Monte Carlo Simulation of OPS(On-Base Plus Slugging) Using the Multinomial Distribution

Andrew Shin

2024-10-19

Abstract

In modern baseball analytics, sabermetrics plays a crucial role in objectively measuring player performance through advanced statistical metrics. One key metric, On-Base Plus Slugging (OPS), combines On-Base Percentage (OBP) and Slugging Percentage (SLG) to provide a comprehensive view of a player's offensive capabilities, capturing both their ability to reach base and generate extra-base hits. This study approaches OPS analysis from a probabilistic perspective, modeling both OBP and SLG as outcomes of a multinomial distribution. Each at-bat and plate appearance is considered a trial with multiple possible results, such as singles, doubles, or outs for SLG, and hits, walks, hit by pitch ball, or out for OBP. Using Monte Carlo simulations based on these probabilities, we predict future player performance by estimating OBP and SLG, which are then combined to approximate OPS. This method allows for the probabilistic modeling of offensive outcomes, offering performance evaluation into a player's expected contribution over the course of a season.

Introduction

In modern baseball analytics, sabermetrics plays a crucial role in objectively quantifying player performance using advanced statistical measurements. One of the most widely recognized metrics in recent years is OPS (On-Base Plus Slugging), which combines a player's on-base percentage (OBP) and slugging percentage (SLG). OPS is highly regarded because it integrates two essential components of offensive performance into a single statistic, capturing a player's ability both to reach base and to generate extra-base hits.

In Major League Baseball, an OPS of .800 or higher puts a player in the upper echelon of hitters, and the league leader in OPS typically scores near or above 1.000. Players with high OPS values are usually the most valuable offensive contributors in the league.

When analyzing OPS from a probabilistic perspective, I came up with the idea that both OBP and SLG can be modeled as a multinomial distribution. In OBP, each plate appearance is considered a trial with four possible outcomes: hit, walk, hit by pitch ball, or out, $OBP = \frac{\#Hit+\#Walks+\#HBP}{(\#At-Bats+\#Walks+\#HBP+\#Sacfly)=\#plate\ apps.}$. Similarly, slugging—calculated as SLG $= \frac{1\times 1B+2\times 2B+3\times 3B+4\times HR}{\text{At-Bat}}$ —has multiple possible outcomes (singles, doubles, triples, and home runs). Therefore, applying a multinomial distribution is a proper approach to model the various types of hits a player might achieve over the course of a season, reflecting the likelihood of different hit categories in a sequence of independent at-bats.

A Monte Carlo simulation based on players' stats from March to July allows for player evaluation and comparison through probabilistic analyses, such as CDF plots, density plots, and box plots. These visualizations provide insights into the variability and distribution of player performance, highlighting how likely certain outcomes are across a range of scenarios.

Methods

Data Collection

The data for this analysis was sourced from Baseball Reference, a website that provides up-to-date player statistics from Major League Baseball. I randomly selected the stats of 10 players who stood out to me during the 2024 MLB season. In baseball, the regular season is typically divided into two halves: the first half runs from March to July (All-Star Game), and the second half from August to September. For this analysis, I have split the season accordingly and will use the first-half statistics in a Monte Carlo simulation to predict player performance in the second half of the season.

Multinomial Distribution on On-Base Percentage + Slugging Percentage (OPS)

OBP is calculated as $\frac{\#Hits+\#Walk+\#HBP}{\#Plate\ Appearance}$ in which we have four different possible outcomes with distinct probabilities. Similarly, Slugging percentage is $\frac{\#Singles+\#Doubles+\#Triples+\#Homeruns}{\#At-Bat}$ which have five possible outcomes, including number of other outcomes. Since the multinomial distribution is a generalization of the binomial distribution that allows for more than two possible outcomes, we can model the number of occurrences of different outcomes in a series of independent trials (in this case, the number of plate appearance and at-bats).

$$P(X_1 = x_1, \cdots, X_k = x_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

, where:

- n is the number of trials (in this case, the number of plate appearances and at-bats)
- $k_1, \ldots, k_i$ are the counts of each possible outcomes ($i = 1, \ldots, 4$ in OBP and $i = 1, \ldots, 5$ in SLG).
- $p_1, \ldots, p_i$ are the probabilities of each outcomes
- $k_1 + \cdots + k_i = n$ (the total number of trials in the sum of all outcomes)

Based on the stats in the first half of this season, I calculated the probability of each outcome in OBP and will use this for my analysis.

```
kable(obp_probs)
```

Player	H/PA	BB/PA	HBP/PA	OUT/PA
Aaron Judge	0.254	0.179	0.015	0.552
Shohei Ohtani	0.266	0.129	0.006	0.599
Juan Soto	0.249	0.183	0.004	0.564
Bobby Witt Jr	0.317	0.069	0.010	0.604
Kyle Schwarber	0.198	0.173	0.009	0.620
Seiya Suzuki	0.240	0.087	0.016	0.657
Francisco Lindor	0.226	0.083	0.022	0.669
Anthony Santender	0.224	0.075	0.016	0.685
Jose Ramizez	0.253	0.073	0.002	0.672
Marcell Ozuna	0.265	0.106	0.007	0.622

```r
kable(slg_probs)
```

Player	Singles/AB	Doubles/AB	Triples/AB	Homerun/AB	Other_outcomes/AB
Aaron Judge	0.148	0.065	0.003	0.101	0.683
Shohei Ohtani	0.155	0.068	0.010	0.077	0.690
Juan Soto	0.174	0.056	0.010	0.069	0.691
Bobby Witt Jr	0.212	0.069	0.023	0.044	0.652
Kyle Schwarber	0.155	0.030	0.000	0.061	0.754
Seiya Suzuki	0.160	0.055	0.009	0.046	0.730
Francisco Lindor	0.142	0.062	0.000	0.050	0.746
Anthony Santender	0.123	0.043	0.005	0.078	0.751
Jose Ramizez	0.152	0.056	0.002	0.065	0.725
Marcell Ozuna	0.173	0.050	0.000	0.078	0.699

Result

Monte Carlo Simulation on On-Base Percentage and SLG

The Monte Carlo simulation repeatedly simulates a fixed number of plate appearances for OBP and at-bats for SLG using the multinomial distribution. In each simulation, we count the occurrences of each event type and compute OBP and SLG for that trial. We set the number of iterations to n = 10,000, calculating OBP and SLG for each simulation based on the total plate appearances and at-bats.

```r
kable(monte_carlo_OBP)
```

Player	exp_OBP	actual_OBP
Aaron Judge	0.448	0.482
Shohei Ohtani	0.401	0.393
Juan Soto	0.436	0.383
Bobby Witt Jr	0.396	0.375
Kyle Schwarber	0.380	0.360
Seiya Suzuki	0.343	0.404
Francisco Lindor	0.332	0.376
Anthony Santender	0.315	0.296
Jose Ramizez	0.328	0.345
Marcell Ozuna	0.378	0.379

The average values from the simulation provide an estimate of the player's expected performance. In addition to the On-Base Percentage (OBP) estimated through the Monte Carlo simulation, we also have the actual OBP values for players from August to September. By comparing these values, we can observe whether players' ability to get on base has improved or declined during this period.

```r
kable(monte_carlo_OPS)
```

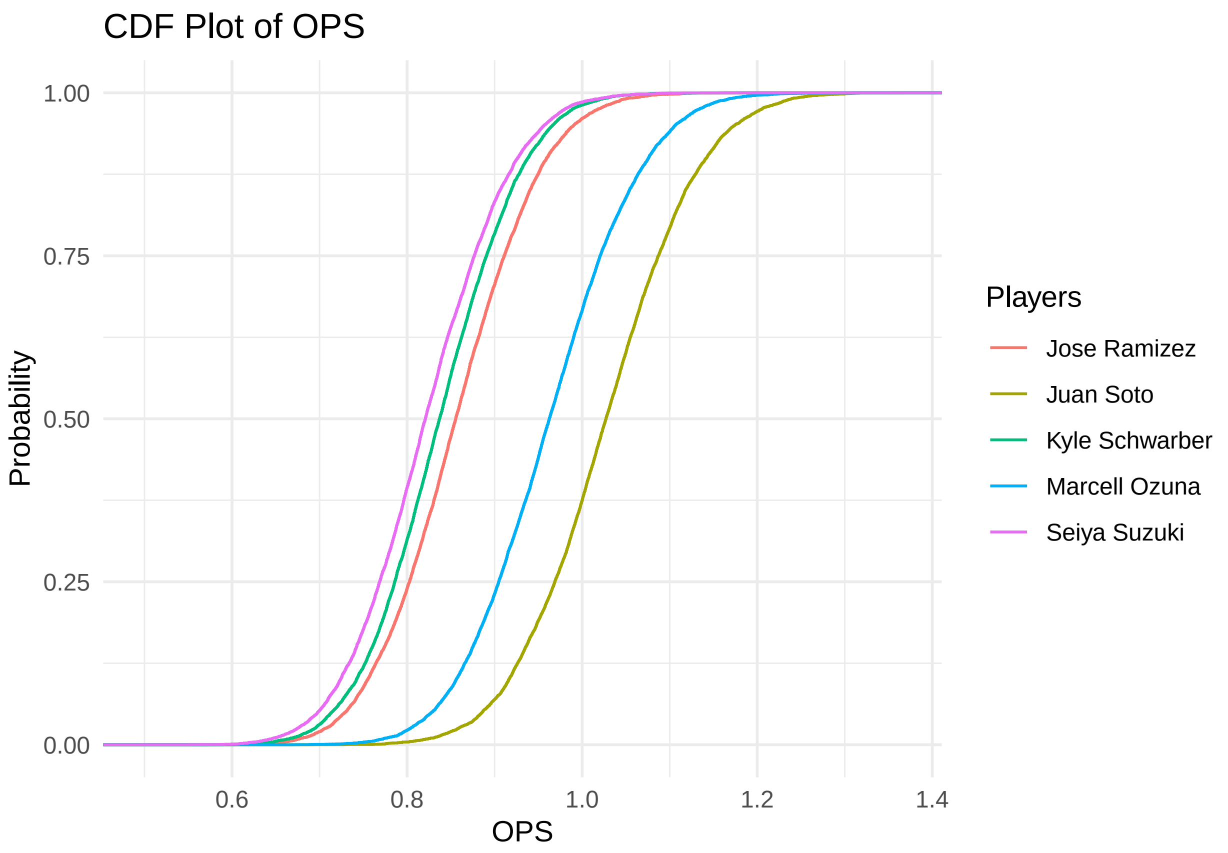
Player	exp_OPS	actual_OPS
Aaron Judge	1.1378434	1.2103237

Player	exp_OPS	actual_OPS
Shohei Ohtani	1.0310468	1.0731802
Juan Soto	1.0284470	0.9019189
Bobby Witt Jr	0.9922177	0.9464286
Kyle Schwarber	0.8390976	0.8908057
Seiya Suzuki	0.8233608	0.8878710
Francisco Lindor	0.7986934	0.9561105
Anthony Santender	0.8508418	0.7449796
Jose Ramizez	0.8571641	0.8935437
Marcell Ozuna	0.9641329	0.8524300

Cumulative Distribution Function (CDF) Plot

CDF plots allow us to make probabilistic statements about our simulated variable. We can use it to answer questions like "What is the probability of the OPS being above or below a certain value?" or "What is the range of likely outcomes?".

```
OPS_cdf_plot
```



As shown in the plot, the NY Yankees Aaron Judge's median OPS (where $y = 0.5$) is $P(X_{Judge} \leq 0.5) = 1.143$, meaning he is likely to achieve an OPS around 1.143. This median OPS is higher than both the median and mean OPS of 10 other players, which is around 0.933. His actual OPS in August and September 2024 was 1.210, which corresponds to approximately $P(1.18 \leq X_{\text{Judge}} \leq 1.23) = 0.1486$, indicating a 14.86% chance of falling within this range. Notably, his OPS in the second half of the season exceeded the expected value predicted by the Monte Carlo simulation. This improvement in OPS highlights his exceptional performance as a player.

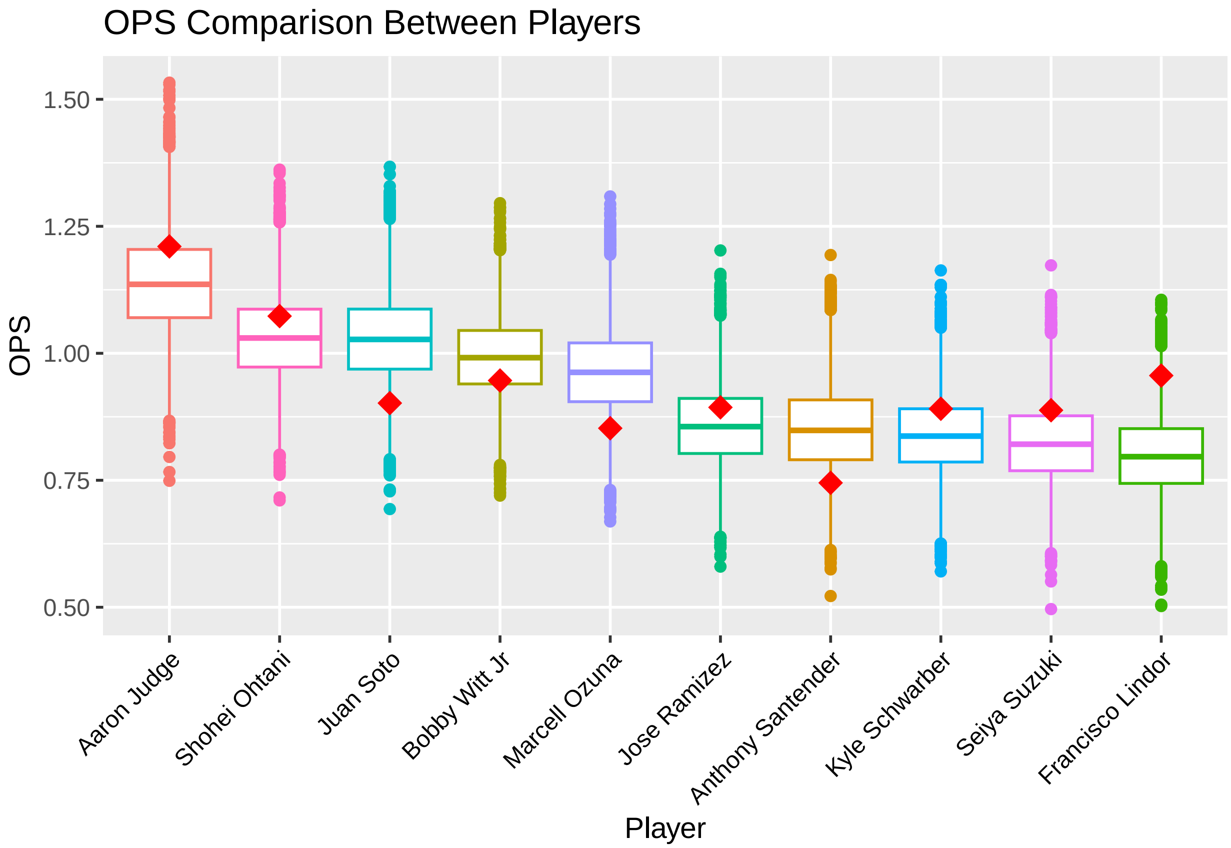
Likewise, the CDF is particularly useful because it allows us to visualize the probability that a variable falls within a specific range. It helps quantify risks and ask probabilistic questions based on the simulation results.

OPS Comparison Among Players

Comparing the Monte Carlo simulated OPS with its actual value offers valuable insights into player performance. If the actual OPS is higher than the expected value, it indicates the player is performing above expectations, and if it is lower, the player is under-performing relative to the projection. Due to the length of the MLB season, it is challenging for players to maintain consistent performance throughout. Most players experience fatigue as the season progresses, particularly toward the end. Despite this, a player's ability to sustain their performance until the season's conclusion is a strong indicator of their consistency and resilience.

As shown in the plot, each boxplot represents the simulated OPS for individual players, with the red dots indicating their actual OPS values during the second half of the season. By observing the positions of these red dots relative to the boxplot, we can assess how players performed compared to their simulations. The middle line within each box represents the simulated median, while the lower and upper edges correspond to the 25th and 75th percentiles, respectively.

```
pl_ops
```



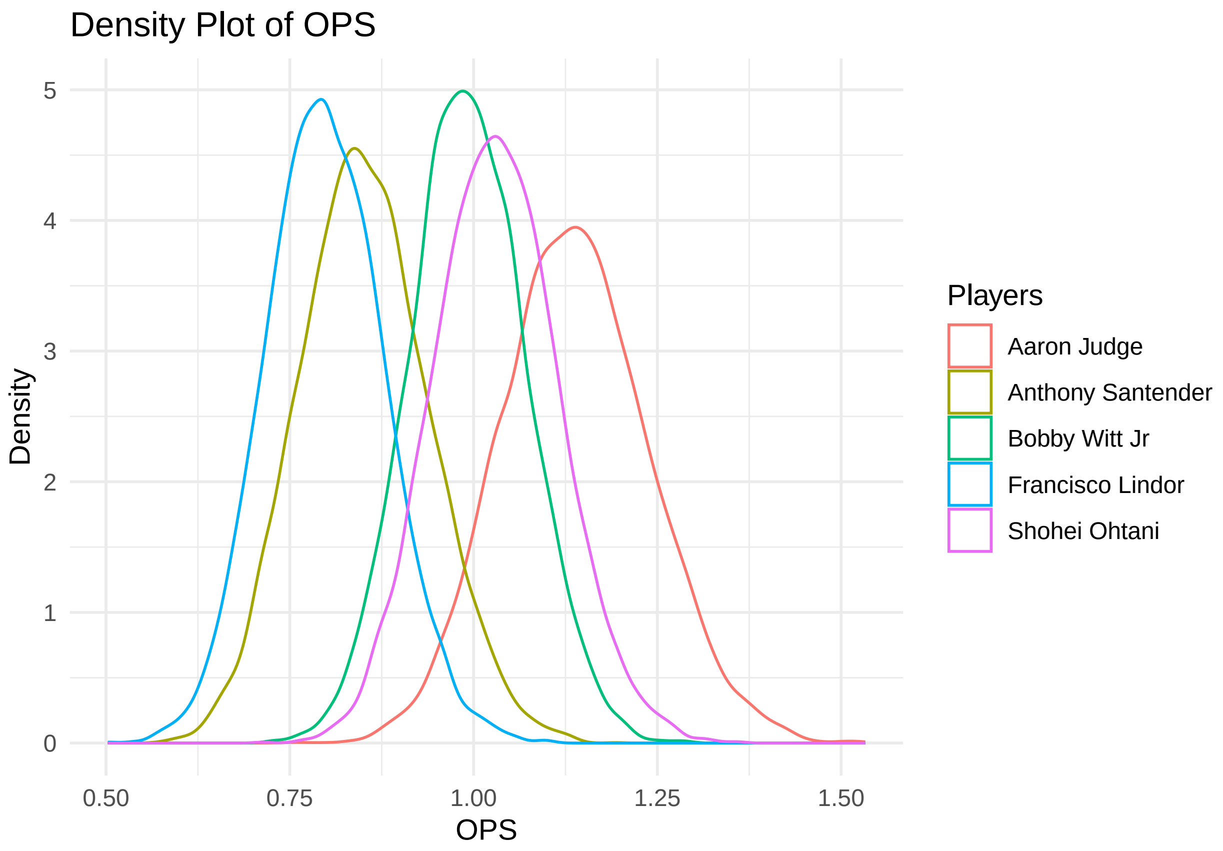
Francisco Lindor of the New York Mets showed remarkable improvement in his OPS, recording a 0.956 OPS from August to September. This represents a 20% increase compared to the median and ranks as the third-highest among the ten players analyzed. Lindor's exceptional OPS performance has carried into the MLB postseason, where his OPS currently stands at 0.893 after Game 4 of the NLCS against the Dodgers.

On the other hand, players like Soto, Ozuna, and Santander showed a significant decline in OPS. Based on this data, we can assess them as having under=performed compared to their expected performance. Moreover, Shohei Ohtani who made to the historic 50/50 club (50 Homeruns and 50 Stolen Bases) and a homerun leader, Aaron Judge, both performed better than expected from Monte-Carlo simulation.

Kernel Density Estimate (KDE) after Monte Carlo Simulation

In the context of Monte Carlo simulations, KDE (Kernel Density Estimate) plots in R are highly useful for visualizing the distribution of simulated OPS values. This allows us to estimate and visualize the probability density function of the simulated results, providing insight into where the most likely outcomes are and the overall spread of the results.

```
OPS_den_plot
```



As shown in the plot, the top two players are Aaron Judge and Shohei Ohtani, both of whom are strong contenders for the 2024 MLB American and National League MVP awards. This visualization not only highlights their exceptional performance but also provides a clearer comparison with other players.

Discussion & Conclusion

The probabilistic approach to OPS analysis provides a unique perspective on player performance by modeling offensive outcomes using a multinomial distribution. Through Monte Carlo simulations, we can estimate future performance by approximating On-Base Percentage (OBP) and Slugging Percentage (SLG), which together offer a comprehensive view of a player's offensive capability (OPS). This method has several important implications for player evaluation and decision-making in sabermetrics.

The use of multinomial distribution to model OBP and SLG offers a more nuanced way of understanding player performance compared to traditional methods. Each at-bat is treated as a probabilistic event with multiple possible outcomes, such as singles, doubles, or walks, reflecting the inherent randomness in baseball. This allows us to account for the variability in a player's performance over a large number of at-bats. By running Monte Carlo simulations based on these probabilities, we can generate a distribution of likely OPS values, offering a comprehensive understanding of a player's offensive contributions to the teams.

Nonetheless, this approach assumes independence between at-bats and doesn't account for external factors like pitcher quality or game situations. Future models could integrate these factors for more accuracy. The model also focuses on offensive performance, disregarding the fact that defensive contributions and base-running could be included for a more comprehensive evaluation.

By applying Monte Carlo simulations to OPS analysis, despite limitations, this study offers a detailed, probabilistic framework for evaluating player performance. This method enhances the ability to estimate better offensive contributions and assess players' performances, with potential for wide-ranging applications in baseball sabermetrics.