Ethan Newcomb
Matthew Murnane

STAT 550 Project Proposal: Does Pitch Velocity Determine Elite Pitching?

Our project focuses on developing a multivariate statistical model to distinguish Cy Young Award–caliber MLB pitchers from the general population of qualified pitchers. The analysis is based on two sources: (1) the AwardsSharePlayers.csv table from the Lahman Baseball Database, which provides Cy Young voting data such as votes earned and (2) manually collected pitch-velocity datasets from Baseball Savant for the 2017–2023 seasons (excluding 2020). For every season in this window, we extracted four main pitch types (fastball, slider, changeup, and curveball) using Baseball Savant's "Qualified" filter to ensure that each pitcher had a sufficient sample size. The resulting dataset included the average pitch speed for those pitches, from each qualified pitcher. We then merged the Baseball Savant velocity data with the Cy Young voting dataset. Only pitchers who remained on a single team for the full season were kept. See the attached dataset "master_data." There are 301 total pitchers, where 22 of them received votes for the Cy Young. This is expected as only a handful (10-20) pitchers per season receive votes.

The objective of our project is to build a multivariate profile of pitcher performance using these modern indicators and to examine whether velocity-based characteristics align with Cy Young Award voting. We will use Principal Component Analysis (PCA) to identify latent dimensions in the pitch-velocity data (examples being power or balance) and then apply Discriminant Analysis (DA) to classify pitchers into two groups: those who received at least one Cy Young vote and those who received none. PCA will allow us to visualize clusters of pitchers and understand the dominant structure of their pitch characteristics, while Discriminant Analysis will quantify how well these latent velocity profiles distinguish award-caliber seasons from non-award seasons. Finally, we will validate our approach using 2024 as a test year for prediction. This project aims to reveal whether the velocity component of pitching corresponds to Cy Young voting outcomes. One may assume that "the pitchers who throw the hardest will be most commonly receiving votes," but this is a misconception as there is much more that makes an elite pitcher such as accuracy, speed differential (difference between fast and slow pitches), unpredictability, etc. However, this project only focuses on whether pitch speed alone can determine a Cy Young-caliber pitching profile.