Determining Optimal Pitch Selection Using Run Values and PITCHf/x Data

Benesch, Cameron

Objective: To create a mathematical model which, when applied to a Major League Baseball hitter's pitch-by-pitch data and a pitcher's pitch selection, returns the optimal pitch selection given the number of balls and strikes. Procedure: Data from Retrosheet.org are used to calculate average run values of at-bat outcomes. Individual pitch outcomes are used to calculate corresponding Bayesian changes in the probabilities of each at-bat outcome. These two numbers are multiplied to return the expected change in run value caused by any given pitch outcome in any given count. This new metric, Expected Run Value Added (ERVA), is determined for each combination of pitch outcome and count. PITCHf/x data are used to create a standardized similarity formula, which is then used to quantify the "similarity" between any two pitches. ERVA values and similarity weights are appended to the hitter's 2016 PITCHf/x data, and the target pitch is continually adjusted in order to calculate multiple weighted averages of the hitter's ERVA against each pitch type and location in the pitcher's repertoire. These pitch-by-pitch averages reveal the opposing hitter's demonstrated strengths and weaknesses. Conclusion: The ERVA table by count and pitch outcome is extremely viable. In 180 out of 180 values generated, the sign (positive or negative) matches the play's direction, and the magnitude matches intuitive expectations. The similarity weights, as expected, closely follow a lognormal distribution. Finally, the pitch selection results show some correlation with simpler measures of hitter's strengths and weaknesses, and interpolation within the training data yields convincing results. However, until the model is exposed to new test data in future MLB games, its precise predictive ability is unclear.

Awards Won: