STAT 3494W Project Proposal

Ajay Natarajan

October 09, 2023

1 Introduction

Statistical analysis in sports has been a burgeoning field over the past few decades. In baseball in particular, "analytics" have become especially prominent, with several professional organizations using advanced statistical methods to achieve success and gain a competitive edge. In baseball, one area where it is commonly used is on the side of pitching, enabling pitchers to optimize spin rate, grip, movement, and deception. However, an area with relatively limited study is pitcher injuries. Pitchers are at a relatively high risk of injury, and it has become a growing issue through recent years. Therefore, my research goal is to analyze several baseball factors, such as size, pitch types, pitch velocities, and volume, to try to determine which factors contribute most significantly to pitcher injuries in Major League Baseball and whether these are controllable by teams and players.

2 Aims

This study aims to focus on three major objectives. Firstly, my primary objective is to answer my research question: "What factors increase risk of pitcher injury, and are these controllable?" I am aiming for this query to serve as my main point of reference for my statistical investigation.

Additionally, I'm hoping for this overall research question to translate into specific statistical queries that I am able to analyze and test. I plan to use R for any of my data analysis, and am hoping that I can use R markdown to come up with a functioning document. I plan to use regressions and correlation analysis to establish connections, and potentially do further research to potentially establish evidence for causation.

Finally, I'm hoping to extend the data and my results gathered into a meaningful discussion on pitcher injuries and the factors that contribute to causing them. I may look over time to see whether these factors have changed. I also plan on extending out my discussion into which factors might be more controllable by teams.

3 Data

I plan on relying on data from various reliable baseball data sites, particularly featuring Fangraphs, Baseball Reference, and Baseball Savant, and potentially including Brooks Baseball and Pitcher List. I also will see if there are any updated, useful R libraries that I might be able to import for an easily usable set of functions and data to use. I am aware that Lahman exists and is a fairly common dataset used in this sort of analysis, but I do not know how far it is updated through. I plan on utilizing data from 2004 through 2023 for two reasons: most of Fangraphs' pitcher value metrics go back to 2003, and this stretch marks a convenient two-decade interval to track changes. I can use these websites and libraries for certain factors. I also plan on utilizing MLB.com transaction updates for reliable data on injury date and type.

4 Research Design and Methods

I initially plan on doing a thorough analysis of the data and organizing together relevant information within the websites, as well as setting an overall sample of pitchers to utilize in my analysis. I plan on focusing on variables such as pitch speed, pitch type, pitcher size, injury type, and pitcher volume for my sample of pitchers. Ideally, I will do a season-by-season analysis, but this may have some flaws, such as multi-seasonal injuries, so I might have to find a different method of analysis to see a trend over time. A potential strategy I might do is to create a statistic measuring how "injury-prone" a pitcher is based on games played out of their time in the league, excluding suspensions/leave.

I plan on using statistical methods such as correlation analysis and regression analysis to discover significance of variables and their respective weights on injury risk, and ideally will be able to differentiate these by type of injury.

On the basis of this data analyzed, I also plan on trying to use predictive modeling to potentially determine pitchers who may be at risk of injury in the future, as well as crafting a "profile" of an at-risk pitcher, pinpointing elements most responsible.

I plan on first collecting data, then analyzing it and sorting it into usable formats, before executing data analysis, and then drawing conclusions. I hope to spread this out over a reasonable timeframe so I have time to spend on each subject.

5 Discussion

There may be challenges in utilizing this data. For one, I still need to find a metric to determine a player's risk of injury, or otherwise find a way to incorporate that into my analysis without running into issues such as multi-seasonal injury. Additionally, the data will need careful collection and processing to ensure it is correct and in usable formats, since typically it will be imported as a csv file.

In the event of any challenges, I will document my difficulties and potentially step back and follow a new method to work past those difficulties.

6 Conclusion

Overall, my study will conduct a thorough investigation into pitcher injuries in the MLB, utilizing statistical analysis to determine which factors are most responsible for pitcher injuries in the league. I'm hoping to provide insights into how pitching injuries happen, and eventually how they can be limited, since they take opportunities away from some of the best talent in the league, and lead to optimization issues for teams in roster management.