Adi Srikanth - April 30, 3018

A Statistical Analysis of Home Team Performance in High-Scoring MLB Games

Introduction

In Major League Baseball, America's premier baseball league, the home team is considered to have an advantage by virtue of having the support of the crowd. This is backed up by years of statistics that show that teams, almost always, have a higher winning percentage playing at home versus playing on the road. However, in this report, we explore a more nuanced question: do home teams have an advantage in high scoring games compared to in normal scoring games? Before running any statistical tests or analyses, we hypothesize that indeed *when games are high scoring, the home team is especially at an advantage*. We claim this because high scoring games may involve many swings in leads and bursts of offense, which the home crowd can help sway in the favor of their home team. Additionally, as more runs are scored, the home crowd perhaps becomes more rambunctious, energizing the home team. To test this beyond just theory, we move on to the methodology of this analysis.

Methodology/Roadmap

In order to test our hypothesis, we use a dataset, that will from now on be referenced as *MLB Data.* This dataset has information on almost every game in the 2016 MLB season (it does not include unfinished games and other anomalies). It contains the number of runs scored, and whether or not the home team one, along with other extraneous independent variables. MLB Data contains over 2,000 games from the season and was scraped from baseballreference.com, a reputable website for baseball information. Thus we claim that MLB Data is a trustworthy data set. A more detailed explanation of the data set can be found in the appendix.

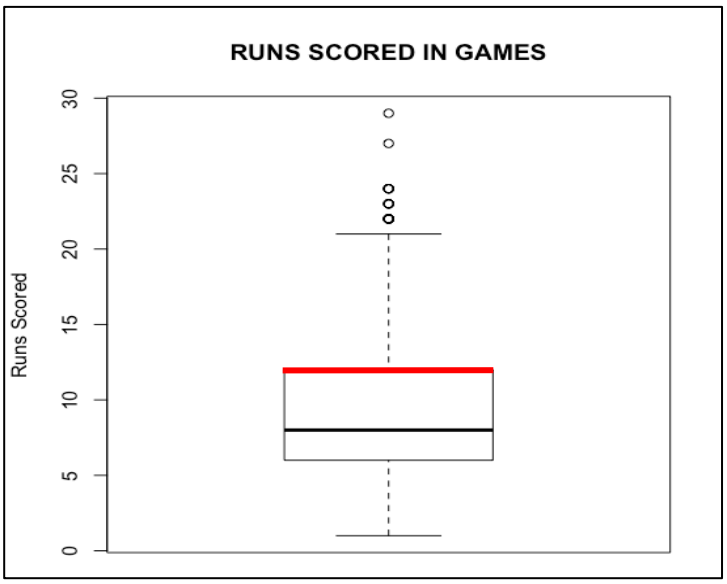
Testing the aforementioned hypothesis is challenging because as noted, the home team almost always has a built-in advantage. Thus, we conduct this analysis by first noting that all winning percentages must be evaluated relatively. Then, we subset MLB Data on games that were high scoring and on games that were normal scoring and check to see if the home team win percentage is higher when games are high scoring compared to when games are normal scoring; in other words we check to see if the home team has an added advantage when the games are high scoring. Finally, if there appears to be a significant relationship, we will attempt to create a model that will predict the likelihood of the home team winning based on the number of runs scored to see if we can further quantify any correlation. All of these steps are elaborated upon in the analysis section.

Analysis

We begin by defining the built-in advantage that the home team has in MLB games. This is a valuable benchmark because it tells us, across all MLB games, the proportion of games that are won by the home team. In order to do this, we count the number of games in MLB Data, sum the number of wins by the home team, and find the proportion of wins to total games. This value is reported as 0.529 (rounded).

The next step is to define a threshold for a "high scoring" game. The figure below, *Figure One*, graphically shows the distribution of the runs scored in all MLB games. Based on this graph, the 75th percentile seems like a reasonable threshold (denoted by the upper boundary of the box) as it is clearly above the mean but not too close to the unusually high values signified by the circles towards the top of the graph.

Figure One



The 75th percentile of the runs scored in all games was calculated to be 12 runs. Thus, 12 runs is our threshold for a high scoring game (anything lower than 12 runs is a normal scoring game). Our next step is to find the home team's winning percentage in high scoring games. We do this by using a temporary dataset composed of games with 12 or more total runs. Then, we find the proportion of home team wins to total games played; this value is calculated to be 0.518 (rounded). Next we find the winning percentage of home teams in normal scoring games by using another temporary dataset composed of games with less than 12 total runs and repeating the previous method. By doing this, we find that in normal scoring games, the home team winning percentage is 0.533. Of course, this seems to indicate that home teams do worse in high scoring games, but to make an accurate conclusion we first must test for statistical significance of our results.

To test for statistical significance, we employ the two-sample z-test for proportions. This test allows us to see if a difference in two proportions is simply by chance or by statistical significance. (We note that the z-test is generally reserved for populations with known parameters, but the MLB Data set is large enough that we can claim to essentially know the population of all MLB games well enough to use a z-test; the Law of Large Numbers will confirm this line of thinking). So, we employ this test using the win proportion of normal games and the win proportion in high scoring games. After running this test, we compute a p-value of 0.7404. Essentially, the p-value tells us how likely our results are if we assume that what we believe is not true actually is true. In this case, we believe that it is *not* true that the home team win proportion is higher normally than in high scoring games. Thus, there is roughly a 74% chance that we obtain our results if in fact the home team win proportion in normal games is greater than or equal to the win proportion in high scoring games. Given this high likelihood, we must say that what we thought was not true could actually be true. Moreover, what we thought was true is likely not true. In plainer words, from this test, we cannot conclusively say that home teams fare better in high scoring games compared to normal games.

While we could not make the conclusion we wanted to make, it is hard not to notice that the computed win percentage in normal games was actually higher than the computed win percentage in high scoring games. So, we attempt to make the opposite conclusion of our initial hypothesis: home teams fare better in normal scoring games than in high scoring games. In our attempt to make this conclusion, we employ another statistical technique called bootstrapping. This technique involves taking the observations we currently have (in this scenario one observation is the result from one game) and using them to estimate new, hypothetical

observations. This is useful because it allows us to generate a data set that has more observations, but retains the same trend that existed before. More observations are useful because as a data set has more observations, trends become clearer. For example, if two people say the prefer a certain flavor there is a weak trend, but if 10,000 people say they prefer a certain flavor, there is a stronger trend; we attempt to apply this principle to our analysis. (We did not attempt this method on the previous test because the previous hypothesis was so far from being concluded as true). This method is essentially answering the question: if the trend of MLB Data continued for several more years, could we conclude anything of statistical significance?

In order to employ bootstrapping we first take our two populations, normal scoring games and high scoring games, and define the general trend they exhibit. *Figure Two* and *Figure Three* below show the patterns that the two populations follow.

Figure Two

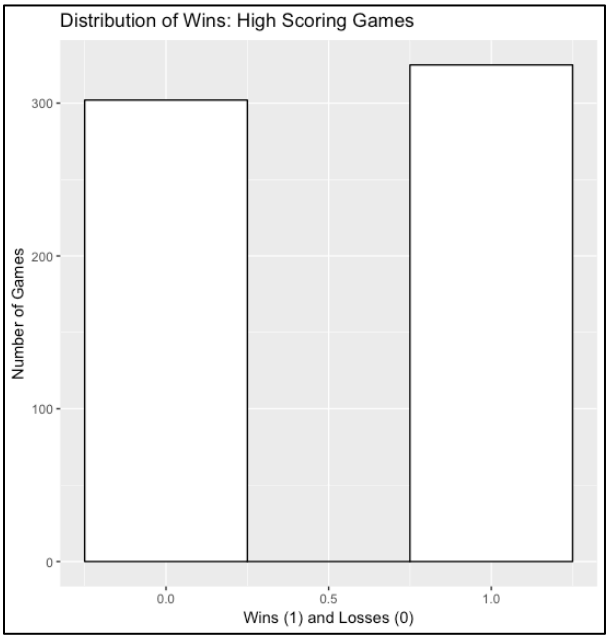
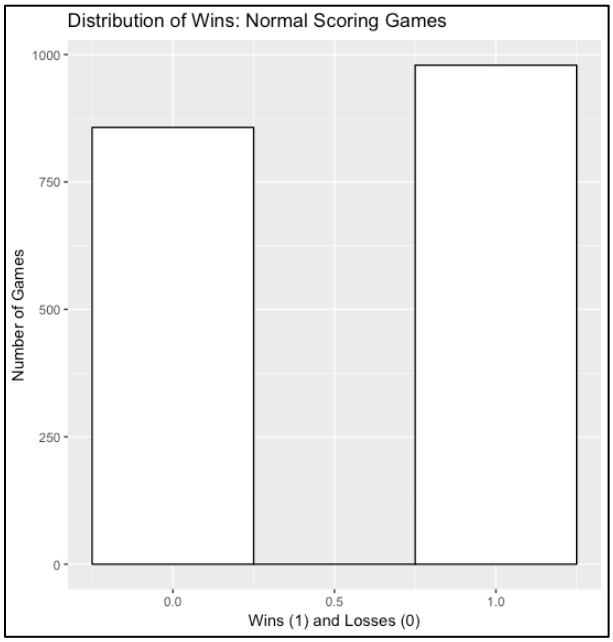


Figure Three



After obtaining these approximations, we implement bootstrapping by producing hypothetical game results for high scoring and normal scoring games that follow the distributions above. For both types of games, we produce 10,000 hypothetical game results. After bootstrapping, as expected, both win percentages are the same, but with more observations we now attempt to make another test of two proportions.

The test of two proportions, again, is the two-sample z-test for proportions and is carried out the same way as before. After running this test, we compute a p-value of 0.01. In theory, this means that if the home team win percentage in normal scoring games was *not* greater than the home team win percentage in high scoring games, there would be around a 1% chance of obtaining the results we obtained. According to the test, there is statistical significance to the claim that home teams fare better in normal scoring games than in high scoring games. However, because we used bootstrapping, we are apprehensive to claim this as generally true. This will be further addressed in the Limitations section of the report.

A Note on Attempting to Produce a Prediction Model

We briefly note here that this report contains no attempt to produce a prediction model. The tests previously conducted with the given data did not yield a high enough significance to warrant a model. Had a model been created, the correlation would have likely been weak and any prediction used based off of the model could not truly have been reliable.

Limitations

In this section we address limitations of the two tests we conducted. The first test conducted was the two-sample z-test for proportions to check to see if the win proportion for home teams in high scoring games was greater than the win proportion for home teams in normal

scoring games. In this test, the clear limitation is that there was no statistical significance found in the test. As a result, we are not able to make any definitively conclusive statements about home teams faring better in high scoring games than in normal scoring games. This said, being unable to conclude something is still quite telling. If there was a clear trend indicating that home team win proportions were greater in high scoring games, we would most likely have concluded so. Thus, based off this test, we can practically say that home teams do not fare better in high scoring games.

The second test was another two-sample z-test for proportions, but this test used data that was produced through bootstrapping. While bootstrapping is a valid statistical method, in this scenario it can be misleading. Through bootstrapping we essentially turned one season's worth of data into over five seasons worth of data. The sample size of one season is large so variation is mostly controlled for. However, the MLB rules and style of play vary from season to season, so assuming that this trend will hold in the future, which is implicitly assumed in bootstrapping, could be incorrect. As a result, it would not be prudent to claim, despite the significance of the test, that home teams fare better in normal scoring games than in high scoring games because the bootstrapped data assumes the continuation of a trend that may not continue. However, we do note that the result of this test is still useful because it tells us that if this trend *does* continue, we can begin to say with confidence that home teams fare better in normal scoring games.

Conclusion

After all analysis, exploration, and measurement of limitations, there are a couple conclusions that we can draw. First and foremost, we can say that the data does not show that home teams tend to have an advantage in high scoring games compared to normal scoring

games. Moreover, we can conclude that if the trend of the 2016 MLB season continues over the next few seasons, we can begin to say that the data show that home teams in fact fare better in normal scoring games opposed to high scoring games.

For future exploration, using multiple seasons worth of data can yield more conclusive results. Furthermore, exploring factors outside of runs scored, such as weather, attendance, and other factors could yield intriguing results.

Overall, we find that there are useful results from this study, but we urge users to understand the complexities of the results and use them with caution and care.

Sources

1. https://www.statisticssolutions.com/sample-size-calculation-and-sample-size-justification-resampling/
2. https://www.r-bloggers.com/r-tutorial-series-multiple-linear-regression/
3. https://www.statmethods.net/advstats/glm.html
4. https://www.theanalysisfactor.com/r-tutorial-part-13/
5. http://www.sthda.com/english/wiki/ggplot2-title-main-axis-and-legend-titles
6. Professor Gretchen Martinet, University of Virginia Department of Statistics

Appendix

An excerpt of the MLB Data dataset (rows 0 through 33) is found below

ID	attendance	runs	hours	win
0	40030	7	3.21666667	1
1	21621	5	2.38333333	1
2	12622	6	3.18333333	1
3	18531	4	2.88333333	0
4	18572	7	2.65	0
5	28386	12	3.5	0
6	12757	5	3.11666667	1
7	28329	3	2.6	0
8	26049	11	3.45	1
9	10478	9	3.46666667	0
10	47820	8	3.28333333	0
11	24123	3	2.48333333	0
12	36911	15	3.78333333	0
13	13468	10	3	0
14	45229	7	2.85	0
15	13371	5	3.08333333	0
16	24318	13	3.75	0
17	40638	5	3.08333333	0
18	43332	10	2.98333333	1
19	26271	11	3.28333333	0
20	40882	8	3.05	1
21	21203	8	2.68333333	0
22	21226	17	3.43333333	1
23	34898	9	3.11666667	1
24	19400	3	2.2	1
25	27938	9	2.71666667	1
26	16783	7	2.83333333	0
27	14160	9	2.95	1
28	45432	5	2.53333333	0
29	41432	16	4.58333333	0
30	21585	7	2.85	0
31	21078	11	2.35	1
32	16112	7	2.71666667	1
33	12777	6	2.68333333	0

ID – the unique identification number of each observation

Attendance – attendance at the game (unused)

Runs – total runs scored in the game

Hours – duration of the game (unused)

Win – whether or not the home team won (1 if yes, 0 if no)