Tennis

Olivia Beck, Emma Lewis, Ryan Volkert

Decemeber 19, 2019

Introduction

For our report, we replicated the methods outlined in Newton and Keller (2005) for predicting the outcome of a tennis tourement.

Motivation

The probability of winning a game, a set, and a match in tennis are calculated based on the each players probability of winning a point on a serve. This paper and the reference paper,Newton and Keller (2005) both make the assumption that each serve is an independently and identically distributed (iid) random variable. This assumption has been made in other reports relating to the probability of winning a game in other racket sports and even though serves in tennis are not iid random variables in reality. For most purposes this is not an outragous assumption as the divergence from iid is small.

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 a player in woman's format a player must win two out of three sets and in men's format a player must win three out of five sets.

Methods

Data for the 2002 Wimbledon for men's and women's as well as 2002 US Open for men's and women's was provided in Newton and Keller (2005). Data for 2019 Wimbledon for men's and women's 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)

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

- P(A Winning a Game | A Served)

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

- P(A Winnning a Set | A Served)

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

- P(A Winning the Tie Breaker | A Served Initially)

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

- P(Winning a Women's Match, Best of 3)

$$p_A^M = (p_A^S)^2 + 2(p_A^S)^2 p_B^S$$

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

$$p_A^M = (p_A^S)^3 + 3(p_A^S)^3 p_B^S + 6(p_A^S)^3 (p_B^S)^2$$

Where, $p_A^S(i,j)$ 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.

And, $p_A^T(i,j)$ 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

Reproducability

First we will consider data from the 2002 US Open for both men's and women's. We want to reproduce the probability of winning a game. Table 1 compares the empirical probablity of winning a game versus the theoretical probability of winning a game calculated in Newton and Keller (2005) verus the probability of winning a game that we calculated for the 4 semi finalists in this tournament. We can see that the vales that we caluculated and the values calculated in Newton and Keller (2005) are the same. These values do differ slightly from the empiricial probabilities. Table 2 outlines the same comparisons except for the men's 2002 US Open.

Next, we want to comapare the probability of winning a tournement. Newton and Keller (2005) does not provide their predictions of tournament winners, only their probabilies comparted to what actually happened in the tournement. The paper also only considers the 4 semifinalists when predicting the tournement, so we will as well.

Table 1: Data for the Womens Semifinalists in the 2002 U.S. Open Tournament

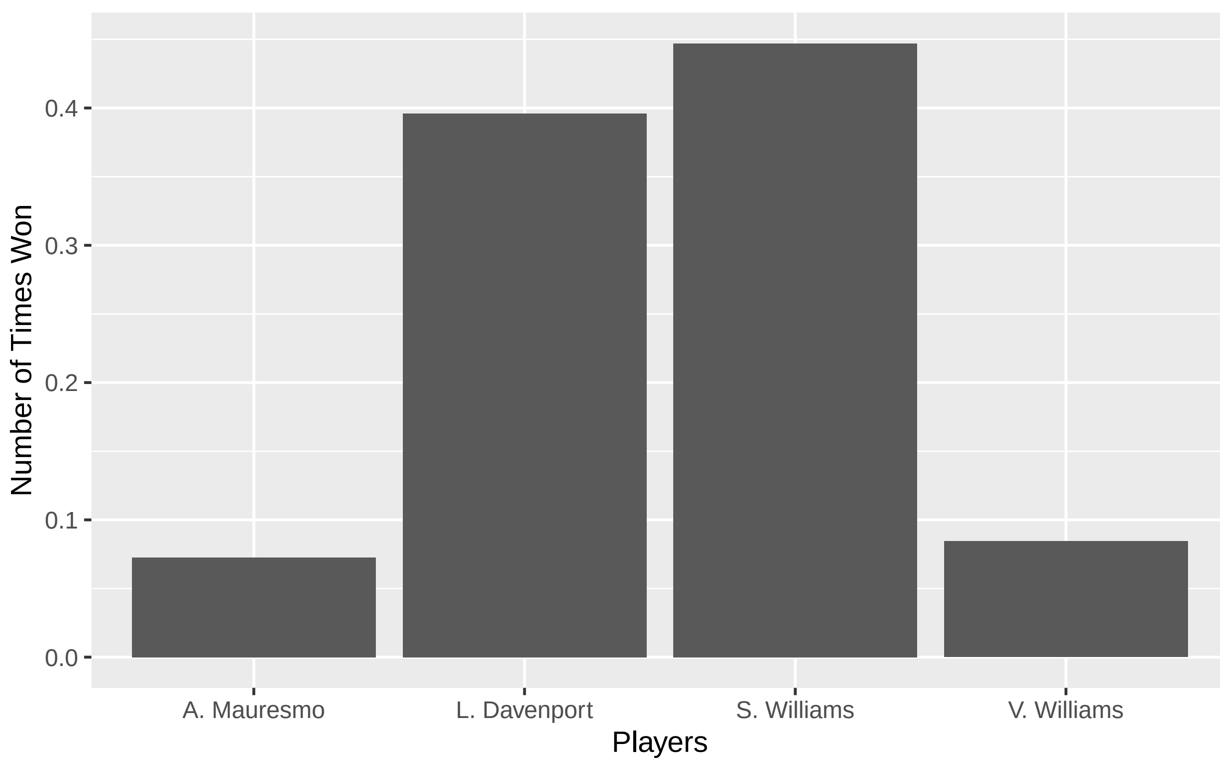
	P(Win a Rally)	Empirical P(Win a Game)	Paper P(Win a Game)	Our P(Win a Game)
S. Williams	0.69	0.71	0.89	0.89
V. Williams	0.63	0.80	0.79	0.79
L. Davenport	0.65	0.85	0.83	0.83
A. Mauresmo	0.63	0.77	0.79	0.79

Table 2: Data for the Mens 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 Women's US Open

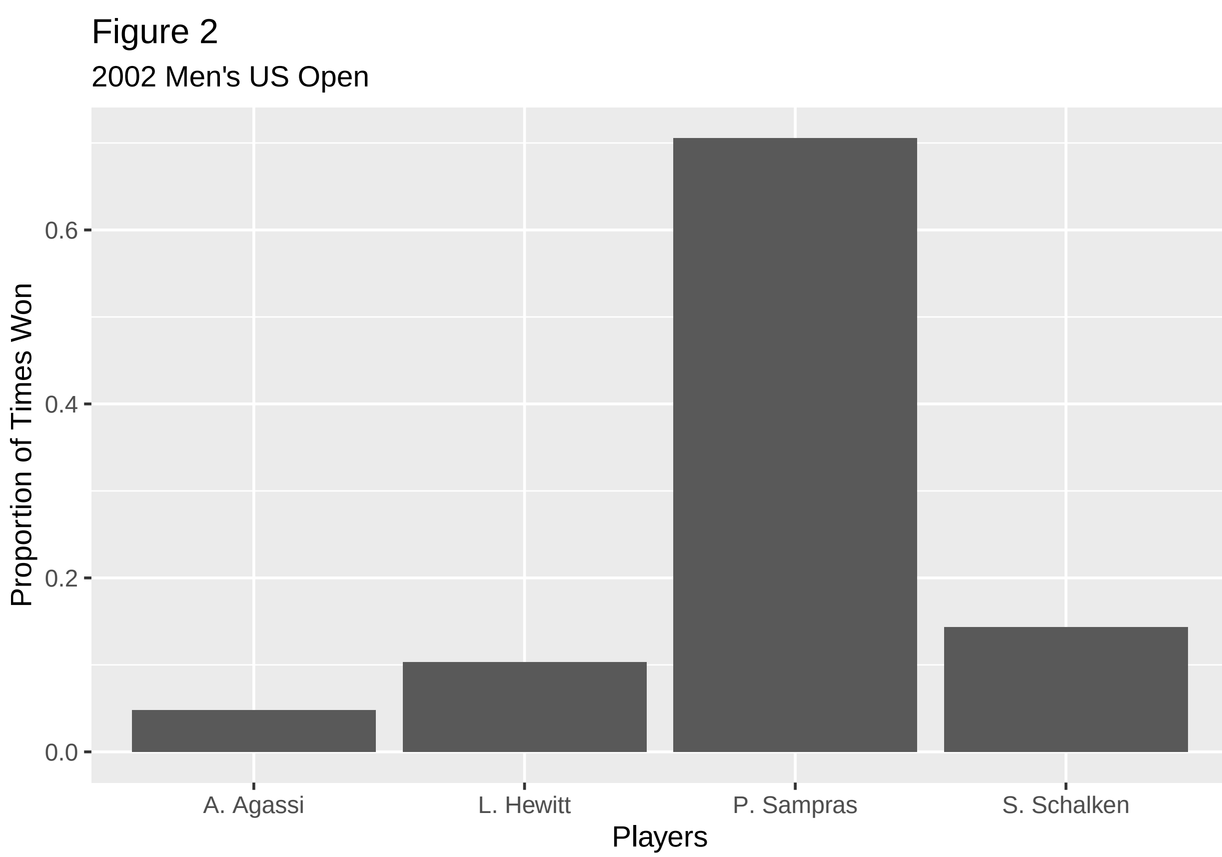


We fist look at 2002 US Open's Womens tournement for the 4 semi finalists. We calculate the probability each semi finalist has of winning the tournament, then we sample one player to win the tournament with the probabilites we calculated. Figure 1 shows the results from 2000 tournements played with these 4 semi finalists. From our simlations we predict that S. Williams wins the tournement, and in fact she did win this tournement.

Similarly, we want to consider the 2002 US Open for the men's semifinalists. Using the same method outline above, we can see in Figure 2 that we predict P. Sampras to win the tournement, and in fact he did win this tournament.

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



Next, we consider data from the 2002 Wimbeldon. The probability of winning a rally for the 4 semifinalists for both women's and men's can be found in Table 3. Figure 3 shows the results for the 2002 Women's Wimbledon. We predict that V. Williams wins, but in reality S. Williams had an upset win. Some speculations as to why we predicted incorrectly can be found in the Limitations section. Figure 4 shows the same methods as Figure 3, but for the men's semi finalists. We predict L. Hewitt to win the tournament, and in fact he does.

Figure 3

2002 Women's Wimbledon

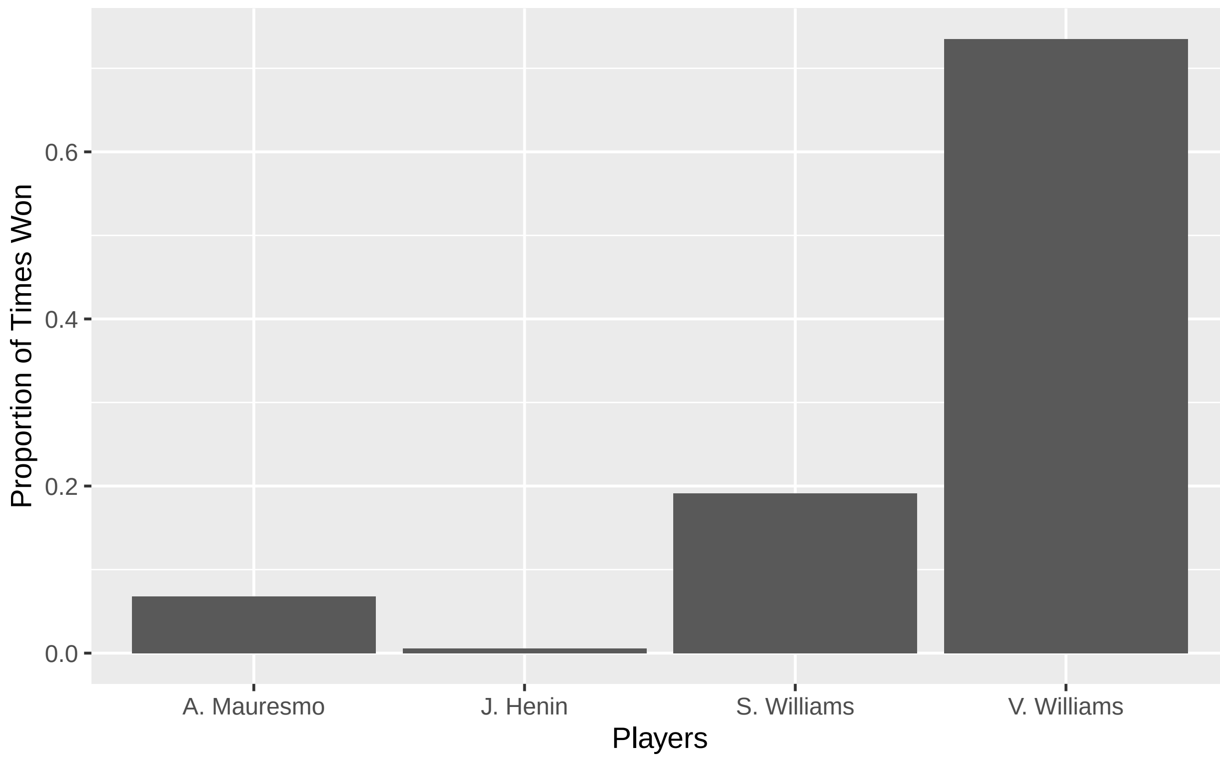
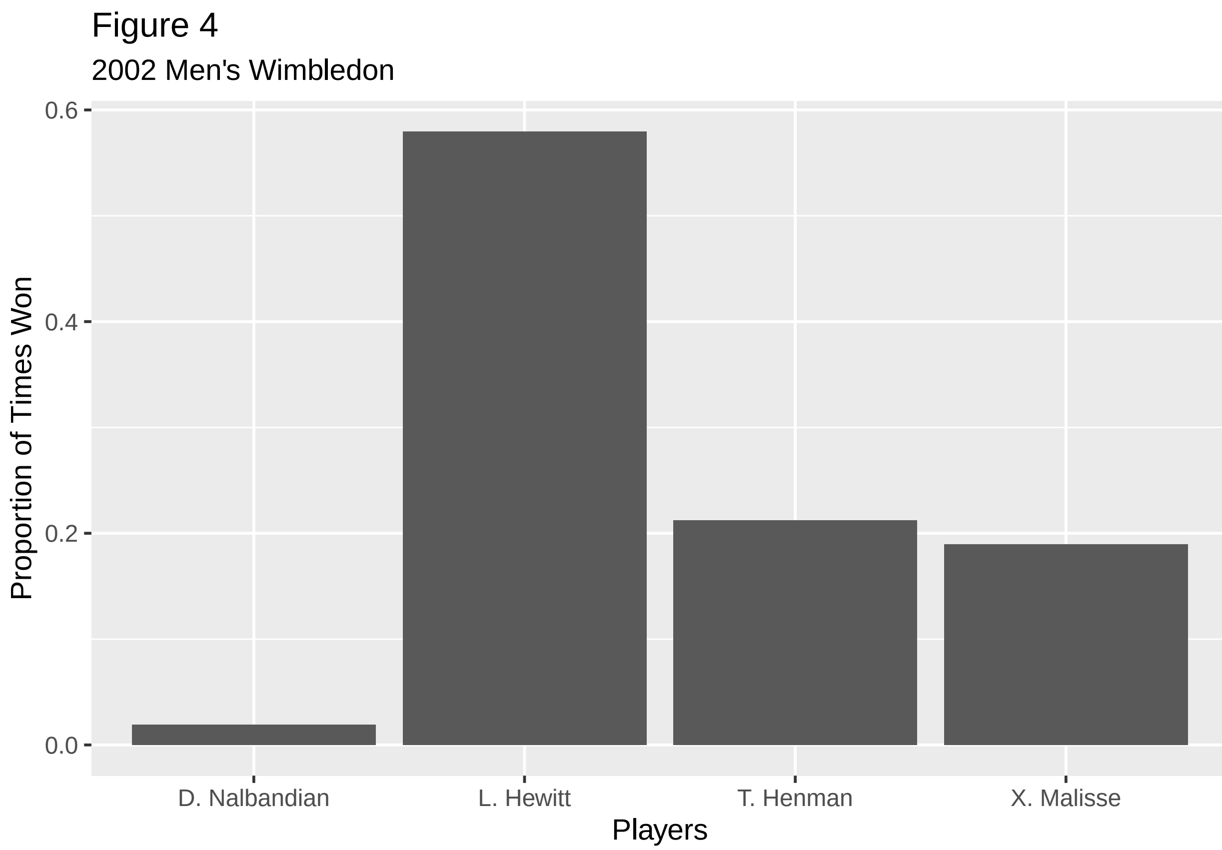


Table 4: Data for the Semifinalists in the 2019 Wimbledon Tournament

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



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 tournement, but it was Novak Djokovic who actually won the tournement. Again, the discrepencies between who won the tourment and what we predicted are discussed in the limitations section.

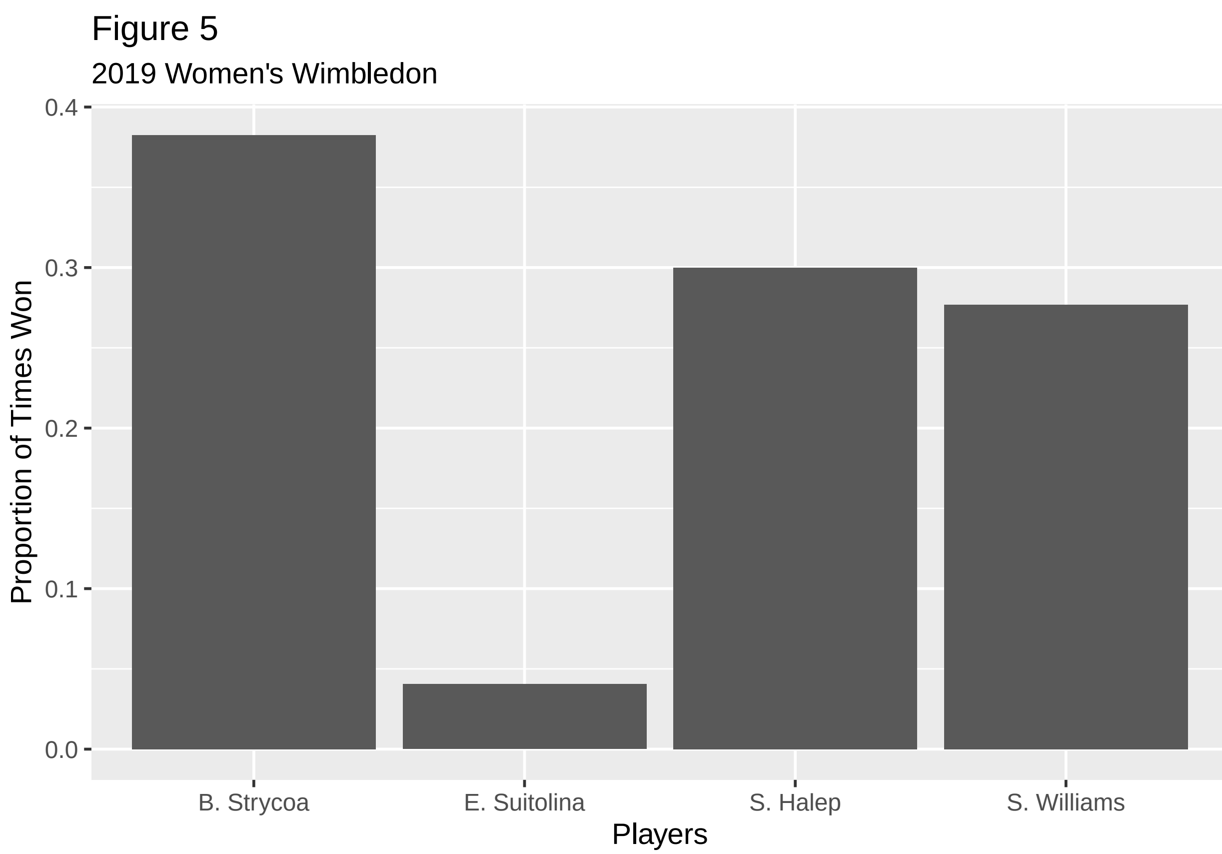
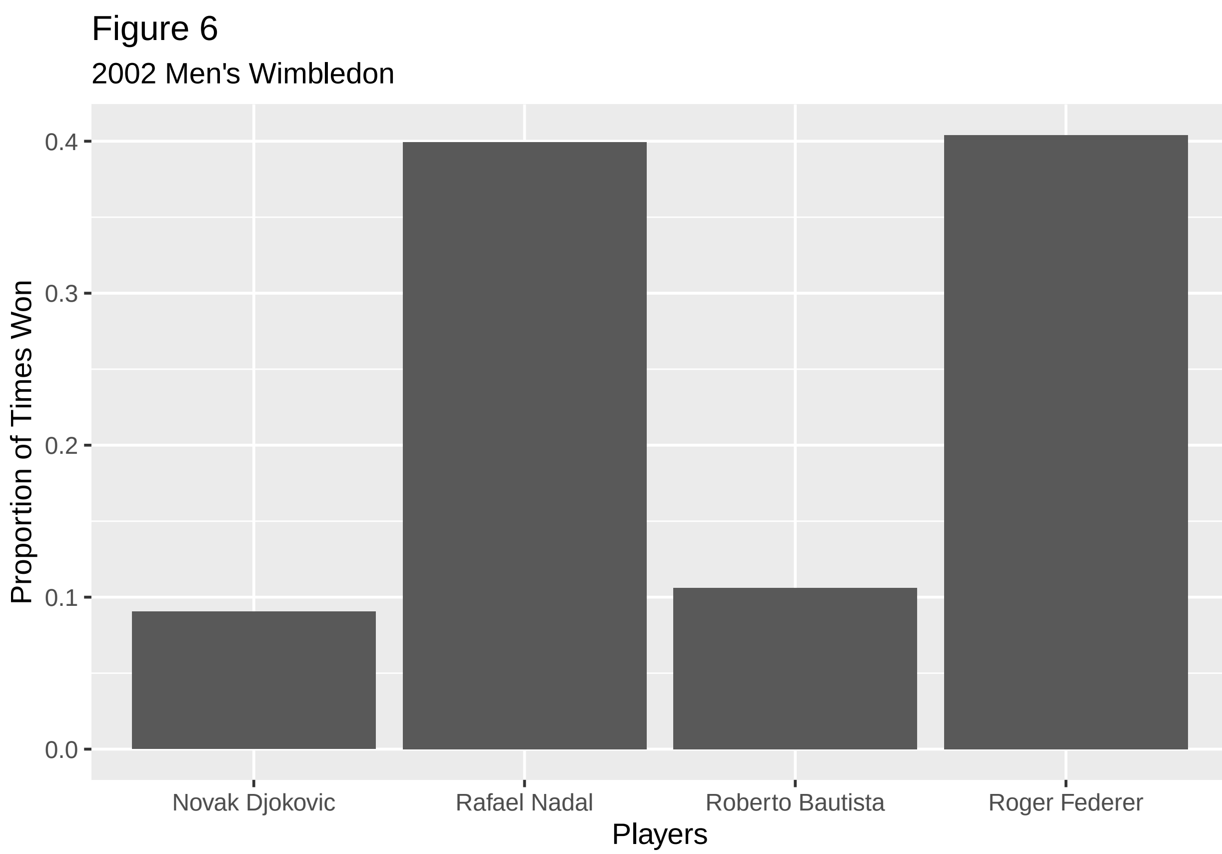


Figure 5
2019 Women's Wimbledon
0.4
0.3
0.2
0.1
0.0
Proportion of Times Won
B. Strycoa
E. Suitolina
S. Halep
S. Williams
Players



Conclusions

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.

Bibliography

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). Society for Industrial; Applied Mathematics: 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. doi:10.1111/j.0022-2526.2005.01547.x.