ISEN 427/627: Final Project
Due date: Upload onto *Canvas* by 5/6/25 (5:00 pm)

NOTE: You may work in groups of at most four (4) students. There shall be no collaboration outside each group.

Baseball umpiring is a skilled job, requiring sustained mental effort. The most significant task that the home plate umpire faces in his working day is "calling" the game: deciding which pitches are balls and which are strikes. A pitch should be called a strike if any portion of the baseball passes through the strike zone.

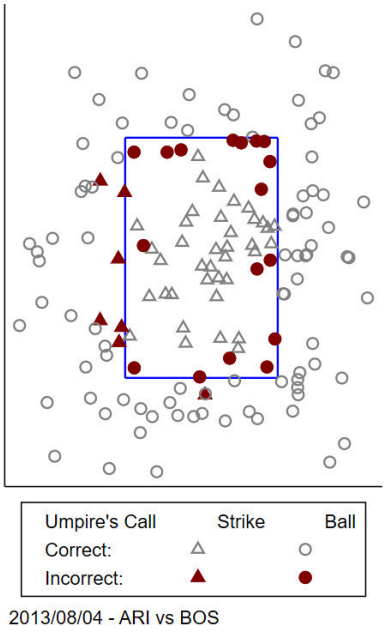

Figure 1

In an average game, an umpire makes calls on around 120 pitches. A high-precision pitch-tracking technology called PITCHf/x has been in operation at every MLB ballpark since 2008. Figure 1 presents a spatial scatterplot of the true locations of pitches upon which the umpire had to make a call in one game, as generated by PITCHf/x. Umpires make both Type 1 and Type 2 errors. A solid triangle in the plot denotes a pitch that passed outside the zone that an umpire erroneously called a strike. A solid circle indicates that a pitch passed through the zone, but the umpire called it a ball. Note that pitches close to the strike zone boundary are more likely to be called incorrectly.

The purpose of this project is to build a Bayesian logistic model of call precision for individual MLB umpires. The dataset consists of all registered pitches from MLB games spanning several seasons officiated by three umpires, downloaded from the source: `baseball.savant.mlb.com`.

The data is available here. There are 3 datasets, each representing calls made by a chosen home plate umpire. The strike box is divided into 14 zones. An umpire's call is incorrect if for example a pitch that lands on zones $\{11, 12, 13, 14\}$ is called a strike. An umpire's call is also incorrect if a pitch that lands in zones 1 to 9 is called a ball. The description of all other columns is available at `https://baseballsavant.mlb.com/csv-docs` and at `https://www.mlb.com/glossary`.

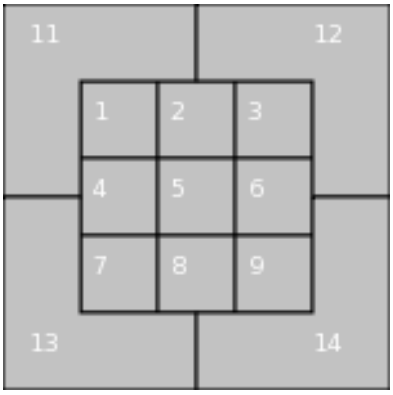

Figure 2: Zone codes in Savant

Specifically, in a logistic regression model the log-odds of correctly calling a pitch is written as:

$$\log \frac{\mathbb{P}(y_i = 1|\beta)}{\mathbb{P}(y_i = 0|\beta)} = X_i^\top \beta$$

where $y_i = 1$ if the i-th pitch is correctly called (and $y_i = 0$ otherwise), $X_i^\top \in \mathbb{R}^p$ are the features describing the i-th pitch in the dataset and $\beta \in \mathbb{R}^p$ is the parameter vector. Examples of features include pitch type, ball-strike count (# balls and strikes before the pitch), Leverage index (measures the importance of a particular event), horizontal (resp. vertical) position of the ball when it crosses home plate from the catcher's perspective, etc.

Final Project

1. Build several Bayesian logistic models of precision for each umpire by selecting different subsets of features for example including *pitch location, zone, pitch number, hitter's stand, pitcher is left or right handed.*
2. Compare the models. Evaluate posterior predictive.
3. What are the most relevant features to explain an umpire's precision ?

Instructions for Report

1. Your results, analysis and conclusions should be included in a final report not to exceed 10 pages. The contribution of each individual in the project should be clarified.
2. Describe step by step all the derivations and computation needed for obtaining results (any code used must be included and does not count in the 10 page limit).