A PGA Golf Simulation Framework Using Statistical Distributions and Decision Modeling

Shusuke Hashimoto

Indiana University Bloomington

Abstract

This paper introduces a simulation framework for PGA golf tournaments using statistical modeling and binary decision-making algorithms. Leveraging player statistics from the PGA Tour's official website, the simulation employs lognormal, truncated normal, and uniform distributions to reflect realistic shot outcomes across various scenarios, such as driving, fairway shots, approach shots, and putting. A binary decision model further refines shot location outcomes based on player-specific statistics like Greens in Regulation Percentage and Scrambling. While the results demonstrate the utility of this approach in simulating player performance, they also highlight areas for improvement, including fine-tuning parameter weights and resolving imbalances in the current implementation. This study underscores the potential for enhancing golf performance simulations through statistical and probabilistic modeling.

1 Introduction

Golf is a sport defined by variability, where outcomes are influenced by distance, terrain, and player precision. This project simulates a PGA tournament by integrating player performance metrics with statistical modeling to replicate realistic shot scenarios. Data obtained from the PGA Tour's official website includes metrics like Scoring Average, Strokes Gained (SG), and Greens in Regulation.

The simulation applies statistical distributions tailored to specific shot types: lognormal and uniform distributions for longer shots, and truncated normal distributions for approach shots and putting. Binary decision models determine shot outcomes, such as landing on the green or rough, using weights derived from player statistics.

By running the simulation 10,000 times, this project highlights the potential for realistic golf performance modeling while identifying areas for improvement, such as parameter fine-tuning and addressing imbalances favoring top-ranked players.

2 Methodology

2.1 Datasets

The datasets for this project were obtained from the PGA Tour's official website, which provides a comprehensive array of performance statistics for 175 professional players. To effectively simulate realistic golf scenarios, A range of key performance was utilized. These metrics were grouped into broader categories to reflect the different phases of play in golf. Below is a table summarizing the datasets and their specific contributions to the simulation.

Performance Scenario	**Corresponding Data**
Driving	• Driving Distance • Driving Accuracy Percentage
Fairway shot	• Greens in Regulation Percentage • SG: Approach the Green
Rough shot	• Scrambling • Birdie or Better Percentage
Approach shot	• SG: Around-the-Green
Putting	• Putting Average • SG: Putting • 3-Putt Avoidance
Overall	• Scoring Average • SG: Total • Birdie Average • SG: Tee-to-Green • SG: Off-the-Tee

Table 1: A table of expected performance scenarios and their corresponding datasets

SG stands for Strokes Gained, a metric that measures how many strokes a player gains or loses relative to the field average across specific aspects of play. To ensure concision and consistency in calculations, I converted certain statistical values, particularly averages like Scoring Average—the average number of strokes a player takes per round—into comparative values. This was necessary because raw averages can be tricky to incorporate directly into formulas. The comparative value was calculated using the formula:

$$\text{Comparative Value} = \frac{\text{Original Value} - \text{Mean}}{\text{Standard Deviation}}$$

This transformation standardizes the data, allowing for smoother integration into calculations and better comparisons across players' performances.

The data was originally separated into smaller CSV files, so it was essential to merge these files into a single cohesive dataset using the pandas library. This step ensured that all necessary statistics for individual players were accessible for subsequent analysis and implementation.

2.2 Libraries

To implement the simulation, I employed key Python libraries: NumPy for statistical calculations, pandas for managing and analyzing data, random for stochastic probability modeling, and PyQt5 for visualizing the final tournament leaderboard. Specifically, PyQt5 enabled the creation of leaderboard that displayed the results in a clear and interactive format.

2.3 Algorithms

2.3.1 Distributions

At the heart of the simulation lies the concept of statistical distributions, which were applied to different shot scenarios to capture realistic variability. Three distinct distributions—lognormal, truncated normal, and uniform—were utilized based on the nature of each scenario.

To demonstrate the effectiveness of these distributions, I sampled 4 to 5 players for each histogram below based on their scoring averages. Scottie Scheffler, who ranks 1st, was included as his statistics are exceptional and provide a clear benchmark. Adam Hadwin, positioned in the middle of the scoring average rankings, and Camilo Villegas, ranked at the bottom, were selected to show the contrast across different performance levels. Chris Kirk was also included because his scoring average is closest to the mean of all players, serving as a reference point for average performance. Additionally, Xander Schauffele, ranked 2nd, was sampled to balance the effect of Scottie's dominance and avoid the distribution appearing overly biased toward a single outstanding performer.

Lognormal Distribution

First, the lognormal distribution was used to simulate the scenario of shots to the green. A lognormal distribution is characterized by the logarithm of its variable following a normal distribution, resulting in a right-skewed shape. This makes it particularly suitable for modeling comparatively longer-distance shots, as the remaining distance to the hole often clusters at shorter values for skilled players but still allows for occasional large errors.

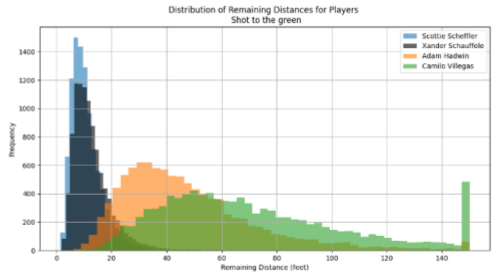

Figure 1: A histogram of remaining distances for players' shot from the fairway to the green

In Figure 1, the x-axis represents the remaining distance after the shot, while the y-axis shows the frequency of outcomes. For players like Scottie Scheffler and Xander Schauffele, who are top-ranked, the peaks occur around 10 feet, reflecting their high level of accuracy. However, the right-skew ensures that occasional missed shots result in remaining distances as high as 40 feet, reflecting the inherent variability in golf performance. This right-skew captures the essence of inconsistency, where even top players are not immune to poor shots.

Truncated Normal Distribution

The truncated normal distribution was used for approach shots and putting scenarios. A truncated normal distribution is derived from a normal distribution but with its tails "cut off" at specified limits. This ensures that the values remain within a realistic range. For instance, when a player misses an approach shot, the ball typically lands near the green rather than far away. Similarly, a missed putt still lands near the hole, reinforcing the consistency of short-distance shots. The truncated normal distribution reflects these limitations, ensuring that extreme values, which would be unrealistic in such scenarios, are eliminated. This approach was particularly effective in maintaining realism for approach shots and putting, where precision is generally higher than for longer-distance shots.

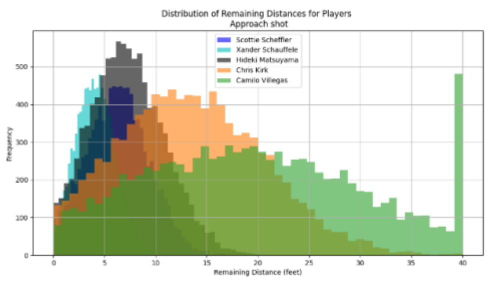

Figure 2: A histogram of remaining distances for players' approach shot

Figure 2 shows the distribution of remaining distances for approach shots across five players. Scottie Scheffler and Xander Schauffele demonstrate sharper peaks around shorter distances, indicating their exceptional accuracy on approach shots. Meanwhile, players like Chris Kirk and Camilo Villegas show a broader spread, reflecting less precision.

Hideki Matsuyama is included because he has a good value for approach-related metrics like SG: Around-the-Green. Including him helps visualize how effectively the truncated normal distribution works, as parameters such as bias, standard deviation, and additional weights to reflect realistic shot outcomes, are manipulated.

Uniform Distribution

For driver shots, the uniform distribution was employed due to its simplicity and consistency. A uniform distribution gives equal probability to all values within a specified range, making it ideal for modeling the driver shot, which is the most consistent club in golf. Drivers are generally used to achieve maximum distance off the tee, and the resulting outcomes fall within a relatively narrow range. The uniform distribution reflects this consistency, as players hit the ball to similar distances with minor deviations.

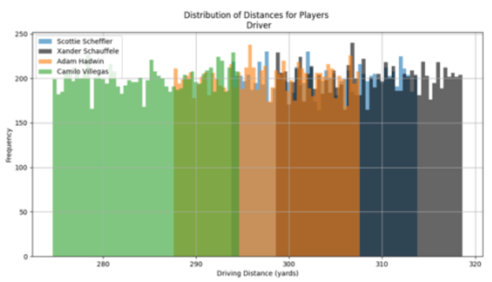

Figure 3: A histogram of distances for players' driving shot

In Figure 3, the x-axis represents the driving distance after a driver shot, while the y-axis indicates the frequency of these outcomes. The distribution shows that driver shots result in relatively uniform outcomes within a specified range, which reflects the consistency of this club compared to others. Each player, including Scottie Scheffler, Xander Schauffele, Adam Hadwin, and Camilo Villegas, exhibits a distinct band that aligns with their average driving distances. Scottie and Xander, being top-ranked players, produce results that cluster toward longer distances, while Camilo Villegas, ranked lower, has outcomes concentrated in the shorter range. This uniform distribution ensures equal probability across a defined range, making it suitable for modeling driver shots, which tend to be more consistent and less prone to extreme variability compared to shorter shots like approaches or putts.

2.3.2 Binary Decision Modeling

In addition to distribution modeling, binary decision modeling was introduced to determine the outcome locations of shots. For example, from the fairway to the green:

$$\text{Fairway\_to\_Green\_Outcomes} = \{\text{Green}, \text{Rough\_around\_the\_Green}\}$$

$$\text{Weights for "Green"} = \text{Greens in Regulation Percentage}$$

$$\text{Weights for "Rough\_around\_the\_Green"} = 1 - \text{Greens in Regulation Percentage}$$

there are two primary outcomes: the ball lands on the green, or it ends up in the rough around the green. The probabilities of these outcomes were weighted based on player-specific stats such as greens-in-regulation percentage. If a player has a 90% greens-in-regulation stat, there is a 90% chance their shot will land on the green.

For scenarios where stats were less impactful, such as putting, bias was introduced to ensure realistic variability. In the case of putting outcomes, two possibilities are considered: "In the hole" or "Green," which means remaining on the green after a missed putt:

$$\text{Putting Outcomes} = \{\text{"In the hole"}, \text{"Green"}\}$$

The weights for these outcomes are determined using a combination of the 3-Putt Avoidance stat and a small bias adjustment. Specifically, the weight for "In the hole" is calculated as:

$$\text{Weights for "In the hole"} = 1 - (\text{3-Putt Avoidance}) + 0.1$$

which slightly increases the likelihood of a successful putt. Conversely, the weight for "Green" is adjusted as:

$$\text{Weights for "Green"} = (\text{3-Putt Avoidance}) - 0.1$$

which ensures that the probability of a missed putt remains realistic while accounting for natural variability. These adjustments help balance precision and inconsistency, reflecting the nature of short-distance shots where performance tends to be more stable but not entirely predictable.

3 Results

Using these methods, the simulation was run 10,000 times to produce a tournament leaderboard, showing the frequency of wins for each player. However, the results revealed some imbalances. For example, Scottie Scheffler won over 8,000 games and the next closest player, Xander Schauffele, won around 1,500 games, while all other players recorded fewer than 10 wins. The result highlights an issue with the current implementation.

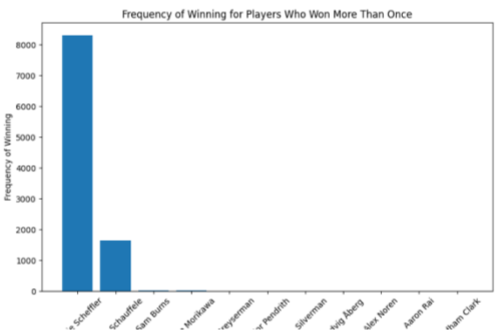

Figure 4: A diagram of the frequency of winning for top eleven players

4 Discussion

The observed imbalance is primarily attributed to fine-tuning issues within the algorithm. The weights applied to certain stats, such as Scoring Average and Greens in Regulation Percentage, disproportionately favored top players like Scottie and Xander. For instance, in the approach shot scenario, A dynamic variance is introduced based on scoring average, where better scoring averages resulted in lower variance. Combined with additional weights assigned to other stats, this created an unfair advantage for top-ranked players. To address this, I plan to reduce the influence of these weights or adjust the parameters to ensure greater parity among players.

Another issue identified was the lack of a playoff scenario to resolve ties. Currently, when two players, including Scottie Scheffler, tie in the simulation, Scottie Scheffler is always declared the winner because he occupies the first position in the dataset. Implementing a sudden-death playoff or tiebreaker logic would resolve this issue and make the simulation more realistic.

Looking ahead, the project has potential applications in machine learning for predictive analytics. For example, machine learning models could be developed to predict driver distances based on driver accuracy or other related stats. By incorporating machine learning, the simulation could adapt dynamically to player performance and external factors, improving accuracy and realism.

5 Conclusion

This project successfully simulated a PGA golf tournament by combining real-world player statistics with tailored statistical distributions and binary decision models. The lognormal, truncated normal, and uniform distributions effectively captured the variability in different shot scenarios, while weighted probabilities determined realistic shot outcomes.

Despite the model's strengths, results showed significant imbalance favoring top-ranked players, emphasizing the need for fine-tuning parameters such as variance, weights, and biases. Future improvements include incorporating playoff scenarios for tied results and exploring machine learning to predict performance metrics. This work demonstrates the potential of statistical modeling to replicate complex sports dynamics while highlighting areas for further refinement.

6 References

Keelin, Thomas W. "The Multivariate Metalog Distributions with Application to Strategic Decision-Making in Golf."

“Stats.” *PGA Tour*, 2024, https://www.pgatour.com/stats. Accessed 20 Oct. 2024.

Stephenson, Paul, et al. "How LO can you GO? Using the dice-based Golf Game GOLO to illustrate inferences on proportions and discrete probability distributions." *Journal of Statistics Education* 17.2 (2009).