample size of pitches thrown to each team in an MLB game.

The mixed results of RE24 in high-leverage situations of high-attendance games is less surprising. The first scenario, bases loaded with two outs, is more consequential than batting in a close game in the 8th inning. In the first scenario, there is a high expected run value, given that the batter can score anywhere from zero to four runs in a single at-bat. As a result, fans are likely to be more animated, cheering for the home team and potentially giving the home batter a confidence boost that translates into better performance. In contrast, the second scenario may have a low expected run value, since it encompasses all game states in the 8th inning. Since each team will have another opportunity to bat, and the situation includes at-bats from teams ahead by one, there is less overall urgency for individual players to maximize the run expectancy as well as less action from fans. Tom Tango's leverage index and matrix could be used to further explore the relationship between attendance and RE24 for home and away teams, as several of the highest leverage situations were not covered in this paper (Tango, n.d.).

In addition to researching different high leverage situations, the relationship between RE24 and attendance could be refined by analyzing the 2020 season. Due to COVID-19 protocols, fans were not permitted to attend live games. Although the season was shortened to just 60 games per team, it offers a unique opportunity to examine the effects of stadium attendance on variables including RE24. In this paper, the 2020 season could be added as a fifth attendance group, "none", in order to discern whether there were differences across RE24 between the other groups, and if there were differences in RE24 across home and away teams. However, the 2020 season had other irregularities that may have impacted performance, including social distancing

protocols that may have influenced training, and perhaps even game outcomes, to some degree (“New rules, features, protocols for 2020 MLB season”, 2020).

Aside from incorporating data from the 2020 season to create a fifth attendance group and expanding the dataset with data from other seasons, further analysis of metrics other than RE24 may build upon the ideas presented in this paper. RE24 was chosen as the primary indicator of player performance due to its more holistic nature in implicitly incorporating pitchers (and defense to some extent) as well as its accounting of context, but the home field advantage may show up in other areas. As mentioned earlier, home teams were more likely to score an additional home run than away teams, which would not be obvious if only looking at average RE24 over multiple, or even single games (Smith et al., 2009). In order to solidify the conclusion that stadium attendance does not create a home field advantage by boosting player performance, other performance metrics should be tested. These may include simple count statistics, like home runs, traditional batting statistics such as batting average, or more complicated metrics such as win probability added.

Conclusion

Stadium attendance has been shown to have little to no bearing on player performance, as measured by RE24, between home teams and away teams. As a result, it is unlikely that the home field advantage reported by past work is solely caused by fans. However, further research should be performed to verify that other performance metrics do not display a relationship with stadium attendance that demonstrates a home field advantage. If the link between fan attendance and home field advantage is discovered,

it may present new incentives for cultivating and retaining fan bases and potentially offer new predictors in MLB game outcomes.

Sources

Footer, A., Bastian, J., & Sheldon, M. (2022, August 12). *The top moments from an enchanting Field of Dreams game*. MLB.com. https://www.mlb.com/news/cubs-reds-field-of-dreams-game-2022

Jones, M. (2015, December). *The home advantage in major league baseball*. Perceptual and Motor Skills. https://www.researchgate.net/publication/286650014_The_Home_Advantage_in_Major_League_Baseball

MLB Doubleheaders - 2022. ESPN. (2022). https://www.espn.com/mlb/stats/doubleheaders/_/year/2022

MLB. (2020, June 29). *New rules, features, protocols for 2020 MLB season*. MLB.com. https://www.mlb.com/news/mlb-announces-new-features-for-2020-season

Rays eliminate upper-deck seating, reduce capacity to 25,000. ESPN. (2019, January 4). https://www.espn.com/mlb/story/_/id/25683771/tampa-bay-rays-eliminate-upper-deck-seating-reduce-capacity-25000

Retrosheet Game Logs. (n.d.). http://www.retrosheet.org/gamelogs/index.html

Silver, Z. (2022, August 21). *Red Sox, O's meet in Little League Classic tonight*. MLB.com. https://www.mlb.com/news/2022-mlb-little-league-classic-red-sox-orioles

Smith, E. E., & Groetzinger, J. D. (2009, October 28). *Do fans matter? The effect of attendance on the outcomes of Major League Baseball games*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1495195

Tango, T. (n.d.). *Crucial situations Leverage Index (LI)*. InsideTheBook. http://www.insidethebook.com/li.shtml

Zimmer, T. E., Snyder, A., & Bukenya, L. (2021, April 9). *American baseball fans do not influence game outcomes*. AccessEcon. http://www.accessecon.com/Pubs/EB/2021/Volume41/EB-21-V41-I2-P67.pdf