Predicting the Outcomes of Tennis Tournaments

A Monte Carlo Approach

Olivia Beck, Emma Lewis, Ryan Volkert

Decemeber 19, 2019

Contents

List of Figures

List of Tables

Introduction

This report replicates the methods outlined by Newton and Keller (2005) for predicting the outcome of a tennis tournament. We used the empirical probability of each player winning a rally given they served to calculate theoretical probabilities that the player wins a game, a set, a match, and a tournament. We first used the same data that Newton and Keller (2005) used for the 2002 Wimbledon and US Open tournaments. Our report then expands on the original study by doing the same analysis for the 2019 Wimbledon tournament.

Motivation

The probability of winning a game, a set, and a match in tennis are calculated based on each player's probability of winning a point on a serve. This paper and the reference paper both make the assumption that each serve is an independently and identically distributed (iid) random variable. This assumption has been made in reports relating to the probability of winning a game in other racket sports. It is important to note that serves in tennis are not iid random variables in reality. For most purposes this the divergence from iid is small, so we ignore the dependency.

A game in tennis is played with one player serving and is won by the first player to score four or more points and to be at least two points ahead of the other player. Player A can win a game against Player B by a score of (4,0), (4,1) or (4,2), or else the score becomes (3,3) which is called "deuce".

In a set the players alternate serving until a player wins at least six games and is ahead by at least two games. If the game score reaches 6-6, a 13-point tiebreaker is used to determine who wins the set.

To win a match in the women's format, a player must win two out of three sets and three out of five in the men's format.

Methods

Data for the 2002 tournaments was provided by Newton and Keller (2005). Data for the 2019 Wimbledon for was found at IBM Corp. (n.d.).

The probability of winning a game, set, match, and tournament is derived in Newton and Keller (2005) as follows:

- P(A Winning a Rally | A Served)

$$p_A^R = \frac{\text{Points Won On Serve}}{\text{Points Served}} \tag{1}$$

- P(A Winning a Game | A Served)

$$p_A^G = (p_A^R)^4[1 + 4q_A^R + 10(q_A^R)^2] + 20(p_A^R q_A^R)^3(p_A^R)^2[1 - 2p_A^R q_A^R]^{-1} \tag{2}$$

- P(A Winning a Set | A Served)

$$p_A^S = \sum_{j=0}^{4} p_A^S(6, j) + p_A^S(7, 5) + p_A^S(6, 6)p_A^T \tag{3}$$

$p_A^S(i, j)$ is defined recursively as:

$p_A^S(0, 0) = 1$, $p_A^S(i, j) = 0$ if $i < 0$ or $j < 0$.

- if i+j -1 is even: $p_A^S(i, j) = p_A^S(i-1, j)p_A^G + p_A^S(i, j-1)q_A^G$
  - omit i-1 term if j=6 and i<6

  - omit j-1 term if i=6 and j<6
- if i+j -1 is odd: $p_A^S(i,j) = p_A^S(i-1,j)q_B^G + p_A^S(i,j-1)p_B^G$
  - omit i-1 term if j=6 and i<6
  - omit j-1 term if i=6 and j<6.
- P(A Winning the Tie Breaker | A Served Initially)

$$p_A^T = \sum_{j=0}^{5} p_A^T(7,j) + p_A^T(6,6)p_A^R q_B^R[1 - p_A^R p_B^R - q_A^R q_B^R]^{-1} \tag{4}$$

$p_A^T(i,j)$ is defined recursively as:

$p_A^T(0,0) = 1$, $p_A^T(i,j) = 0$ if i<0 or j<0.

- if $i + j - 1 \mod 4 \equiv 0$ or 3: $p_A^S(i,j) = p_A^T(i-1,j)p_A^R + p_A^T(i,j-1)q_A^R$
  - omit i-1 term if j=7 and i<7
  - omit j-1 term if i=7 and j<7
- if $i + j - 1 \mod 4 \equiv 1$ or 2: $p_A^T(i,j) = p_A^T(i-1,j)q_B^R + p_A^T(i,j-1)p_B^R$
  - omit i-1 term if j=7 and i<7
  - omit j-1 term if i=7 and j<7
- P(Winning a Women's Match, Best of 3)

$$p_A^M = (p_A^S)^2 + 2(p_A^S)^2 p_B^S \tag{5}$$

- P(Winning a Men's Match, Best of 5)

$$p_A^M = (p_A^S)^3 + 3(p_A^S)^3 p_B^S + 6(p_A^S)^3(p_B^S)^2 \tag{6}$$

- P(Winning the Tournament): Let $\boldsymbol{P}_{ij}$ be the probability that player i wins the match against player j. Then for the 4 semifinalists players i, j, k, and l:

$$p_i^{TC} = \boldsymbol{P}_{ij}(\boldsymbol{P}_{ik}\boldsymbol{P}_{kl} + \boldsymbol{P}_{il}\boldsymbol{P}_{lk}) \tag{7}$$

Note that $\sum_{\forall i} p_i^{TC} = 1$. We will sample one player from the 4 semifinalists with the probabilities calculated above to determine the tournament winner.

Reproducibility

First we will consider data from the 2002 US Open for both the men's and women's tournaments. We want to reproduce the probability of winning a game. Tables 1 and 2 compare the empirical probability of winning a game to the theoretical probability of winning a game calculated by Newton and Keller (2005) and the probability that we calculated for the four semi-finalists in each tournament. We can see that the values that we calculated and the values calculated by Newton and Keller (2005) are the same. These values do differ slightly from the empirical probabilities.

Next, we want to check the probability of winning a tournament. Newton and Keller (2005) do not provide their predictions of tournament winners. They also only consider the four semi-semifinalists in each tournament, so we will as well.

Table 1: Data for the Women's Semifinalists in the 2002 U.S. Open Tournament

	P(Win a Rally)	Empirical P(Win a Game)	Paper P(Win a Game)	Our P(Win a Game)
S. Williams	0.69	0.71	0.89	0.89
V. Williams	0.63	0.80	0.79	0.79
L. Davenport	0.65	0.85	0.83	0.83
A. Mauresmo	0.63	0.77	0.79	0.79

Table 2: Data for the Men's Semifinalists in the 2002 U.S. Open Tournament

	P(Win a Rally)	Empirical P(Win a Game)	Paper P(Win a Game)	Our P(Win a Game)
P. Sampras	0.73	0.95	0.93	0.93
A. Agassi	0.66	0.87	0.85	0.85
L. Hewitt	0.67	0.85	0.86	0.86
S. Schalken	0.68	0.90	0.88	0.88

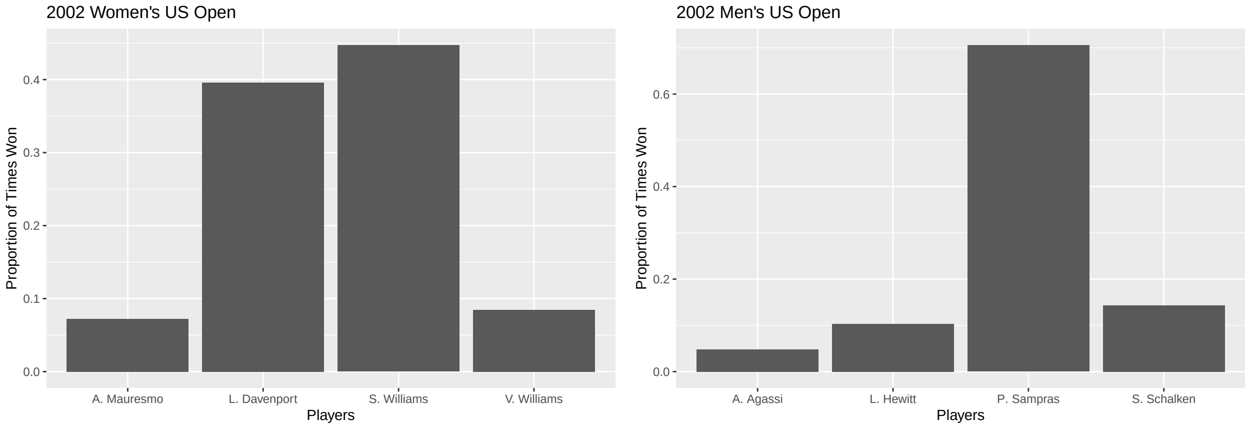

Figure 1: 2002 US Open

We first look at the 2002 US Open. We calculated the probability each semi-finalist has of winning the tournament, then randomly sampled one player to win calculated weights. We repeated this 2,000 times, and took the player with the most sampled wins to be the predicted winner. Figure 1 shows the results from this method for the men's and women's tournaments. Our simulations predict that Serena Williams wins the women's tournament, and Pete Sampras wins the men's tournament. These results are consistent with the outcome of the actual 2002 US Open.

The 2002 Wimbledon is considered below.

Table 3: Data for the Semifinalists in the 2002 Wimbledon Tournament

Women	P(Win a Rally)	Men	P(Win a Rally)
S. Williams	0.71	L. Hewitt	0.77
V. Williams	0.67	D. Nalbandian	0.61
J.Henin	0.59	T. Henman	0.67
A. Mauresmo	0.64	X. Malisse	0.67

Table 4: Data for the Semifinalists in the 2019 Wimbledon Tournament

Women	P(Winning a Rally)	Men	P(Winning a Rally)
S. Williams	0.6416465	S. Williams	0.6416465
B. Strycoa	0.6559140	B. Strycoa	0.6559140
E. Suitolina	0.5791855	E. Suitolina	0.5791855
S. Halep	0.6302895	S. Halep	0.6302895

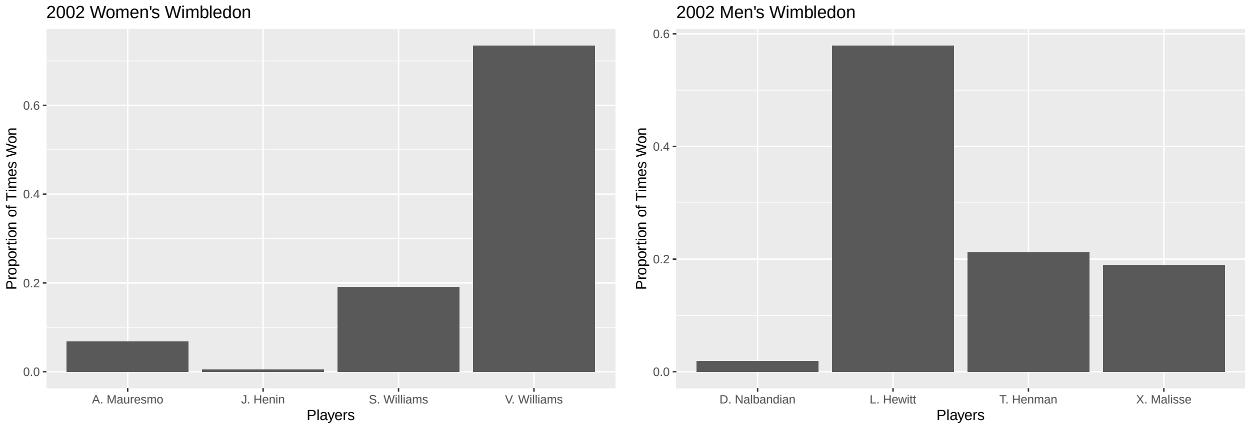

Figure 2: 2002 Wimbledon

The probability of winning a rally for the four semifinalists for both women's and men's can be found in Table 3. Figure 2 shows the results of our method. We predict that Venus Williams wins, but in reality Serena Williams had an upset win. Some speculations as to why we predicted incorrectly can be found in the limitations section. For the men's semi finalists, we predict L. Hewitt to win the tournament, and in fact he does.

Extension

For our extension, we ran the same method on data from the 2019 Wimbledon Tournament for both men's and women's. The data for probability of winning a rally can be found in Table 4. As seen in Figure 5, we predict that B. Strycoa wins the tournament, but S. Halep actually won the 2019 Women's Wimbledon. Similarly, in Figure 6 we predict that Rodger Federer and Rafeal Nadal essentially have an equal probability of winning the tournament, but it was Novak Djokovic who actually won the tournament. Again, the discrepancies between who won the tournament and what we predicted are discussed in the limitations section.

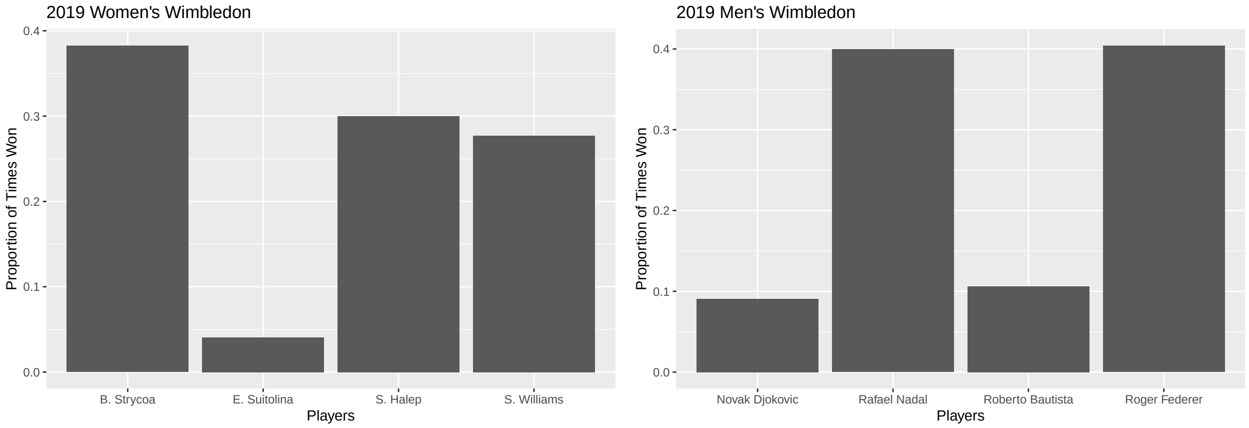

Figure 3: 2019 Wimbledon

Conclusions

Overall, we were able to predict the winner of a tennis tournament from the 4 semi finalists with relative accuracy using monte carlo methods. We used the empirical probabilities of winning a rally on serve going into a tournament to calculate the probability of the 4 semi finalists winning the tournament. The discrepancies between our predictions and the actual winners of the tournament can be contributed to the fact that we assumed all rallies are iid, when in reality, this is not the case. This will be addressed further in the limitations section.

Limitations

When a tennis tournament is played, the empirical probability of winning a point on serve changes every time a rally is completed. This report does not take that into account. For this report we used previous tournament data to calculate the probability each player had of winning the tournament. Our data did not update as the tournament progressed. There is one major draw back to this approach. Our method only considers how a player was performing before the tournament and does not consider how they are performing during the tournament. This is one reason we see discrepancies between our predicted tournament winner and the actual winner of the tournament. For example, in the 2002 Women's Wimbledon, Venus Williams was predicted to win, as she had the highest probability of winning a rally, but Serena Williams actually won the tournament. This is in part due to the discrepancies between how the players performed before the tournament and how they preformed during the tournament. Serena Williams performed much better during the tournament than she did leading up to it. One way to improve our predictions is to update our probabilities each time a rally is conducted.

When a player plays well during a tournament, it gives them momentum which is important to winning in any sporting event. This report also did not take momentum into account. We assumed that all rallies were iid, but in reality, this is not the case. Newton and Keller (2005) assumes that the difference between iid and non-iid rallies is so small that for their purposes, and for ours, they assume that all rallies are iid. Newton and Aslam (2006) considers predictions for winning a tournament where rallies are non-iid.

References

IBM Corp., AELTC. n.d. “Player Statistics.” *Wimbledon.com.* https://www.wimbledon.com/en_GB/scores/extrastats/index.html.

Newton, Paul K., and Kamran Aslam. 2006. “Monte Carlo Tennis.” *SIAM Review* 48 (4): 722–42. http://www.jstor.org/stable/20453873.

Newton, Paul K., and Joseph B. Keller. 2005. “Probability of Winning at Tennis I. Theory and Data.” *Studies in Applied Mathematics* 114 (3): 241–69. https://doi.org/10.1111/j.0022-2526.2005.01547.x.