# Hierarchical Stochastic Modeling of Baseball Hit Outcomes

Evan Goforth

August 8, 2025

**Abstract**

We propose a rigorous hierarchical stochastic model for a batter's hit outcomes in a single game against a starting pitcher. The framework treats each game as a nested composition of random variables and stochastic processes: game-level variables (such as innings pitched and batters faced) define the context for a sequence of at-bats, each at-bat unfolds as a finite-horizon Markov process over pitch-by-pitch states, and each plate appearance terminates in a context-dependent Bernoulli trial representing a hit or out. This path-dependent modeling approach captures known count-based effects on hitting outcomes and integrates modern metrics like expected batting average (xBA) to condition hit probabilities on the quality and context of the final pitch. We ground all modeling assumptions in established baseball analytics literature – from early Markov chain models of at-bats and innings to empirical Bayes shrinkage of batting averages and recent simulation-based approaches. The mathematical formulation of the model is presented in detail, and we discuss how the structure naturally leads to compound distributions (e.g., Poisson-binomial) for total hits. We demonstrate that this framework unifies and extends previous stochastic baseball models, providing a blueprint for simulation, inference, and strategic decision analysis in player performance evaluation.

## 1 Introduction

Modeling offensive performance in baseball has long been a subject of interest for statisticians and analysts. A single batter's performance in a game can be viewed as a compound outcome driven by many latent factors – the pitcher's longevity in the game, the opportunities (at-bats) the batter receives, and the pitch-by-pitch dynamics of each encounter. Traditional metrics like batting average or slugging percentage summarize outcomes but do not explicitly account for the process by which those outcomes arise. In contrast, a *stochastic process model* can capture the sequential, context-dependent nature of each plate appearance. In this paper, we develop a hierarchical stochastic framework for a batter's hits in a game, explicitly modeling the nested structure of game context, at-bat sequences, and terminal outcomes.

Our approach is motivated by and builds upon several strands of prior research in baseball statistics. At the *plate appearance level*, researchers such as Katz (1986) and Hopkins and Magel (2008) pioneered Markov chain models of the count (balls and strikes) to analyze how outcome probabilities shift with each pitch. Their findings demonstrated that the state of the count dramatically affects a batter's success – for example, a hitter's on-base average or slugging percentage is much higher in favorable counts (e.g., 3–1) than with two strikes. This motivates treating an at-bat not as a fixed-probability event but as a sequence of states that influence the final outcome. At the *team and inning level*, early work by Cover and Keilers (1977) used a Markov model of the 24 base-out states in an inning to compute run expectancies and introduce an "Offensive Earned-Run Average" metric. Bukiet et al. (1997) extended this to optimize batting lineups and predict team runs and wins using a Markov chain for innings and games. These studies established the power of nested stochastic models (at-bats within innings, within games) for baseball outcomes.

We also draw inspiration from psychological and decision-making models of the batter–pitcher duel. Gray (2002), for example, developed a cognitive Markov model of an at-bat ("Markov at the Bat") that showed batters continuously update their expectations of pitch type based on the count and prior pitches. Such findings support the idea that pitch sequences can be modeled as stochastic processes influenced by both state (count) and strategic adjustments. Our framework incorporates pitch-level detail (pitch type and location probabilities) to acknowledge these dynamics. Furthermore, our model's hierarchical structure aligns with *Bayesian and compound modeling* approaches in the literature. Jensen et al. (2009) introduced a hierarchical Bayesian model for batting performance that treats each player's true hitting talent as a latent parameter evolving over time. This reflects the view that observed hits are random draws from an underlying probability unique to each batter and season. Even earlier, Efron and Morris (1975) demonstrated the benefits of empirical Bayes shrinkage by treating a handful of at-bats as a sample from a latent batting average, dramatically improving estimation accuracy by pooling information across players. Such approaches reinforce the notion that hit outcomes can be viewed as a sum of Bernoulli trials with an uncertain success probability – an idea we integrate by allowing the hit probability itself to depend on context and be treated as a random variable. Modern analytics have also introduced the concept of *expected outcomes* for batted balls: for instance, Statcast's expected batting average (xBA) assigns each batted ball a probability of becoming a hit based on its exit velocity, launch angle, and other factors. We leverage this idea by making the final hit outcome of an at-bat a Bernoulli trial with success probability informed by such predictive models of batted-ball success.

In summary, our contribution is a comprehensive probabilistic framework that synthesizes these lines of work into a single model. We treat a batter's total hits $Y$ in a game as a compound random variable, composed of a random number of at-bats with random outcomes, each outcome conditioned on the path of the at-bat. By nesting a pitch-by-pitch Markov process within a game-level simulation of innings and opportunities, we create a model that can simulate realistic at-bat sequences, estimate outcome probabilities under various scenarios, and serve as a foundation for inferential or decision-support tools (such as evaluating in-game strategies or betting probabilities).

## 2 Related Work

**Markov chains for at-bats.** The idea of modeling an at-bat as a sequence of pitches with probabilistic transitions dates back to Katz (1986). Hopkins and Magel (2008) investigated batting productivity conditional on the count, specifically slugging percentage in different count situations. Both treated each at-bat as a discrete stochastic process that terminates in a Bernoulli-type outcome. Practical demonstrations include Forman (2017).

**Half-inning and game-level models.** Cover and Keilers (1977) introduced a Markov chain for the 24 base/out states of an inning. Bukiet et al. (1997) extended to full games and seasons, optimizing lineups and predicting team performance. These foundational works motivate our game-layer variables (IP, BF, AB).

**Cognitive and decision models.** Gray (2002) modeled batter cognition as a Markov process. Count and sequence effects justify inhomogeneous transition probabilities in our at-bat layer. Decision-theoretic variants (MDP/RL) build on the same state structure.

**Hierarchical/Bayesian performance models.** Jensen et al. (2009) used hierarchical Bayes for hitters' talent trajectories; Albert (2003) used a beta-binomial mixture for batting average; Efron and Morris (1975) provided empirical Bayes shrinkage foundations.

## 3 Methods: Model Structure and Dynamics

### 3.1 Game-level random variables

Let $Y$ be the total number of hits against the starting pitcher. Decompose

$$Y = \sum_{i=1}^{AB} I_i, \tag{1}$$

where $I_i \in \{0, 1\}$ is the hit indicator in the $i$-th at-bat, and $AB$ is a random number of at-bats determined by pitcher longevity and lineup position.

Let $IP$ be innings pitched by the starter. Define a regression-based expectation $\mu_{IP} = f_{IP}(\text{opponent wOBA})$, and model

$$IP \sim \mathcal{N}(\mu_{IP}, \sigma_{IP}^2). \tag{2}$$

Given $IP$, model batters faced

$$BF \sim \text{Poisson}(\lambda_{BF}), \qquad \lambda_{BF} = f_{BF}(IP). \tag{3}$$

Given $BF$ and lineup slot $L \in \{1, \ldots, 9\}$, model the batter's at-bats

$$AB \sim \text{Binomial}\big(BF,\, p_L\big), \tag{4}$$

where $p_L \approx 1/9$ is the slot share, estimated empirically.

### 3.2 At-bat as a stochastic process

Each at-bat evolves as a discrete-time process $\{X_t\}_{t=0}^{\tau}$ with

$$X_t = (C_t, P_t, Z_t, O_t), \tag{5}$$

where $C_t$ is the count, $P_t$ the pitch type, $Z_t$ the zone, $O_t$ the pitch outcome. The absorbing outcomes are $T = \{\text{BB}, \text{K}, \text{HBP}, \text{BIP}\}$. The initial state is $C_0 = (0, 0)$.

Conditional kernels govern transitions:

$$\mathbb{P}(P_t = p \mid C_{t-1} = c, H, A) = \theta_p(c, H, A), \tag{6}$$

$$\mathbb{P}(Z_t = z \mid P_t = p, C_{t-1} = c, H, A) = \phi_z(p, c, H, A), \tag{7}$$

$$\mathbb{P}(O_t = o \mid P_t = p, Z_t = z, C_{t-1} = c) = \psi_o(p, z, c). \tag{8}$$

The count updates deterministically via $C_t = T(C_{t-1}, O_t)$ (baseball rules for balls/strikes/fouls/HBP/BIP). Define the stopping time $\tau = \inf\{t \geq 1 : C_t \in T\}$.

### 3.3 Terminal hit probability

Let $f : P \times Z \times C \to [0, 1]$ map terminal pitch context to hit probability on contact (xBA-like). Define the at-bat hit indicator

$$I \;=\; \begin{cases} \text{Bernoulli}\big(f(P_\tau, Z_\tau, C_\tau)\big), & \text{if } C_\tau = \text{BIP}, \\ 0, & \text{if } C_\tau \in \{\text{K}, \text{BB}, \text{HBP}\}. \end{cases} \tag{9}$$

### 3.4 Distribution of total hits

With $Y = \sum_{i=1}^{AB} I_i$, we have a random sum of (generally) non-identical Bernoulli trials. If $p_i = f(P_{\tau,i}, Z_{\tau,i}, C_{\tau,i})$, then

$$\mathbb{E}[Y \mid AB = n] \;=\; \sum_{i=1}^{n} \mathbb{E}[p_i], \qquad \mathbb{E}[Y] \;=\; \mathbb{E}[AB] \cdot \mathbb{E}[p_i]. \tag{10}$$

Variance inflates relative to a fixed-$p$ binomial due to variability in $AB$ and $\{p_i\}$, yielding Poisson-binomial / beta-binomial behavior.

## 4 Discussion: Extensions and Applications

- **Pitch sequence modeling.** Extend to higher-order Markov or RNN-based $\theta_p(\cdot)$ to capture sequencing.
- **Spatial dependence.** Add dependence in $Z_t$ (e.g., first-order transitions over zones) for location mixing.
- **Hierarchical estimation.** Pool parameters across players with partial pooling; estimate $f$ from Statcast-like batted-ball data.
- **Compound distribution insights.** Under Poisson $AB$ and constant $p$, $Y$ approximates Poisson($\lambda p$); heterogeneous $p_i$ induces over-dispersion.
- **Simulation uses.** Monte Carlo for prop probabilities (0/1/2+ hits), in-count conditional projections, and strategy evaluation.

## 5 Conclusion

We presented a hierarchical stochastic process model for a baseball batter's hit outcomes against a starting pitcher in a game. The game-level component (IP, BF, AB) connects opponent strength to opportunity; the at-bat component encodes count/pitch context and a path-dependent terminal payoff via $f(P_\tau, Z_\tau, C_\tau)$. The resulting compound distribution for $Y$ accommodates realistic dispersion and aligns with prior Markov, Bayesian, and simulation work while providing a clear generative blueprint.

## References

## References

Efron, B., & Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350), 311–319.

Albert, J. (2003). A batting average mixture model for season and career performance. *Journal of Quantitative Analysis in Sports*, 1(1), 1–17.

Cover, T. M., & Keilers, C. W. (1977). An offensive earned-run average for baseball. *Operations Research*, 25(5), 729–740.

Bukiet, B., Harold, E. R., & Palacios, J. L. (1997). A Markov chain approach to baseball. *Operations Research*, 45(1), 14–23.

Katz, S. M. (1986). Study of "the count" yields fascinating data. *Baseball Research Journal*, 15, 75–79.

Hopkins, T., & Magel, R. C. (2008). Slugging percentage in differing baseball counts. *Journal of Quantitative Analysis in Sports*, 4(4), Article 2.

Gray, R. (2002). "Markov at the Bat": A model of cognitive processing in baseball batters. *Psychological Science*, 13(6), 542–547.

Jensen, S. T., McShane, B. B., & Wyner, A. J. (2009). Hierarchical Bayesian modeling of hitting performance in baseball. *Bayesian Analysis*, 4(2), 219–238.

Forman, C. (2017). *A Markov approach to modeling baseball at-bats and evaluating pitcher decision-making and performance*. Undergraduate thesis, Harvard University.