Tucker Atwood
WGU MSDA
D214 Task 3
6/18/24

A. Executive Summary and Implications

The following executive summary details the analysis and results of research performed on the Lahman Baseball Database. This database is described as containing "complete batting and pitching statistics from 1871 to 2023, plus fielding statistics, standings, team stats, managerial records, post-season data, and more." This quotation and the database were retrieved from seanlahman.com.

A1. Problem and Hypothesis

Every season, Major League Baseball teams spend hundreds of millions of dollars on players, coaches, executives, resources, and more, with the ultimate goal of taking home the World Series trophy at the end of October. Players are collected via free agency, trades, drafts, and international signings, and though there is no strict salary cap in MLB, owner-imposed budget restrictions and league-dictated luxury tax penalties demand precise focus on every consideration necessary to constructing an MLB roster. With each player comes a high-leverage decision with serious financial implications, so it is crucial to develop an all-encompassing organizational vision which ensures every choice is made with a championship in mind.

The problem is, while the path to success in the 162-game regular season is often clearly defined and measured, with most of baseball's inherent randomness regressed to its means over the large sample of games, the victor in a 3-, 5-, or 7-game playoff series is often a matter of luck. While it is clear that a World Series-winning team may as well have been the benefactor of a smooth roll of the dice, in the eyes of team members and diehard fanbases, all teams' success is determined by their ability to win the championship. This leaves the question: when assembling a group of players with that concrete destination in mind, are there any particular factors that matter more to playoff success than they do to regular season success?

This analysis aims to discover whether World Series-winning teams are more successful in certain factors than their counterparts who made the playoffs but could not come away with victory. Since this will be addressed using two-sample hypothesis tests, the null and alternative hypotheses will be as follows:

- **Null hypothesis:** "There is no significant difference, in any factors, between Major League Baseball teams that win the World Series and other playoff teams who do not."
- **Alternative hypothesis**: "There is a significant difference in at least one factor between Major League Baseball teams that win the World Series and other playoff teams who do not."

A2. Data Analysis Process

The process of this analysis can be summarized as follows. The language R in the environment R Studio was used for all data manipulation tasks and hypothesis testing.

- The necessary data was extracted from the Lahman Baseball Database and imported into a data set. This included all basic team statistics for all MLB teams in baseball history, including batting, pitching, and fielding statistics.
- These factors were converted into commonly used rates and percentages. Yearly league averages were then calculated for each of these new factors. Finally, these year-adjusted values were used to define new factors which compared individual teams' statistics with the yearly averages. These were created and expressed such that "100" represents league average; values above 100 identified teams that had higher values than average, while values below 100 identified teams that had lower values than average.
- The resulting adjusted factors were used for the rest of the analysis. Included factors were:
    - Batting average (AVG)
    - On-base percentage (OBP)
    - Slugging percentage (SLG)
    - On-base plus slugging percentage (OPS)
    - Runs per game (RPG)
    - Batter walk rate (BBr)
    - Batter strikeout rate (Kr)
    - Batter home run rate (HRr)
    - Stolen bases per game (SBg)
    - Runs allowed average (RAA)
    - Walks and hits per inning pitched (WHIP)
    - Pitcher walk rate (pBBr)
    - Pitcher strikeout rate (pKr)
    - Pitcher home runs allowed rate (pHRr)
    - Defensive efficiency (DEF)
- These statistics were all expressed as described above, with 100 as "average." These averages were calculated using only regular season statistics; therefore, playoff success was not reflected in these calculations. This removes the possibility of playoff victors having inherently better statistics than their competitors simply because they played better in the playoffs.
- Visualizations were created to examine the correlations between factors, the spread and shape of all factors regardless of playoff success, and the comparison of summary statistics between World Series-winners and other playoff teams. These visualizations confirmed several previously understood connections between certain factors, confirmed that playoff teams generally have better regular season statistics than average, and showed a slight advantage toward World Series-winning teams, which was noteworthy since these visualizations only reflected regular season statistics.
- To test the statistical significance of these slight advantages, a series of two-sample one-sided t-tests were conducted on each individual factor. Results are provided in section A3.

A3. Findings

Hypothesis tests were conducted for each of the year-adjusted statistics listed in section A2. For most statistics, higher values were preferable, and since the analysis attempted to determine whether World Series-winning teams performed better than other playoff teams, one-sided tests were used. The aspects of each test were similar to the following, with a change in only the given factor.

- **Null hypothesis:** The average team slugging percentage for MLB teams that win the World Series is not significantly greater than the average team slugging percentage for other playoff teams that do not win the World Series.
- **Alternative hypothesis**: The average team slugging percentage for MLB teams that win the World Series is significantly greater than the average team slugging percentage for other playoff teams that do not win the World Series.
- The **alpha level** will be **0.05**. A p-value less than this value will be considered statistically significant.

Since there were some factors for which it was preferable to have a lower value, the calculated p-values for these hypothesis tests were subtracted from 1 to adjust for the direction considered to be a "better" performance. The statistics with p-values to be adjusted were Kr, RAA, WHIP, pBBr, and pHRr.

The following factors showed statistically significant results, ordered by significance:

- Runs allowed average, p-value 0.0001.
- Runs per game, p-value 0.0001.
- Defensive efficiency, p-value 0.0003.
- Slugging percentage, p-value 0.003.
- On-base plus slugging percentage, p-value 0.003.
- Batting average, p-value 0.010.
- Pitcher strikeout rate, p-value 0.015.
- On-base percentage, p-value 0.018.
- Batter home run rate, p-value 0.033.
- Pitcher home run rate, p-value 0.048.

The following factors did not show statistically significant results, ordered by significance:

- Walks and hits per inning pitched, p-value 0.147.
- Batter strikeout rate, p-value 0.199.
- Stolen bases per game, p-value 0.213.
- Batter walk rate, p-value 0.281.
- Pitcher walk rate, p-value 0.999.

Notes on these results include the following:

- It is not surprising that runs allowed and runs scored are the most significant measures of success for World Series-winning teams. Even though the data used only includes regular season statistics, runs allowed and runs scored are the most basic measures of team success and it makes sense that the best teams at scoring runs and preventing their opponents from scoring runs continued to do so in the playoffs. However, these results do not provide any information on specific player types or strategies, so they are not particularly illuminating.
- Defensive efficiency is perhaps more significant than expected. Many teams throughout history have sacrificed defensive prowess in order to put a talented batter in the lineup. These results may suggest that defense is more important than commonly believed, and could even imply that more sacrifices should be made to the offensive side of the game in order to ensure capable defenders are in the lineup.
- The significant results of the SLG, OBP, and OPS tests, as well as the insignificant results of the batter walk rate tests, suggest that power hitting skills may be more important than on-base skills. This contradicts ideas described in the book and movie *Moneyball*, and the resulting shift in team strategies in the years since the book was published. Though it is still obviously important for batters to try to get on base, the adage pulled from the movie that "A walk is as good as a hit," may not hold true. With insignificant results in the batter walk rate test, as well as the notable difference in significance between the SLG and OBP tests, it may be beneficial to shift focus from patient hitters to power hitters.
- The most significant pitcher-specific factor was strikeout rate, which was not surprising, but could show that more sacrifices elsewhere could be made in order to maximize pitcher strikeouts. These sacrifices include pitcher walks, which will be discussed later on in this section. A shift toward more strikeout-focused pitchers has already begun in the last decade or two, with teams realizing that the more they can keep their opponents' batted balls from even having the opportunity to fall in for a hit, the better. Therefore, this result may confirm current beliefs about the most desired pitcher type.
- Many teams in baseball history have built their offenses around speed, with the idea that putting runners on base and applying pressure to the pitcher and catcher will create a more replicable stream of runs, instead of relying on home runs, which could come in bunches or be subject to long droughts. However, the insignificance of the stolen base test implies that speed is not an important aspect of the game, or at least not one worth building a team around. There is common belief that only smart, capable baserunners should attempt to steal bases, and that stolen base attempts by other players are detrimental to the offense; the lack of significant results here may confirm this suspicion.
- The most surprising result was technically insignificant: pitcher walk rates. The original results, before accounting for the fact that a lower value is preferable, showed significance, which is obvious from the calculated inverse of the original p-value: 0.999. This implies there is actually statistical significance going the other way: that teams with pitchers who walk more batters have more playoff success than teams who walk less batters. This may be the result of pitcher types: it is commonly believed that pitchers who walk more batters, especially those who are still good enough to stick around in MLB, are likely to have high strikeout numbers as well. However, the visualizations showed a very weak correlation

between pitcher strikeout rate and walk rate. Another theory is that pitchers who walk more batters typically stay away from the middle of the strike zone, where batters are able to make more solid contact, so it is possible that these pitchers give up less hits and less power on those hits. Testing these theories would be a great next step in utilizing this data set.

A4. Limitations

Some limitations of this analysis include the following:

- The baseball landscape has dramatically changed since the first World Series hosted in 1903. Though it is usually beneficial to use as much data as is available to answer a research question, it is likely that factors determining team success have changed significantly throughout the years. Many rules have changed that may impact these results, the number of teams in the league and in the playoffs has expanded, and strategies have undergone major modifications as information and statistics have become more available and sought after. For future research similar to this analysis, it may be preferable to separate statistics into separate eras, perhaps based on the years where the number of teams allowed into the playoffs has changed.
- Some statistics used in the calculations of rate and percentage factors included in this analysis were not properly tracked at earlier points in MLB history and showed up as missing data in the initial data set. These statistics include sacrifice flies and hit-by-pitches, which were imputed using overall averages for the years in which they were missing. Fortunately, these two factors did not have a large effect on the statistics they were used to calculate, but removing them from the formulas would have resulted in the formulas not being properly calculated. For future research, it may be helpful to limit the data to post-1970, where all factors needed for calculations are available.
- Hypothesis tests are limited in that they can only help determine whether a difference in means is statistically significant or not; they do not function as predictors. Any results of tests performed in this analysis are merely describing past events, and they cannot be used as a means of predicting future team success.

A5. Proposed Actions

The results of this analysis have informed the following potential actions to take. All are framed as actions that teams may take in order to increase their chances of playoff success.

- Players' defensive skills should be valued higher than they currently are. With results more significant than all other factors except for runs allowed and runs scored, it is clear that defensively successful teams are more successful in the playoffs than other teams. This implies that often it is wise to sacrifice a certain amount of offensive talent in order to ensure there are capable fielders defending their positions.
- Power hitters should be valued higher than patient hitters. Though there is obviously some overlap in these types, there are definitely hitters whose power outweighs their on-base skills, and vice versa. If considered of equal value otherwise, it is recommended to target players who emphasize slugging percentage over on-base percentage.

- Strikeout pitchers are preferred over other equally valued pitchers. Though some pitchers may experience plenty of regular season success by limiting hard contact and relying on their defense to make plays, it is ultimately better to prevent batted balls in the first place. Importantly, it is apparent that sacrifices made in pitcher walk rates are acceptable in order to ensure a pitcher strikes out more batters.
- An overall team strategy focus should not be spent on baserunning and stolen bases unless the assembled players are also viable at other aspects of the game. These runners should also be deemed smart and capable of minimizing their risks on the basepaths, as it is evident that these risks may outweigh the benefits of focusing on baserunning. Though fans love the excitement of old-school stolen base totals, it is far from the most important aspect when constructing a roster.
- As mentioned, it is abundantly clear that pitcher walk rates are not as important to avoid as commonly believed. Though it is never a good thing to allow an opposing hitter to reach base, it may be useful in limiting hits, including home runs and other hard contact hits, and in achieving higher strikeout rates. The results of the pitcher walk rate hypothesis test were contradictory, so a further proposed action relating to these results would be to further study what may cause this contradiction.

A6. Benefits of the Study

This study aims to provide a better understanding of what constitutes playoff success, and what factors are most important in building a championship-caliber team. While the results and the proposed actions may help a team realize playoff success, specific quantitative measures of any benefits are impossible to predict from hypothesis testing alone. With that said, the potential benefits of utilizing these results in roster construction are as follows:

- A greater probability of success in the playoffs, including winning the World Series.
- An increase in local and national fan support, including ticket sales, merchandise sales, television viewers, and media coverage.
- More interest in the team from potential contributors to future success, including raised probabilities of convincing free agents and drafted players to sign with the team, players already rostered to sign extensions, and sought-after coaches to choose the team.
- Increased revenue from national playoff games, ticket sales, and media recognizability.

B. Presentation of Findings

Please see the attached Panopto video for a PowerPoint presentation of these findings to a non-technical audience.

C. Sources

There is no content in this analysis that has been quoted, paraphrased, summarized, or otherwise requires direct citation. The data set is from the Lahman Baseball Database, which was obtained from seanlahman.com.