Analysis of the Influence of Bat Speed and Swing Length on the Likelihood of Fouling Off 2-Strike Pitches

Objective

This analysis explores how batter mechanics (bat speed and swing length), their interaction with pitch characteristics (release speed and movement), and the mediation of game context (inning, outs, and balls) as covariates influence the likelihood of fouling off 2-strike pitches. This analysis will help in understanding player performance and decision-making in high-pressure situations. One critical skill in baseball is fouling off 2-strike pitches as it allows batter to extend at-bats, tire pitchers, and increase the chances of favorable outcomes and it helps pitchers control the zone and set up strikeouts. Identifying these factors that contribute to fouling off such pitches, teams can improve coaching strategies, player development, and game tactics.

This study could lay the groundwork for a survival model in exploring how batters stay in the game by fouling off multiple pitches with two strikes. Such a model could analyze the factors contributing to prolonged at-bats providing attributes of successful batters. This analysis is limited to the first aspect and the second part can be explored further.

Data Selection

Identifying Two-Strike Pitches

Two-strike situations in baseball are critical moments where a single pitch can determine the outcome of an at-bat. To analyze the likelihood of fouling off 2-strike pitches, the dataset was filtered to include only pitches where the batter faced two strikes. This targeted approach ensures that the analysis focuses on scenarios where fouling is a deliberate strategy rather than a consequence of other circumstances.

Within the dataset, four types of foul outcomes were identified in the description column:

- **foul**: Regular foul ball, which is the most common type of foul.
- **foul_tip**: A foul ball tipped by the batter but caught by the catcher, leading to a continuation of the at-bat unless it's the third strike.
- **foul_bunt**: A foul ball resulting from a bunt attempt.
- **bunt_foul_tip**: A bunt attempt that results in a tipped foul caught by the catcher.

In this research, only the **foul** and **foul_tip** were included as indicators of fouling likelihood. This selection is justified for the following reasons:

- **Relevance to the Research Question**: The study focuses on batter mechanics and pitch characteristics which are directly tied to regular swings and not bunts. Conventional swings result in foul and foul_tip results which makes them ideal for analyzing the mechanics of batting and the interaction with pitch characteristics.

- **Exclusion of foul_bunt and bunt_foul_tip**: They are outcomes specific to bunting with a deliberate strategy that differs significantly from regular swings. Bunting involves a distinct set of mechanics and objectives that do not align with the research focus on swing mechanics. Including these could introduce noise and bias, as the mechanics and context of bunts are unrelated to the conventional batting scenarios being analyzed.

Data Transformation

Creating the Dependent Variable

To simplify the analysis, a binary variable foul_occurred was created to indicate whether a foul occurred (1 for foul or foul_tip, and 0 otherwise). This variable was added to the dataset containing only two-strike pitches, making it possible to quantify the likelihood of fouling off a two-strike pitch.

Data Cleaning

Handling Missing Values

Columns with 100% missing values were identified and removed, as they provided no meaningful contribution to the analysis. Also, rows with missing values in relevant columns (bat_speed and swing_length) were also removed. This process reduced the dataset from 103,692 rows to 58,566 rows which ensures that the analysis was based on complete and relevant data.

Foul Occurrence Distribution

The cleaned dataset was analyzed to ensure a balanced distribution of the dependent variable (foul_occurred) and it showed a reasonably balanced distribution, allowing for meaningful comparisons:

- 34,159 rows where foul 2 off-strikes didn't occur (foul_occurred = 0)
- 24,407 rows where a foul 2 off-strikes occurred (foul_occurred = 1)

Transforming Categorical Variables

Batter Stance (stand) and Pitcher Throwing Hand (p_throws) variables, was transformed into a numeric variable:

- R was encoded as 1
- L was encoded as -1

METHODS

Research Design and Data Sampling

This study employs a causal inference framework using Bayesian econometric models to analyze the likelihood of fouling off 2-strike pitches. It systematically investigates the influence of batter mechanics (bat speed and swing length), pitch characteristics (release speed, vertical and horizontal movement), and game context variables (inning, outs, balls) on fouling behavior. The dataset was sourced from Statcast, comprising pitch-level data from 346,250 MLB plate appearances between April 2, 2024, and June 30, 2024. The analysis focuses on pitches tracked for bat speed and swing length, specifically in 2-strike situations to evaluate plate protection behavior. Data cleaning ensured the removal of incomplete or anomalous entries, leaving 58,566 valid plate appearances.

Statistical Analysis

The statistical analysis used Bayesian logistic regression to estimate the likelihood of fouling off a 2-strike pitch with bat speed, swing length, pitch characteristics, and game context as predictors. Bat speed and swing length interact with pitch characteristics, while game context is treated as a covariate. Directed Acyclic Graphs (DAGs) were used to establish causal pathways and account for confounder and covariate, enabling robust causal inference. Parameter estimation relied on Markov Chain Monte Carlo (MCMC) methods to generate posterior distributions, with High-Density Intervals (HDI) providing credible intervals for the strength and direction of effects. Contrast analyses were conducted to compare fouling probabilities across key conditions (e.g., low vs. high bat speed, short vs. long swing length). Interaction effects between batter mechanics and pitch characteristics were also modeled to evaluate their combined impact.

Data Sampling and Justification

From the 58,566 cleaned plate appearances data, a random sample of 5,000 pitches was selected for the Bayesian causal analysis. This sample size balances computational efficiency with statistical rigor for precise Bayesian logistic regression estimates. The sampling process captured a representative subset of batter-pitcher interactions while maintaining an emphasis on 2-strike situations relevant to the study's focus.

Bayesian inference methods prioritize exploring the posterior distribution of model parameters, such as the effects of bat speed, swing length, and game context variables on fouling likelihood, rather than mirroring the full dataset size. Advanced sampling techniques, like the No-U-Turn Sampler (NUTS), efficiently approximate the posterior distribution using as few as 2,000–5,000 samples. This ensures posterior is a representation of the relationship between the parameters and the data, independent of the dataset's size and Bayesian methods assures convergence to the true posterior distribution even with a smaller sample size.

Model Convergence and Diagnostics

Bayesian analysis emphasizes model convergence over raw sample size. Validating the adequacy of posterior samples are done through diagnostics like R-hat and Effective Sample Size (ESS):

- **R-hat:** An R-hat value close to 1 confirms that the Markov Chain Monte Carlo (MCMC) sampler has converged, indicating additional samples would not significantly alter the results.
- **ESS:** Assesses the number of independent posterior samples. A well-converged model typically achieves sufficient ESS with 2,000–5,000 samples.

These metrics ensure that the posterior samples accurately capture parameter estimates without unnecessary computational overhead. Analyzing the full dataset of 58,566 rows for causal modeling would provide diminishing returns in terms of accuracy while increasing computational complexity.

Using the Full Dataset for Prediction

While a subset of 5,000 samples is sufficient for building and validating the causal model, the entire dataset of 58,566 plate appearances will be used in case of predictive modeling. Prediction tasks, unlike causal inference, rely on leveraging the complete dataset to maximize predictive accuracy.

Bayesian Inference Workflow for Fouling Off 2-Strike Pitches

Defining the Generative Model

The likelihood of fouling off 2-strike pitches is modeled using a Bayesian logistic regression framework. The generative model captures the relationship between fouling likelihood and its predictors, including batter mechanics, pitch characteristics, and game context variables.

Outcome:

- foul_occurred: A binary variable where 0 indicates no foul, and 1 indicates a foul event.

Predictors:

- Batter mechanics: bat_speed, swing_length, stance
- Pitch characteristics: release_speed, pfx_x, pfx_z, plate_x, plate_z, p_throws
- Game context: inning, outs_when_up, balls.

Logistic Regression Framework:

The probability of a foul event is defined as:

$$P(\text{ foul\_occurred} = 1 \mid X\,) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \cdot \text{bat\_speed} + \beta_2 \cdot \text{swing\_length} + \beta_3 \cdot \text{release\_speed} + \beta_4 \cdot \text{pfx\_x} + \beta_5 \cdot \text{pfx\_z} + \beta_6 \cdot \text{inning} + \beta_7 \cdot \text{outs} + \beta_8 \cdot \text{balls})}}$$

Interpretation of Coefficients (βi)

Each coefficient represents the strength and direction of the link between a predictor variable and the likelihood of a foul event. Positive values increase the probability of a foul, while negative values decrease it.

Estimands

The estimands for this study include **Posterior Distributions of Model Coefficients (βi)** which provide insights into each predictor variable's impact on the likelihood of a foul and **Fouling Probability** given as $P(\text{ foul\_occurred} = 1 \mid X)$ which estimates probability of fouling off a pitch under specific conditions.

Estimator Design:

This Bayesian logistic regression estimate model coefficients, with priors for the coefficients β_i set as Normal (0,10) based on expert knowledge (i=0,1,2,3,4,5,6,7,8).

Posterior Distribution: In Bayesian statistics, the posterior distribution represents the updated beliefs about the parameters after observing the data. Mathematically, the posterior distribution is derived using Bayes' theorem, which combines the prior distribution $P(\beta)$ and the likelihood $P(Y \mid \beta, X)$.

Analysis and Interpretation

The model was applied to the actual dataset, which included 5,000 randomly sampled pitches from the cleaned dataset of 58,566 rows. Markov Chain Monte Carlo methods were used to sample from the posterior distributions, providing robust estimates of the coefficients.

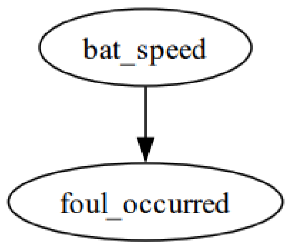
Posterior Distributions: For each predictor, posterior means, 95% credible intervals, and other summary statistics were calculated to evaluate the strength and direction of their effects on fouling likelihood.

Interpretation of Results: Significant predictors were identified based on whether their 95% credible intervals excluded zero.

Research Questions:

This research aims to answer the **primary question**: how batter mechanics (bat speed and swing length), pitch characteristics (release speed and movement), and game context (inning, outs, and balls) as covariates influence the likelihood of fouling off 2-strike pitches. To address this, the study explores the following **seven sub-questions** an each is addressed below:

Q1: Does bat speed directly influence the likelihood of fouling off a 2-strike pitch?



Result:

- **Contrast Between High and Low Bat Speed**:

  Mean Difference (contrast): -0.035 (95% HDI: [-0.062, -0.008]).

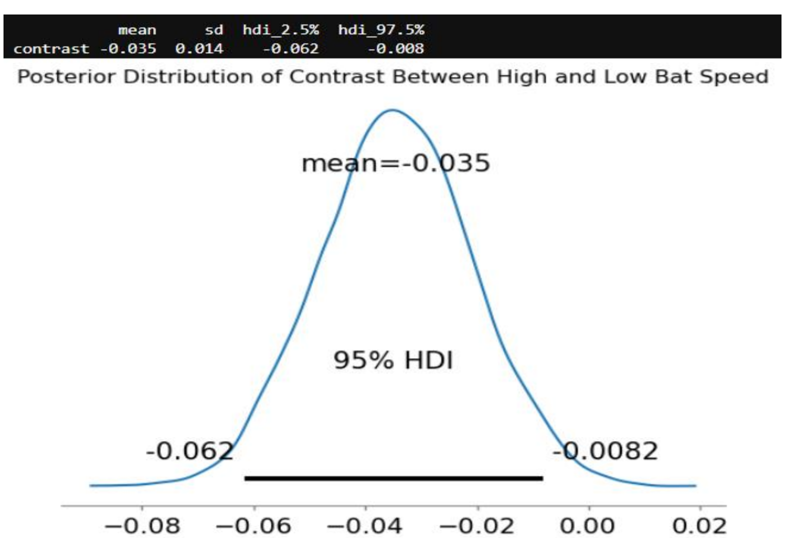
  Significance: The 95% HDI does not include zero, showing that result is statistically significant and not due to random variation.

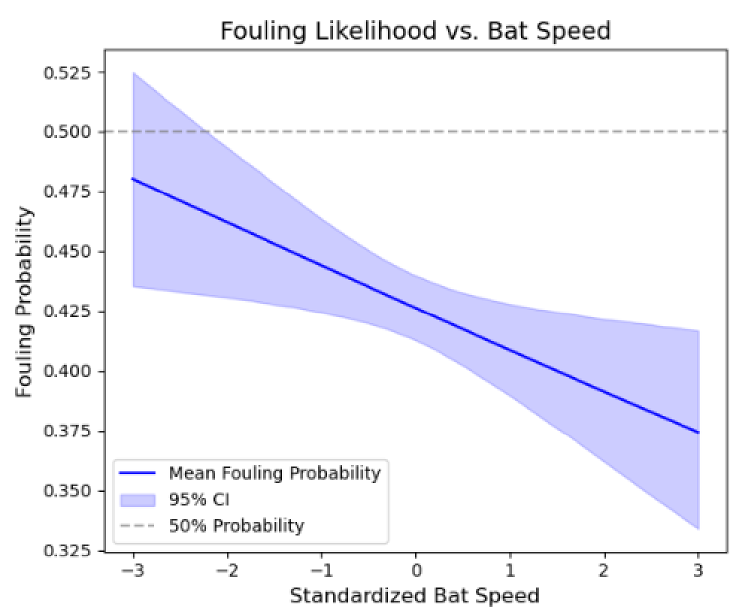
- **Fouling Probabilities**:

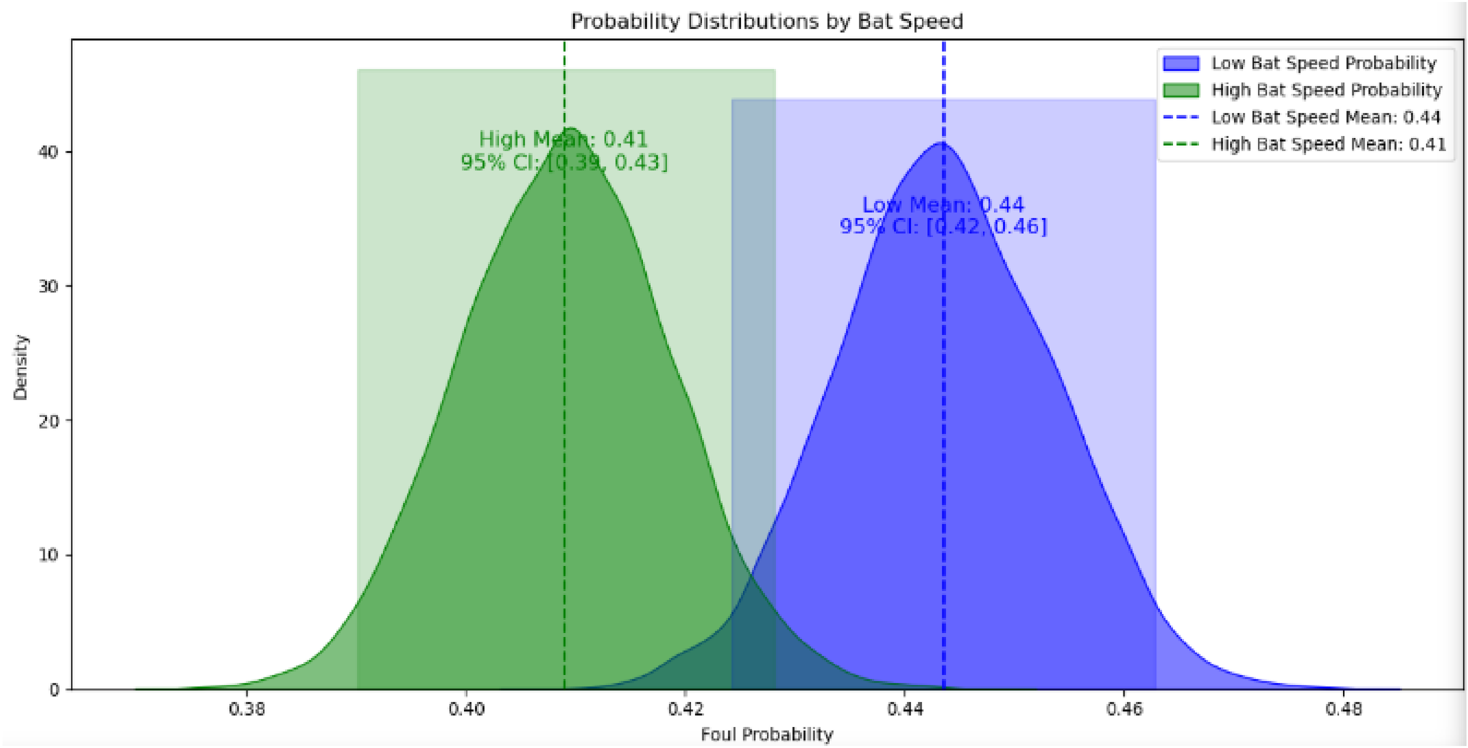
  Low Bat Speed: 44% (95% CI: [0.42, 0.46]).

  High Bat Speed: 41% (95% CI: [0.39, 0.43]).

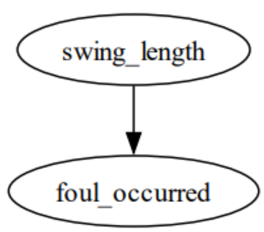
Higher bat speed reduces fouling probability significantly, with the effect being both statistically credible and practically meaningful.







Q2: Does swing length directly influence the likelihood of fouling off a 2-strike pitch?



Results:

- **Contrast Between High and Low Swing Length**:

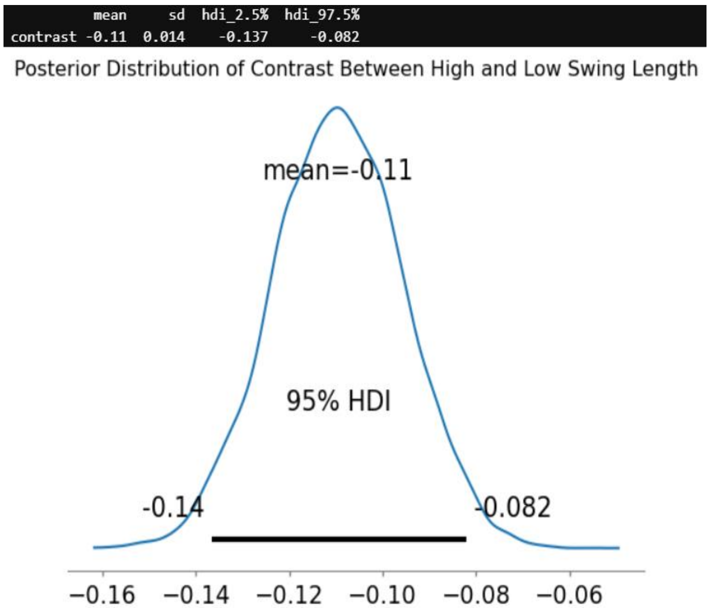
    Mean Contrast: -0.11 (95% HDI: [-0.137, -0.082]).

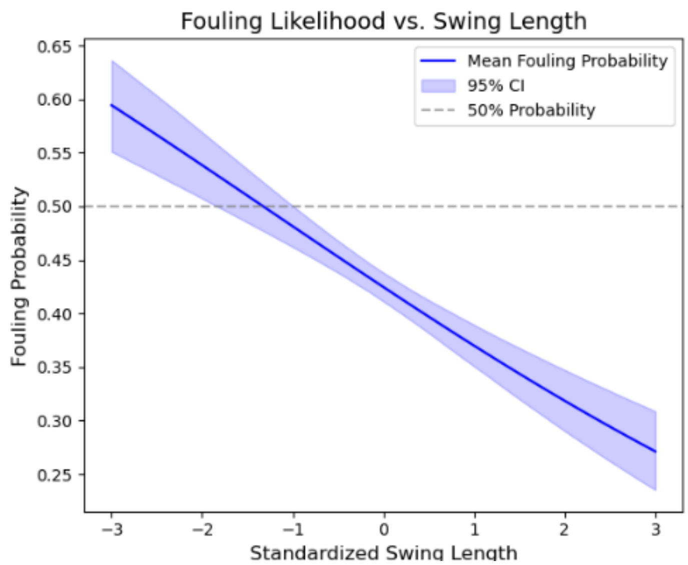
    Significance: The 95% HDI excludes zero, confirming the result is statistically significant.

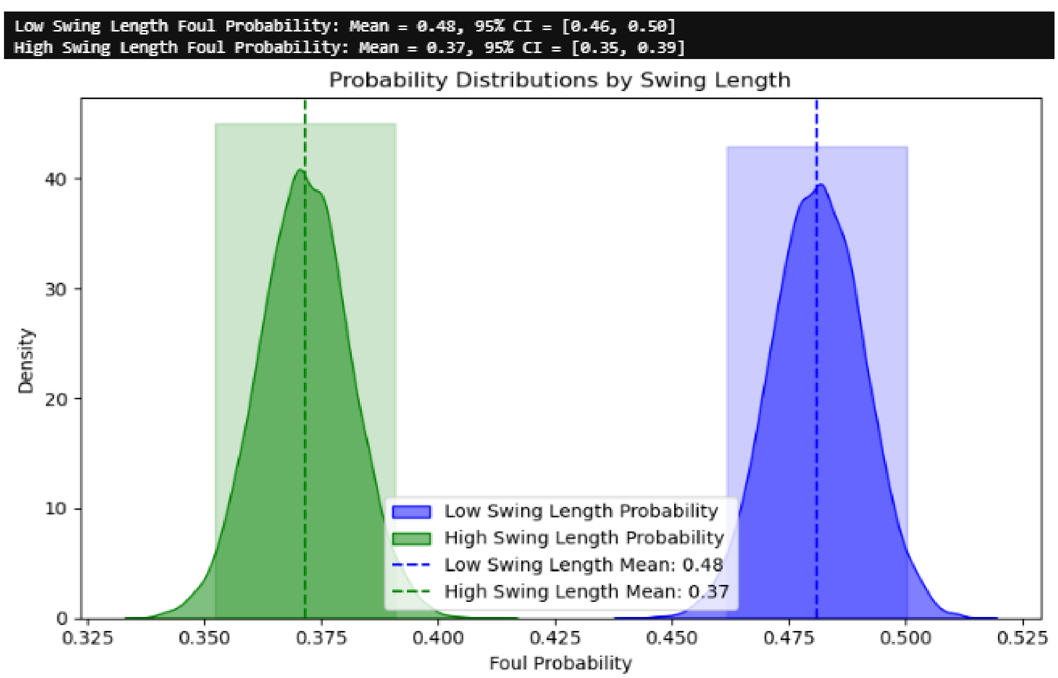
- **Fouling Probabilities**:

- **Low Swing Length**: 48% (95% CI: [0.46, 0.50]).
- **High Swing Length**: 37% (95% CI: [0.35, 0.39]).

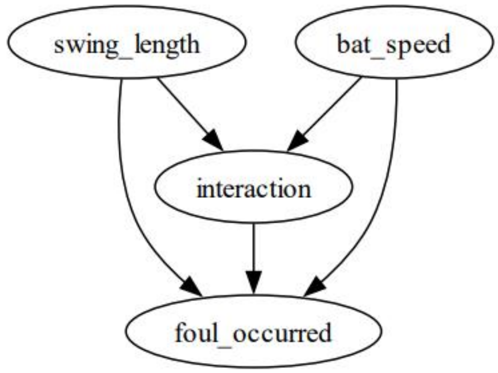
Swing length significantly impacts fouling off a 2-strike pitch., with shorter swings increasing fouling probability due to better control and adaptability. Batters can improve plate protection by shortening their swings in high-pressure situations, potentially extending at-bats and tiring pitchers.







Q3: How do bat speed, swing length, and their interaction influence likelihood of fouling off a 2-strike pitch?



Results:

- **Interaction Effects**:

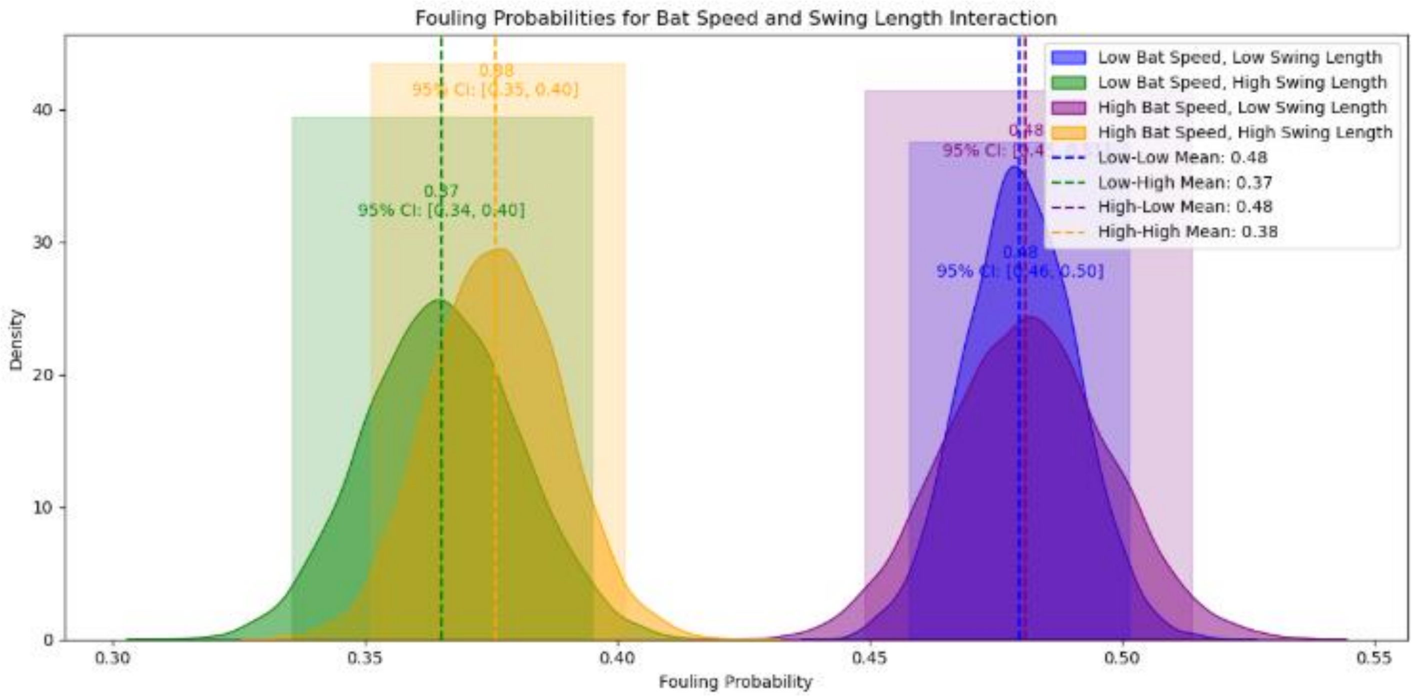
  High Bat Speed & Low Swing Length vs. High Bat Speed & High Swing Length:

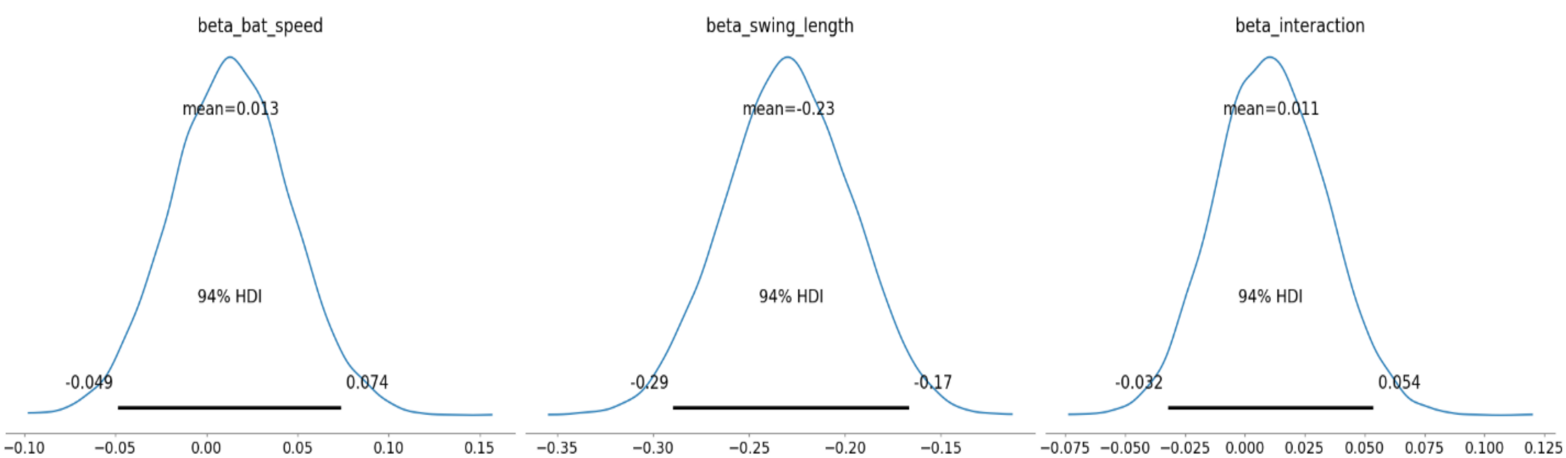
  - Mean Contrast: -0.105 (95% CI: [-0.145, -0.063]).
  - **Significance**: Significant, as the 95% CI does not include zero.

  Low Bat Speed & Low Swing Length vs. Low Bat Speed & High Swing Length:

  - Mean Contrast: -0.115 (95% CI: [-0.146, -0.083]).
  - **Significance**: Significant, as the 95% CI does not include zero.

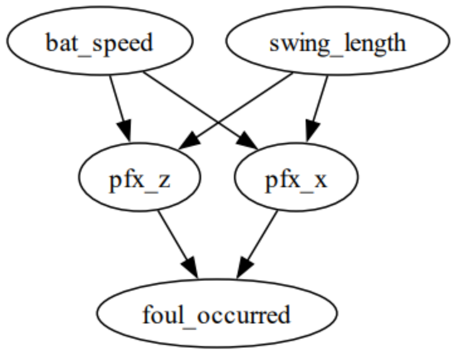
- **Fouling Probabilities**:
  - High Bat Speed & Low Swing Length: 48% (95% CI: [0.46, 0.50]).
  - High Bat Speed & High Swing Length: 38% (95% CI: [0.35, 0.40]).
  - Low Bat Speed & Low Swing Length: 48% (95% CI: [0.46, 0.50]).
  - Low Bat Speed & High Swing Length: 37% (95% CI: [0.34, 0.40]).





Swing length (-0.23) has a stronger and statistically significant effect on fouling probability than bat speed (0.013). The interaction term (0.011) is modest but statistically credible, suggesting a combined influence of mechanics on fouling behavior.

Q4: How do batter mechanics (bat speed, swing length) influence pitch movement (pfx_x, pfx_z), and how do these movements impact the influence likelihood of fouling off a 2-strike pitch?



Results:

Fouling Probabilities Across Combinations:

- **Highest Fouling Probabilities**:

  Combination of Low Bat Speed, Low Swing Length, Low pfx_x, High pfx_z.

  Mean = 0.50, 95% CI = [0.48, 0.53].

  Interpretation: Higher fouling likelihood occurs when vertical movement (pfx_z) is high, and batter mechanics are weak (low bat speed, low swing length).

- **Lowest Fouling Probabilities**:

  Combination: High Bat Speed, High Swing Length, High pfx_x, High pfx_z.

  Mean = 0.39, 95% CI = [0.36, 0.43].

  Interpretation: Strong mechanics reduce fouling probability even when pitch movement is challenging.

Pairwise Contrasts:

Low Bat Speed, Low Swing Length, Low pfx_x, Low pfx_z vs. High Bat Speed, High Swing Length, High pfx_x, High pfx_z:

Mean Difference = 0.07, 95% CI = [0.00, 0.13].

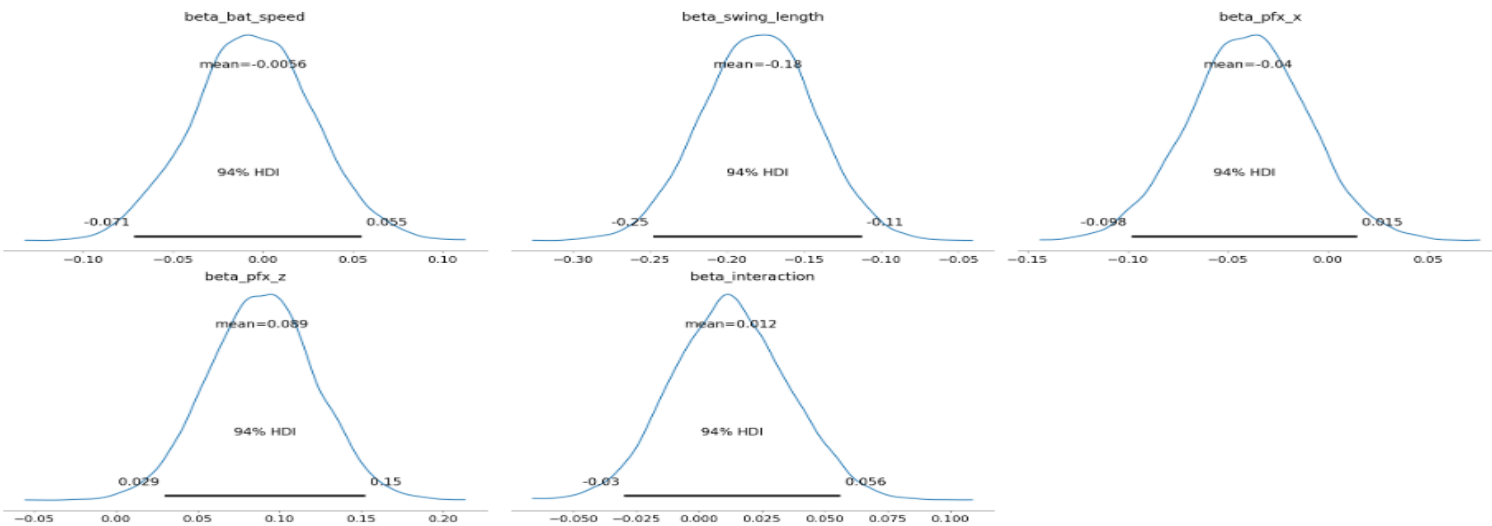
Interpretation: The 95% CI barely excludes zero, indicating borderline significance.

Other contrasts show non-significant differences.

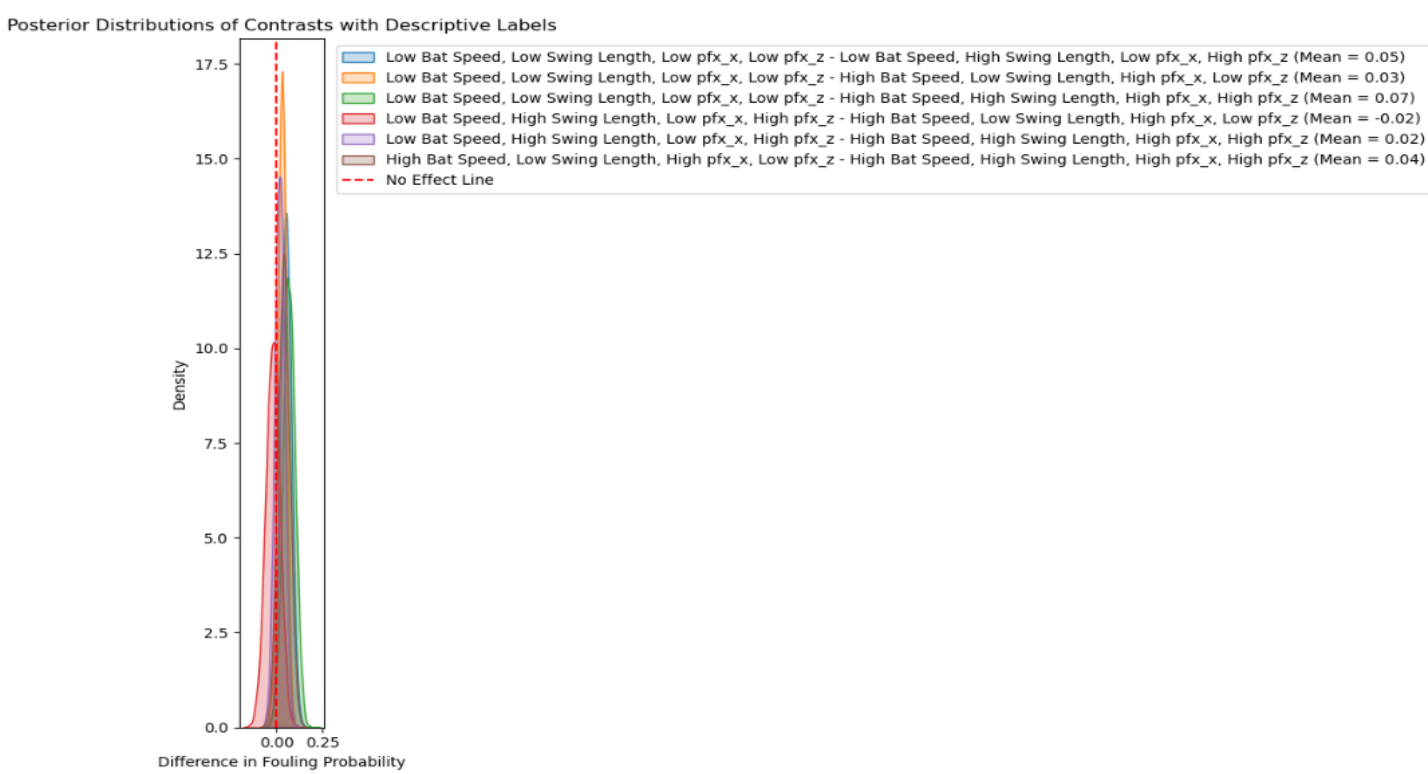
Effects of Pitch Movement:

Vertical Movement (pfx_z): Higher pfx_z consistently increases fouling probabilities across all combinations.

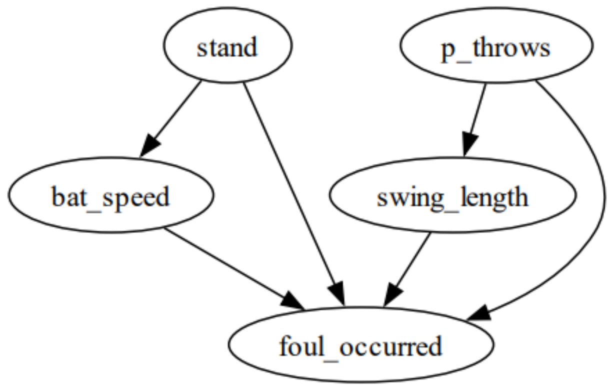
Horizontal Movement (pfx_x): Higher pfx_x reduces fouling probabilities.



Findings: Higher vertical movement increases fouling likelihood, while higher horizontal movement decreases it. Effective batter mechanics (high bat speed and swing length) reduce fouling probabilities in combination with challenging pitch movements



Q5: How do batter stance (stand) and pitcher throwing hand (p_throws) mediate the relationship between batter mechanics (bat speed, swing length) and influence likelihood of fouling off a 2-strike pitch?



Results:

Fouling Probabilities Across Conditions:

Bat Speed and Stance:

- Low Bat Speed, Left Stand: Mean = 0.43, 95% CI = [0.40, 0.47].
- High Bat Speed, Left Stand: Mean = 0.44, 95% CI = [0.40, 0.47].

Interpretation: Differences between low and high bat speed are minimal and within overlapping credible intervals. Stance (Left vs. Right) does not significantly mediate the effect of bat speed on fouling likelihood.

Swing Length and p_throws:

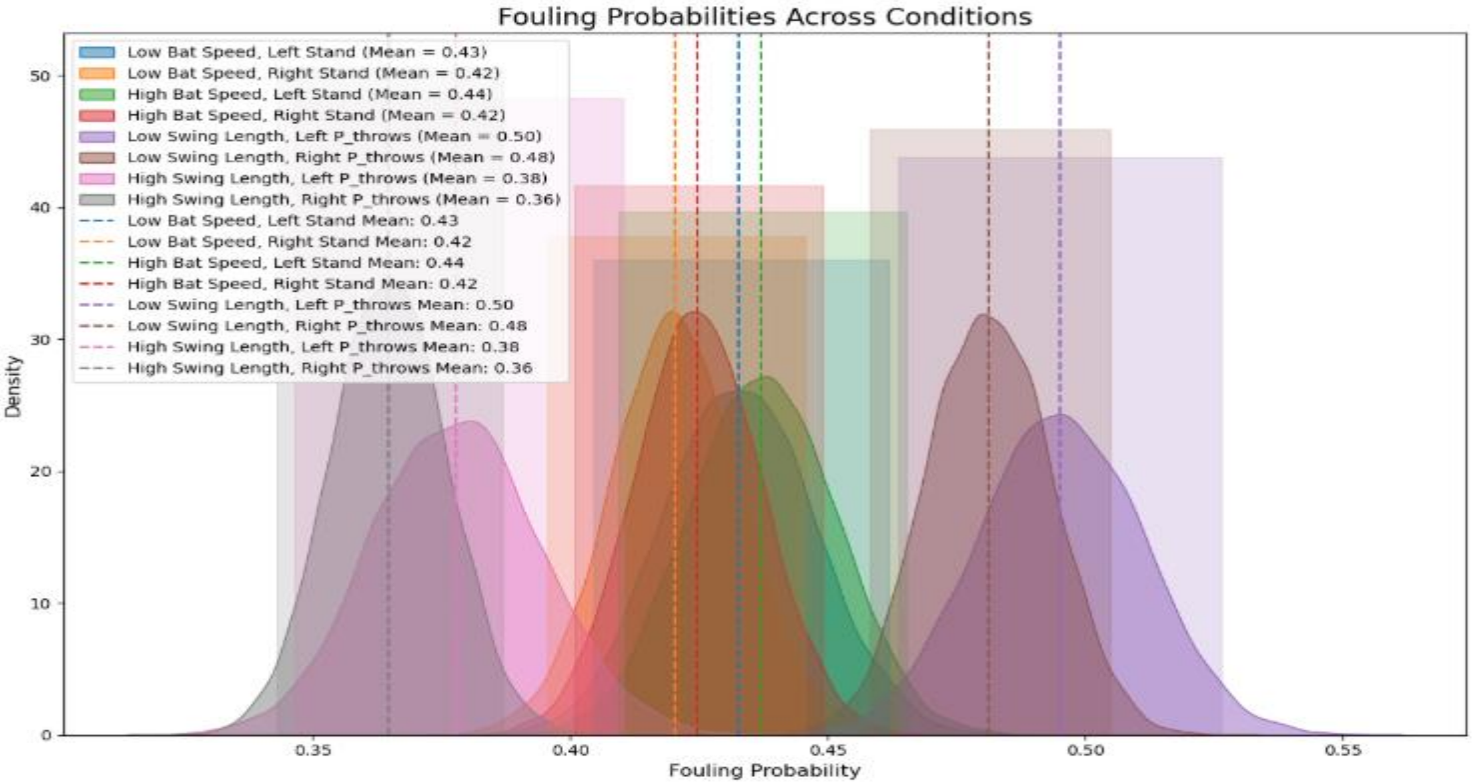
- Low Swing Length, Left p_throws: Mean = 0.50, 95% CI = [0.46, 0.54].
- High Swing Length, Right p_throws: Mean = 0.37, 95% CI = [0.35, 0.39].

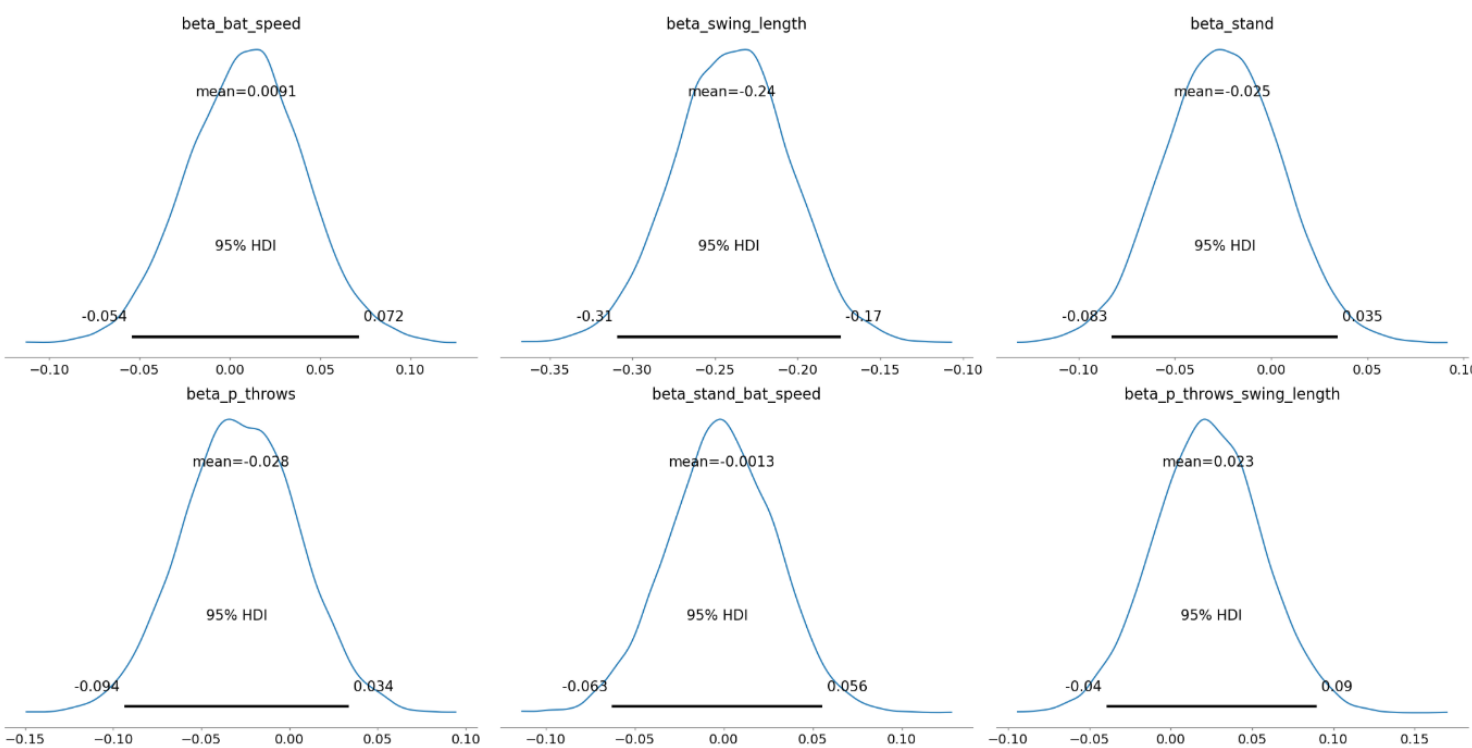
Interpretation: Swing length has a stronger effect and strongly affects fouling likelihood regardless of the pitcher's throwing hand. Low swing length increases fouling likelihood, while high swing length reduces it. The pitcher's throwing hand shows small, non-significant differences.

Contrasts Between Conditions:

Differences in fouling probabilities between specific conditions (e.g., Low Bat Speed, Left Stand vs. High Bat Speed, Left Stand) are small and statistically insignificant.

Interpretation: The mediation effects of stand and p_throws are negligible compared to the direct effects of bat speed and swing length.

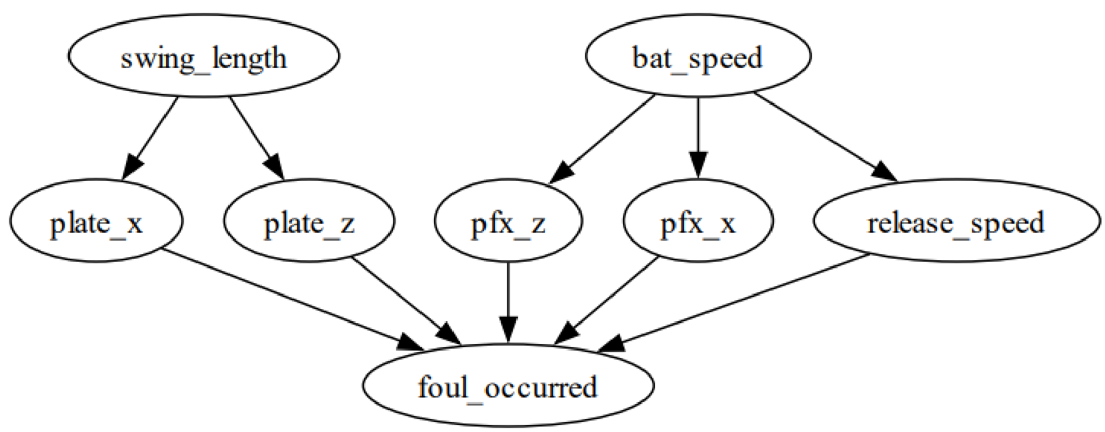




Findings: The results reflect the combined (interaction) effects of **stance, throwing hand, bat speed, and swing length,** but the interaction effect between **Bat Speed and Stance** is negligible, as differences in fouling probabilities are minimal and statistically insignificant. In **Swing Length and p_throws**, swing length has a dominant effect, overshadowing any small differences due to the pitcher's throwing hand.

- **Implication**: Training should focus on improving swing length as the primary factor influencing fouling likelihood. Batter stance and pitcher throwing hand do not substantially alter the relationship between mechanics and fouling.

Q6: Do pitch characteristics (e.g., release speed, movement, and location) mediate the relationship between batter mechanics and the influence likelihood of fouling off a 2-strike pitch?



Findings

- **Swing Length:**

  Shorter swings: Increase fouling likelihood to **59%**.

  Longer swings: Reduce fouling likelihood to **41%**.

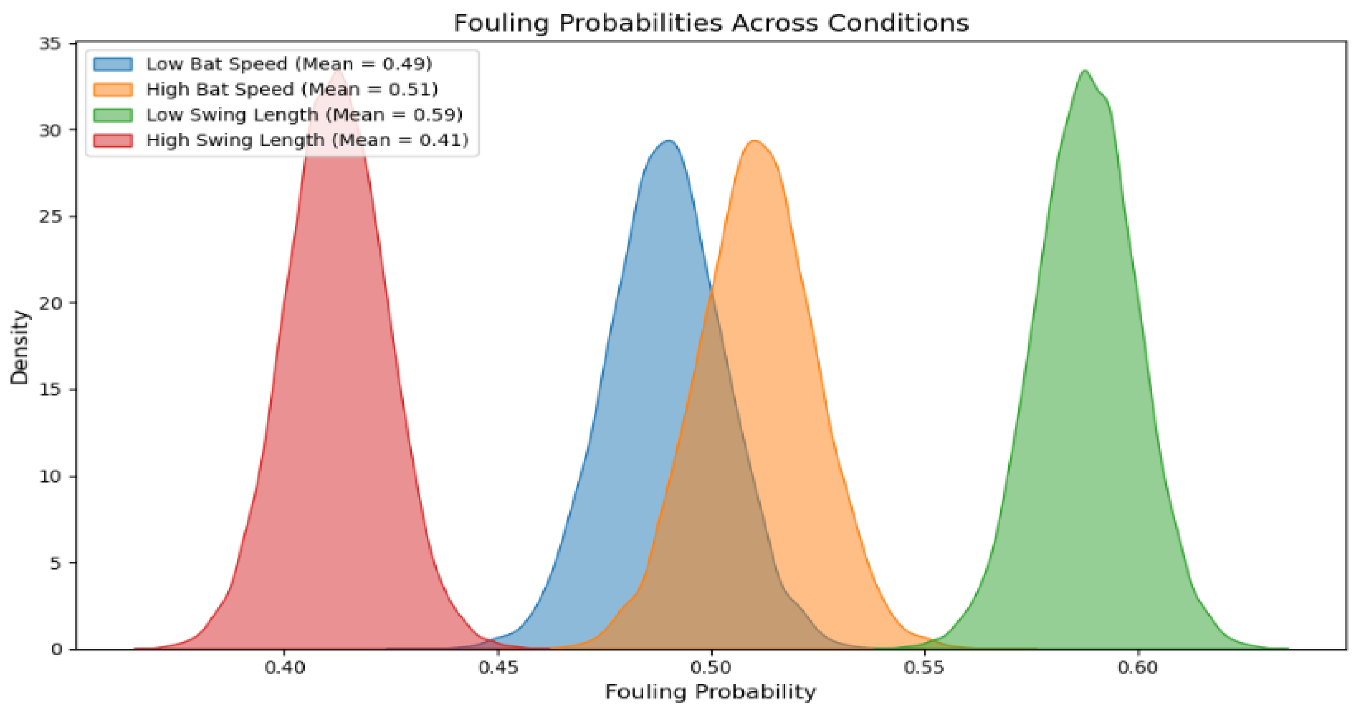
- **Vertical Pitch Location (plate_z):** A critical factor influencing fouling likelihood.
- **Bat Speed:**

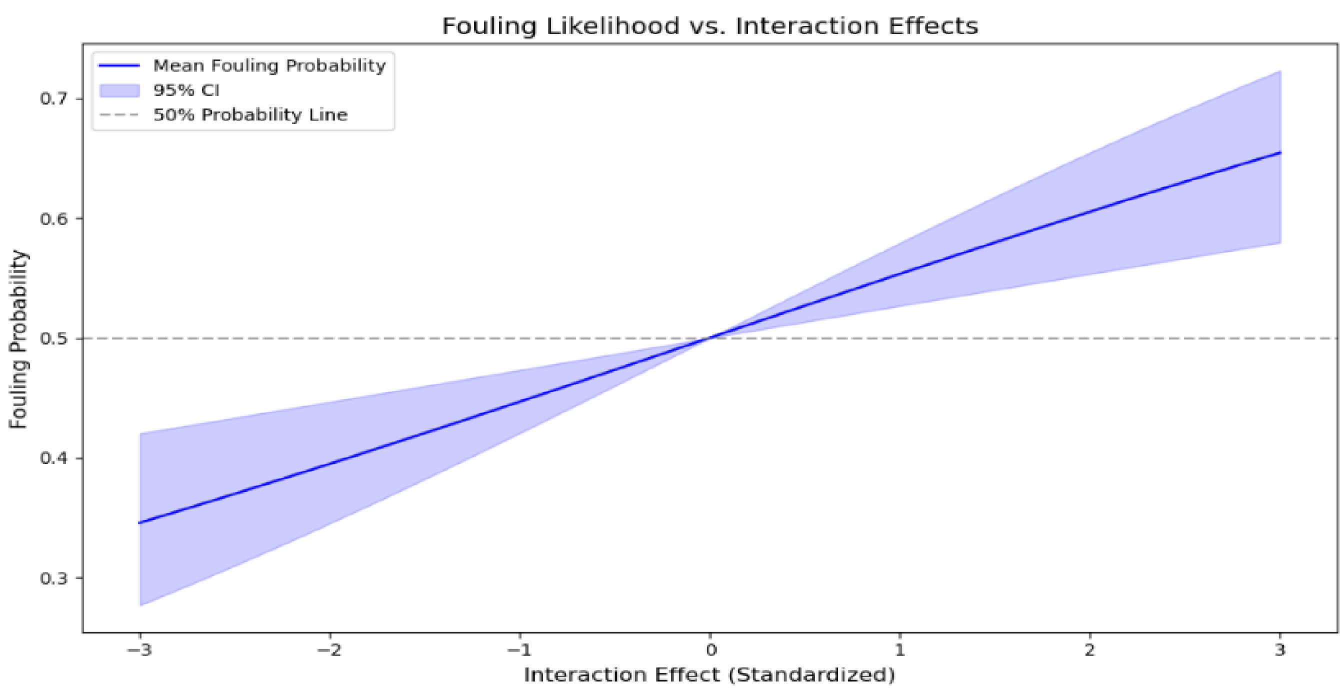
  Smaller effect compared to swing length.

  Higher speeds slightly raise fouling likelihood from **49%** to **51%**.

- **Interaction Effects:** Batter mechanics and pitch characteristics moderately influence fouling and probabilities rise slightly (from **0.35 to 0.37**) with stronger interactions.

For pitchers, targeting batters with longer swings and exploiting vertical movement or disrupting batter-pitch interaction can minimize fouling, improve strikeout chances, and control at-bats effectively.

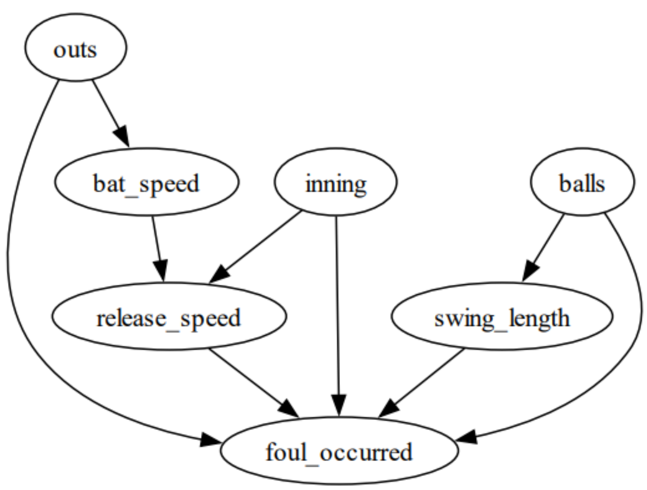




Interaction Range Summary For Foul Likelihood And Interaction Effect

	Interaction	Mean Probability	95% CI Lower	95% CI Upper
1	-3.0	0.35	0.28	0.42
2	-2.39	0.35	0.28	0.42
3	-1.79	0.35	0.28	0.42
4	-1.18	0.35	0.29	0.43
5	-0.58	0.36	0.29	0.43
6	0.03	0.36	0.3	0.43
7	0.64	0.36	0.3	0.43
8	1.24	0.37	0.3	0.43
9	1.85	0.37	0.31	0.43
10	2.45	0.37	0.31	0.43

Q7. How do game context variables (inning, outs, balls) interact with batter mechanics (bat speed, swing length) and pitch characteristics (release speed) to influence the likelihood of fouling off a 2-strike pitch?



Findings:

Interaction Effects:

- **Inning-Bat Speed**:
  - Mean Effect = **-0.014**, 95% CI = [-0.035, 0.0061].
  - **Interpretation**: Weak interaction; as innings progress, the influence of bat speed on fouling decreases.
- **Balls-Swing Length**:
  - Mean Effect = **0.073**, 95% CI = [0.017, 0.13].
  - **Interpretation**: Significant interaction; higher ball counts amplify the impact of swing length on fouling likelihood.

Covariate Effects:

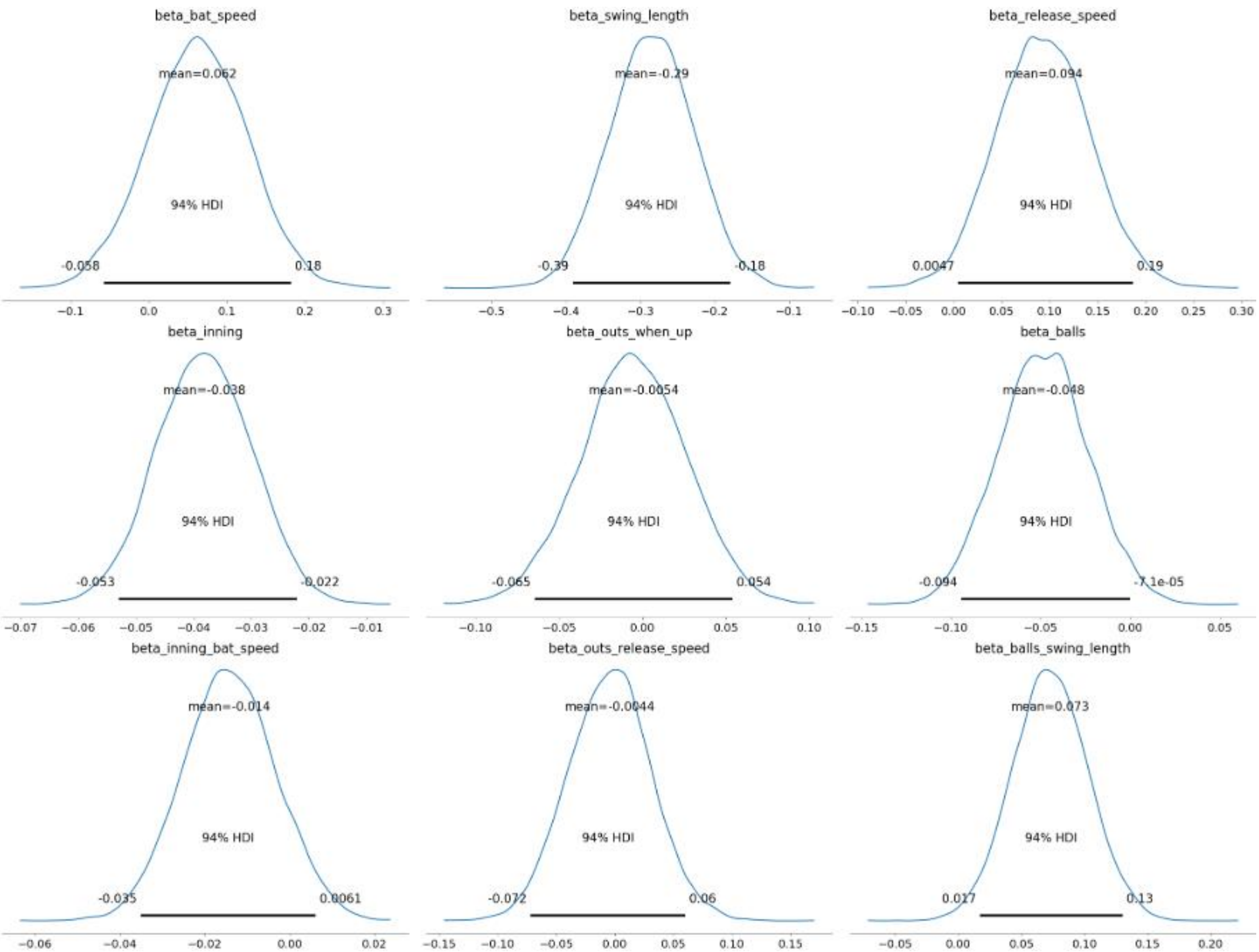
- **Inning**:
  - Mean Effect = **-0.038**, 95% CI = [-0.053, -0.022].
  - Later innings reduce fouling likelihood due to fatigue or strategic adjustments.
- **Balls**:
  - Mean Effect = **-0.048**, 95% CI = [-0.094, -0.000071].
  - Higher ball counts slightly decrease fouling likelihood as batters become more selective.
- **Outs**:
  - Mean Effect = **-0.0054**, 95% CI = [-0.065, 0.054].
  - Outs have no significant effect on fouling likelihood.

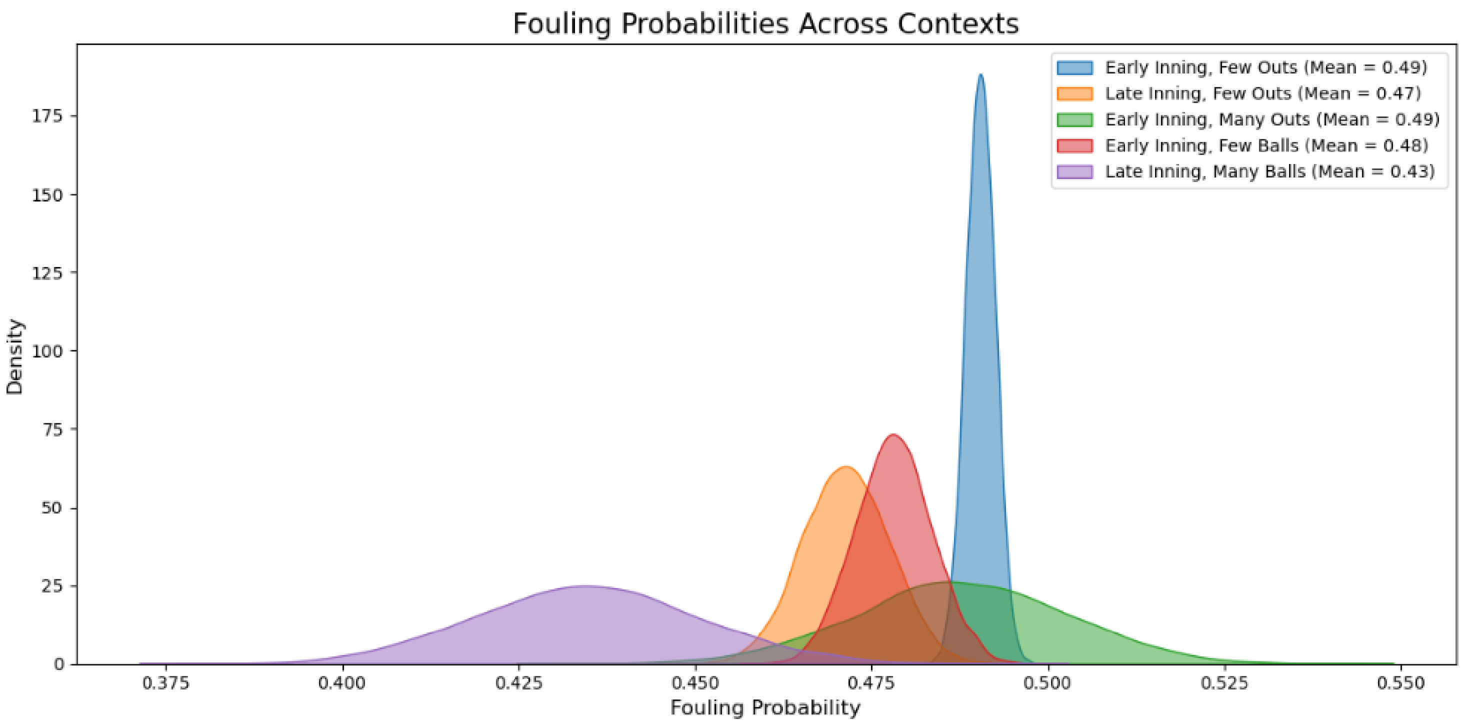
Conclusion:

- **Interaction Effects** between game context variables with batter mechanics is significant such that **Ball counts amplify swing length's impact**, while innings weakly interact with bat speed, reducing fouling likelihood.
- **Covariates**: Later innings and higher ball counts decrease fouling likelihood, while outs remain insignificant.

- **Implications**: Batters should adjust strategies in late innings and favorable counts, while pitchers can exploit early innings and low-ball counts to induce fouls. Swing length remains the strongest determinant of fouling 2-offstrikes across contexts.

Result: Game context variables interact with batter mechanics and pitch characteristics significantly with the strongest effects observed for innings and ball counts.





Final Conclusions

Batter mechanics (bat speed, swing length) are the primary determinants of the likelihood of fouling 2-strike pitches. Swing length plays a dominant role, with shorter swings significantly increasing fouling probabilities. Bat speed has a smaller but complementary influence. While pitch characteristics (release speed, vertical movement) and game context (inning, ball counts) mediate fouling likelihood, their effects are secondary to the mechanics themselves. Swing length and vertical pitch location (plate_z) stand out as key factors in all scenarios.

Recommendation: Batters should focus on shortening swings to enhance plate protection while pitchers should take advantage of early counts and aim for tough pitch locations to force fouls or strikeouts