Caution: These lecture notes are under construction. You may find parts that are incomplete.

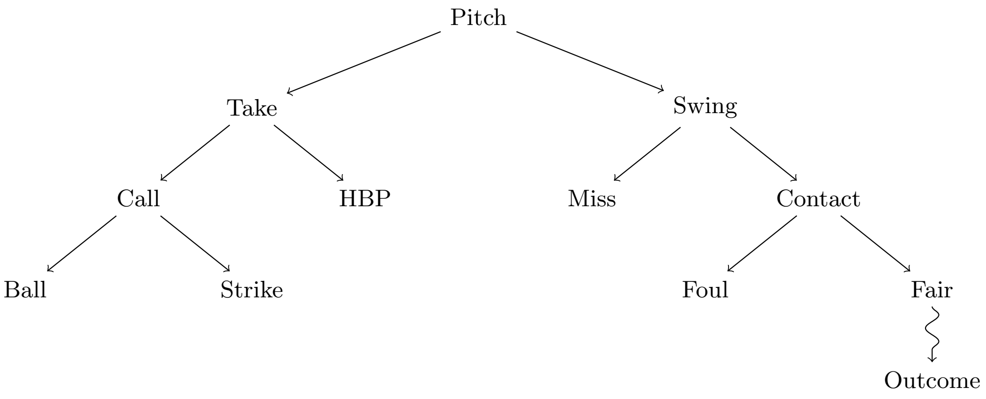
4 Introduction to Pitch-Level Analysis

So far, everything we have discussed has been on the plate appearance level. We put run values on plate appearances and evaluated players on the basis of plate appearance outcomes. Now, we transition to the more granular level of pitch-by-pitch data.

4.1 Pitch-by-Pitch Data

We observe a dataset of pitches indexed by $i \in \{1, ..., n\}$. By way of introduction, here are some of the most important variables we observe on each pitch:

- $c_i = (c_i^{\mathrm{B}}, c_i^{\mathrm{S}}) \in \{0, 1, 2, 3\} \times \{0, 1, 2\}$ is the ball-strike count before pitch i is thrown.
  - $c_i' \in \{0, 1, 2, 3, 4\} \times \{0, 1, 2, 3\}$ is the ball-strike count *after* pitch i is thrown.
- $h_i \in \{\mathrm{L, R}\}$ denotes the handedness of the batter (left or right)
- $x_i \in \mathbb{R}$ is the left/right location (in feet) of the pitch as it crosses the front of home plate.
  - $x = 0$ is the middle of the plate; $x = 17/24$ is the edge of home plate on the side of 1B.
- $z_i \in \mathbb{R}$ is the height (in feet) of the pitch as it crosses the front of home plate.
  - $z = 3.4$ and $z = 1.6$, respectively, are the top and bottom of the strike zone for the median batter.
- $o_i \in \{\text{Strikeout, ..., Home Run}\}$ is the outcome of the *plate appearance* in which pitch i occurred.
- Per the outcome tree below, $\{\text{node}\}_i \in \{0, 1\}$ is the indicator that pitch i runs through $\{\text{node}\}$.
  - Example #1: $\text{Call}_i \in \{0, 1\}$ indicates that the batter did not swing and was not hit by the pitch.
  - Example #2: $\text{Strike}_i \in \{0, 1\}$ indicates the pitch was no swing, no hit by pitch, and called strike.



This outcome tree is very important for evaluating the effectiveness of a pitch. On baseball broadcasts (and, frankly, within MLB front offices), you'll hear stats such as, "Batters are hitting .XXX against this pitcher's curveball." Aside from batting average being a poor measure of batter performance (see Chapter ???), this stat only includes pitches that terminate an at bat. That excludes roughly 80% of all pitches! And what if the pitch is only thrown in two-strike counts? We can do much, much better.

4.2 Count Value

The ball-strike count of a plate appearance is somewhat analogous to the base-out state of an inning. We can think of a plate appearance as a Markov chain transitioning from count to count. When modeling an inning as a Markov chain, our goal was to estimate base-out run expectancy. Here, we want to estimate *count value*: Given that a plate appearance is in count c, what is the expected linear weight of the outcome of the plate appearance?

To estimate count value, we could follow the same steps as for base-out run expetancy: set up the simplified Bellman equation and iteratively update the value function until convergence. However, in this case we have a property that makes the problem easier. The graph of transitions between counts is *acyclic*, meaning that once we leave a count, we cannot come back to it in the same plate appearance. Because of this property, the solution to the simplied Bellman equation is the same as the empirical average outcome linear weight from each count:

$$v(c) = \frac{\sum_{i=1}^n \mathbb{I}\{c_i = c\} \cdot \ell(o_i)}{\sum_{i=1}^n \mathbb{I}\{c_i = c\}}. \tag{1}$$

Note that equation (1) isn't *precisely* correct. Because of two-strike foul balls, some plate appearances will include multiple observations of the same pre-pitch count c. To be precise, we would only include each count once per plate appearance. We are ignoring this detail for notational convenience. The practical difference is insignificant.

An alternative to using the ball-strike count as the state of the Markov chain would be to expand our state to include ball-strike count *and* base-out state (modeling the whole inning, rather than the plate appearance). In this case, we would define the count value to be the inning run expectancy, and it would depend on the base-out state. Here, we choose to ignore base-out state when calculating count value for the same reason we prefer LW over RE24 in Chapter ???. The difference between base-out-specific performance and base-out-neutral performance is mostly noise (it takes several years of data to be equal parts signal and noise). As a shortcut, we choose to model and evaluate players on the basis of plate appearance outcomes rather than inning outcomes.

4.3 A Simple Strike Probability Model

While we can observe the full trajectory of a pitch (and we will in Chapter ???), the most important variables for estimating strike probability are plate location, count and batter handedness. We can introduce a simple model for the probability of a strike (conditioned on no swing, no HBP) as a function of these variables:

$$\mathbb{P}(\text{Strike}_i = 1 \mid \text{Call}_i = 1) = p(x_i, z_i, c_i, h_i).$$

Estimating $p(\cdot)$ is a supervised regression problem, and there are many possible approaches. For now, we ignore that task and assume that $p(\cdot)$ is known. For pitch i, we use p_i as a shorthand to denote $p(x_i, z_i, c_i, h_i)$.

4.4 Catcher Framing

We define the framing run value on pitch i to be zero if the pitch is not a called a ball or a strike by the umpire. If the pitch is called by the umpire, the framing run value is the difference between the value of the resulting count and its expected value, given the strike probability p_i. We use $c_i^+ = (c_i^{\mathrm{B}} + 1, c_i^{\mathrm{S}})$ and $c_i^- = (c_i^{\mathrm{B}}, c_i^{\mathrm{S}} + 1)$ to denote the counts resulting from adding a ball or a strike, respectively, to count c_i.

$$r_i^{\mathrm{F}} = \begin{cases} v(c_i') - \left[(1 - p_i) \cdot v(c_i^+) + p_i \cdot v(c_i^-)\right] & \text{if Call}_i = 1 \\ 0 & \text{otherwise} \end{cases}.$$

For the sake of this calculation, we define the value of any four-ball count to be $\ell(\mathrm{BB})$ (the linear weight of a walk) and the value of any three-strike count to be $\ell(\mathrm{K})$, the linear weight of a strikeout.

4.5 Discussion Questions

1. For this question, refer to the Pitch Arsenal Stats Leaderboard on Baseball Savant.
   (a) How does RV/100 relate to the material covered in this chapter?
   (b) What is the lowest batting average you can find against a below-average pitch (RV/100 < 0)?
   (c) What is the highest batting average you can find against an above-average pitch (RV/100 > 0)?
   (d) What accounts for the discrepancy between batting average against and RV/100?
2. A popular metric in online baseball analytics discourse is "called strikes plus whiffs" (CSW%), defined as the percentage of pitches that result in called strikes or swinging strikes. What are the strengths and weaknesses of CSW% relative to RV/100, for evaluating pitches?
3. How could a baseball club use pitch-level analysis to improve their decision-making?
4. For this question, refer to the Catcher Framing Leaderboard on Baseball Savant
   (a) How does Catcher Framing Runs relate to the material covered in this chapter?
   (b) Who led the league in Catcher Framing Runs this year?
   (c) Who lagged the league in Catcher Framing Runs this year?
   (d) Is the difference in Catcher Framing Runs between these two players bigger or smaller than the difference between their Batting Run Value? (see Batting Run Value Leaderboard)
5. If a pitch with low strike probability ends up being called a strike, to what extent is the outcome attributable to the catcher, and to what extent is the outcome attributable to the pitcher?
6. If MLB adopted fully automated ball-strike calls, how would this change the catching position?