Tennis: The sport where 1% REALLY matters

Tuna Lokman

Introduction

Being a Tom Rocks Maths viewer myself, the recent video about goalball being the most mathematical sport was certainly intriguing. As a person who has been playing tennis since the age of four, I could not help but think, in what way can I apply maths and statistics to my passion?

Thinking about this reminded me of a time when I was watching one of the grand slams on my television. Up come a statistic, which, seemed almost fake:

Roger Federer has only won 54% of the points he has played.

Now, to you, if this seems ordinary, let me clarify why I find this statistic mindblowing. Federer is arguably the greatest tennis player of all time: 20 Grand Slams, 8 Wimbledon Championships… he is one of the names you simply do not see lose. I grew up watching him win, win, win, and win again:

Roger Federer has a win-rate of 82%.

These 2 completely vast percentages were almost begging me to further investigate. I could predict that having a point-win rate of 54% would lead to a higher match win percentage, but was 82% an outlier? Something extraordinary that only Federer could achieve? Or was it something caused by the magic of maths and probability… a percentage that a 54% point-win rate was bound to lead to.

Let's investigate, using data from https://www.wheeloratings.com/index.html.

Format

Before we begin, let me clarify the scoring format of tennis:

- Best of 3 sets used in all ATP (Association of Tennis Professionals) matches except Grand Slams.
- Each set is first to 6.
- Each game is scored: 0, 15, 30, 40, Win.

There are a few exceptions:

- In the case of a score of 5:5 in a set, the set will become first to 7 games.
- In the case of a score of 6:6 in a set, a tiebreaker to 7 points is played.
- In the case where the game goes to 40-40 (Deuce), the scoring becomes: Advantage, Win. If a point is lost on Advantage, the Advantage is lost, and score goes back to 40-40.

Now, time for the fun part!

Assumptions:

- Roger has a chance of 54.1% to win every point.
- The match follows the format above.

Probability of Roger winning a game

Let:

$$p = P(\text{Roger wins a point}) = 0.541$$
$$q = P(\text{Roger loses a point}) = 0.459$$
$$k = \text{the number of points Roger loses}$$

<u>Winning before deuce:</u>

Roger could win the game (win 4 points) before deuce by losing 0, 1, or 2 points. However in each case, he must have won the last point. We can model this using 3 separate binomial distributions, which we can combine using the variable k:

$$\sum_{k=0}^{2} \binom{3+k}{k} p^4 q^k$$

$$P(\text{Win game before deuce}) = 0.42341$$

<u>Game reaching deuce:</u> Before we analyse the probability of Roger winning the game after deuce, we must find the probability of reaching deuce in the game.

$$\binom{6}{3} p^3 q^3$$

$$P(\text{Reaching deuce}) = 0.30624$$

<u>Winning the game after deuce:</u>

This is when the modelling gets tricky. As mentioned above, from now, to win the game, Roger must win 2 points in a row from the deuce position. Theoretically, the game could go on forever. Let $G = P(\text{Roger winning the game from deuce})$:

$$P(\text{Roger winning the game instantly from deuce}) = P(X) = p^2$$
$$P(\text{Game returns to deuce after 2 points played}) = P(Y) = 2pq$$

Therefore:

$$G = P(X) + P(Y) * G$$
$$G = p^2 + 2pq * G$$

Factor out G:

$$G(1 - 2pq) = p^2$$

$$G = \frac{p^2}{1 - 2pq}$$

To make this neater, we can use:

$$(p+q)^2 = 1$$

$$1 - 2pq = p^2 + q^2$$

$$G = \frac{p^2}{p^2 + q^2}$$

Great! Subbing in our values of p and q:

$$G = 0.58145$$

Therefore, the probability of Roger winning a game that goes to deuce is:

$$P(\text{Win game after deuce}) = P(\text{Reaching deuce}) * G$$

$$P(\text{Win game after deuce}) = 0.17806$$

Roger winning the game:

$$P(\text{Win game}) = P(\text{Win game before deuce}) + P(\text{Win game after deuce})$$
$$P(\text{Win game}) = 0.60147$$

Probability of Roger winning a set

Roger could win the set in 3 different ways:

- Win 6 games with at least a 2 game difference (6:4, 6:3, 6:2, 6:1, 6:0)
- Win by 7 games to 5 (7:5)
- Win in the tiebreak if the score reaches 6 games to 6 (7:6)

Let:

$$\alpha = P(\text{Roger wins a game}) = 0.60147$$
$$\beta = P(\text{Roger loses a game}) = 0.39853$$
$$n = \text{number of games Roger loses}$$

Win 6 games with at least a 2 game difference:

As Roger must win the last game, we can model the games up to Roger reaching 5, and multiply by α to achieve:

$$\sum_{n=0}^{4} \binom{5+n}{n} \alpha^6 \beta^n$$

$$P(\text{Win 6 games with min.2 game difference}) = 0.63679$$

Win by 7 games to 5: First, we must calculate P(Score reaching 5:5):

$$\binom{10}{5} \alpha^5 \beta^5$$

$$P(\text{Score reaching } 5:5) = 0.19942$$

To win by 7 games to 5, Roger must win both of the next 2 games.

$$\alpha^2 = 0.36177$$

Therefore:

$$P(\text{Win } 7:5) = P(\text{Score reaching } 5:5) * \alpha^2$$

$$P(\text{Win } 7:5) = 0.07214$$

Win in the tiebreak if the score reaches 6 games to 6 (7-6):

First, we must calculate P(Score reaching 6-6). We can model this as P(Score reaching 5-5), and then Roger and his opponent winning 1 of the following 2 games each.

$$P(\text{Score reaching } 6:6) = P(\text{Score reaching } 5:5) * 2\alpha\beta$$

$$P(\text{Score reaching } 6:6) = 0.09560$$

Here comes the tiebreak... in theory another aspect of tennis which could last until infinity!

Format:

- First to 7 points
- Must win by 2 points
- If the score reaches 6-6, the tiebreak carries on until a player is 2 points ahead.

Win 7 points with a min. 2 point difference (7:k):

Roger must still win the last point – therefore, this can be modelled as:

$$P(\text{Win } 7:k) = \sum_{k=0}^{5} \binom{6+k}{k} p^7 q^k = 0.50144$$

$$P(\text{Win set with tiebreak score } 7:k) = P(\text{Win } 7:k) * P(\text{Set score reaching } 6:6) = 0.04794$$

Win after score reaches 6-6:

As above, we need to calculate:

$$P(\text{Tiebreak score reaching } 6:6) = \binom{12}{6} p^6 q^6 = 0.21664$$

The end of a tiebreak, where the first player to achieve a 2 point lead wins, could also go on forever. Similarly to deuce, we can model this as:

$$\text{Let } T = P(\text{Roger winning the tiebreak from 6-6})$$

$$T = \frac{p^2}{p^2+q^2}$$
$$T = 0.58145$$

Therefore, the probability of Roger winning a tiebreak that reaches 6-6:

$$P(\text{Win tiebreak after } 6:6) = P(\text{Tiebreak score reaching } 6:6) * T$$

$$P(\text{Win tiebreak after } 6:6) = 0.12597$$

Knowing the probability of Roger winning a tiebreak after it goes to 6-6:

$$P(\text{Win set with extended tiebreak score } (a+2):(a)) =$$
$$P(\text{Win tiebreak after } 6:6) * P(\text{Set score reaching } 6:6)$$
$$P(\text{Win set with extended tiebreak score } (a+2):(a)) = 0.01204$$

Roger winning the set:

Now, we have all the information to calculate the probability of Roger winning a set.

$$P(\text{Win set without tiebreak}) = P(\text{Win 6 games with minimum two game difference}) + P(\text{Win } 7:5)$$

$$P(\text{Win set without tiebreak}) = 0.70893$$

$$P(\text{Win set with tiebreak}) = P(\text{Win set with tiebreak score } 7:k) + P(\text{Win set with extended tiebreak score } (a+2):(a))$$

$$P(\text{Win set with tiebreak}) = 0.05998$$

And finally…

$$P(\text{Win set}) = P(\text{Win set without tiebreak}) + P(\text{Win set with tiebreak})$$

$$P(\text{Win set}) = 0.7689$$

Rogers finally won the set! But the match isn't over…

Probability of Roger winning the match

Probability that Roger wins a best-of-3 set match:

$$w = P(\text{Win set}) = 0.7689$$

$$l = P(\text{Lose set}) = 0.2311$$

Roger could win the match the following ways:

- Win the first 2 sets
- Win either one of the first 2 sets, then win the last set

$$P(\text{Win match}) = w^2 + 2wl(w) = w^2(1 + 2l)$$

$$P(\text{Win match}) = 0.8645$$

Level	**Probability of Winning**	**Edge (%)**
Game	0.6015	10.15%
Set	0.7689	26.89%
Match	0.8645	36.45%

Table 1: Winning probabilities and percentage edge at game, set, and match level

As a reminder… Roger Federer has a win-rate of 81.9%.

Well, this is awkward!

As you can see, our theoretical win-rate for Federer is 4.55% higher than his actual win rate. Does this mean Federer is an underperformer? Does this mean he is not one of the greatest tennis players to live?

Well… no. All this shows is that, Federer did not magically defy the maths and specifically probability behind tennis.

To further investigate the results of our model, we can apply it to different players on the ATP tour. You can see how probability of winning a game, set and the match differs based on P(Win point) in Figure 1.

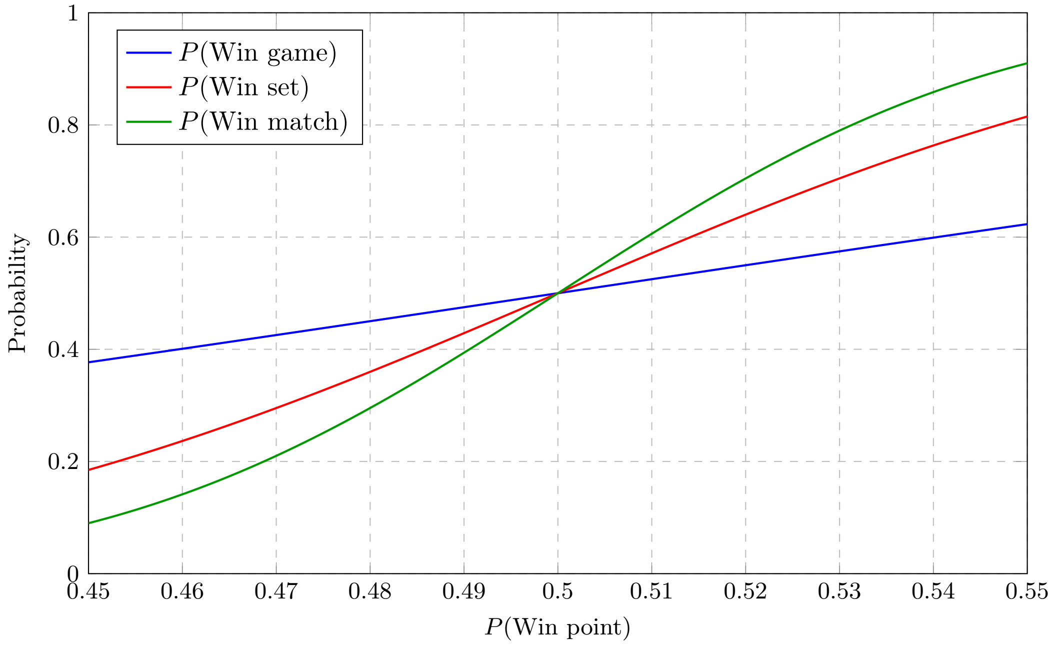

Figure 1: Relationship between point, game, set, and match win probabilities.

Expectations vs Reality: Across the Tour

Our large values of n in these scenarios means a binomial distribution requires an excessive amount of factorial operations - therefore we can apply the **Central Limit Theorem** to model our results as a normal distribution instead, using a continuity correction (± 0.5).

$$\mu = np \quad \sigma = \sqrt{np(1-p)}$$
$$P(X = x) \approx P(x - 0.5 < Y < x + 0.5)$$

Roger Federer
Point Win Probability: 0.541
Calculated P(Win match): 0.8645
Matches Played: 1545
Expected Record: 1336–209
Actual Record: 1265–280

Based on Figure 2, we can interpret that Federer's career win-rate, is significantly lower than what we would predict using our model. In fact, his actual career result can be viewed as a huge outlier. Using this to conclude that Federer is an underachiever seems slightly harsh... so I tested our model with another of the world's other top players:

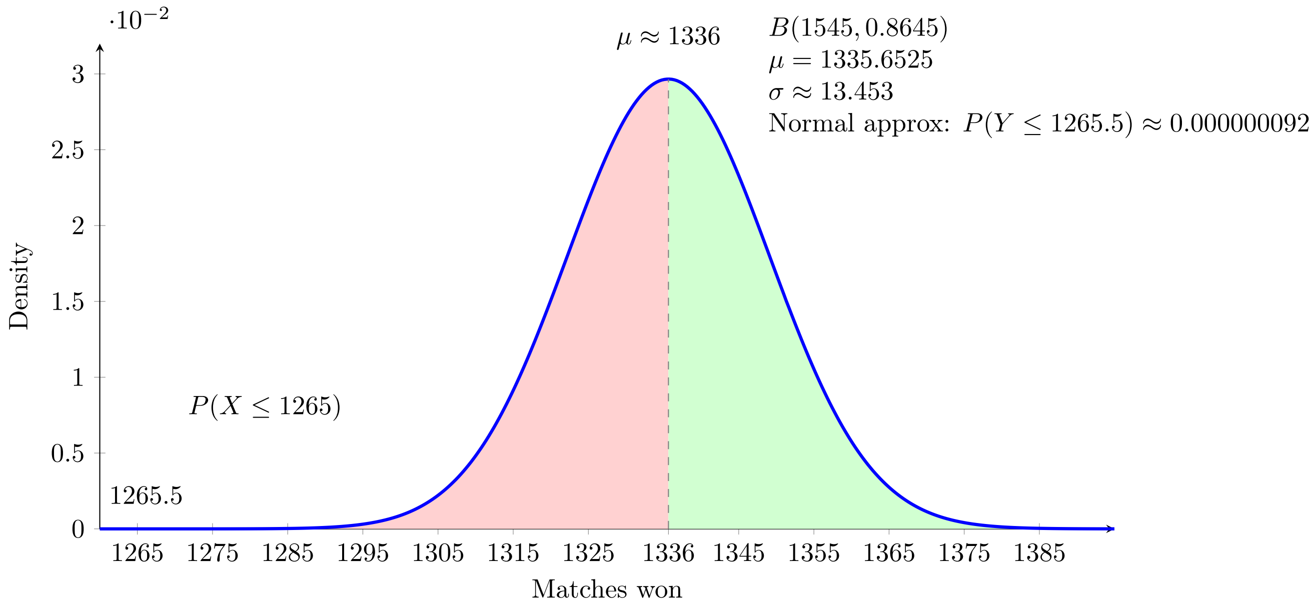

Figure 2: Normal approximation showing Federer's expected amount of wins

Novak Djokovic
Point Win Probability: 0.545
Calculated P(Win match): 0.8864
Matches Played: 1363
Expected Record: 1208–155
Actual Record: 1139–224

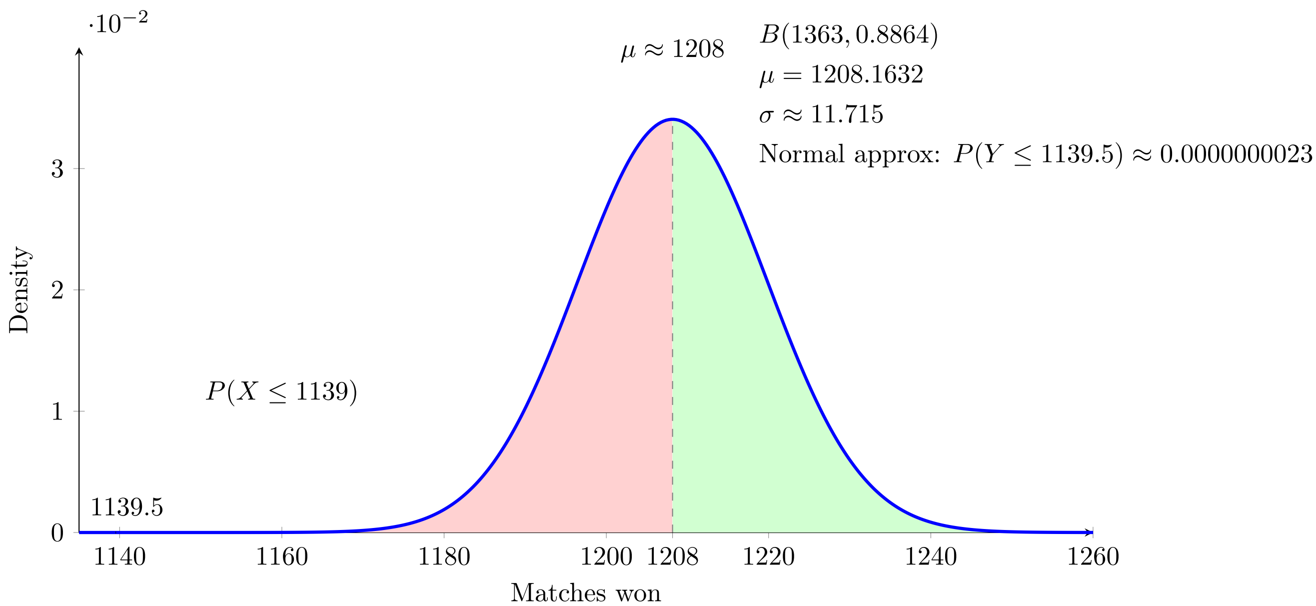

Figure 3: Normal approximation showing Djokovic's expected number of wins

As seen in Figure 3, Djokovic also severely underperforms when compared to the predictions of our model. I can confidently say that our model overestimates the number of wins a player should be getting as the point win probability deviates upwards from 0.50. However, the reason for this is unclear. To get a better idea of where this separation between our estimation and real results occurred, we can compare these player's actual game and set win results with our expected probabilities. This will help us get a better insight into which stage of our estimation has led to the most discrepancy compared to real life.

Figures 4 and 5 show that the biggest discrepancy between reality and our expectations comes in the "win match" stage of the graph. In Figure 4, we can see that Federer's set win probability is almost on par with his expected value, yet his match win probability takes a dip below our expectations. For Djokovic, he is increasingly below the expected results as the graph progresses, yet, the main discrepancy is in the match win stage.

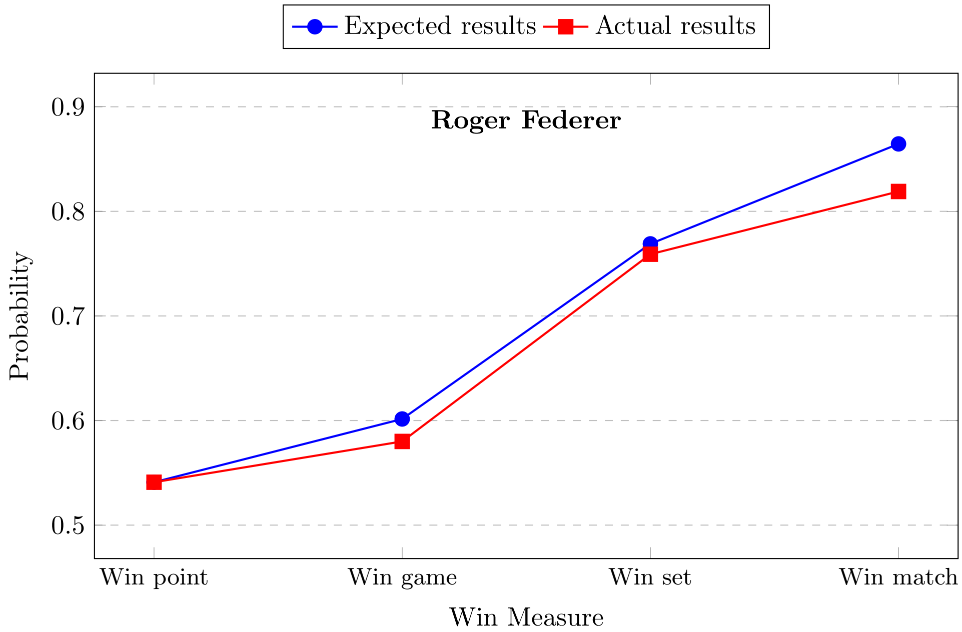

Figure 4: Comparison of Federer's expected and actual probabilities.

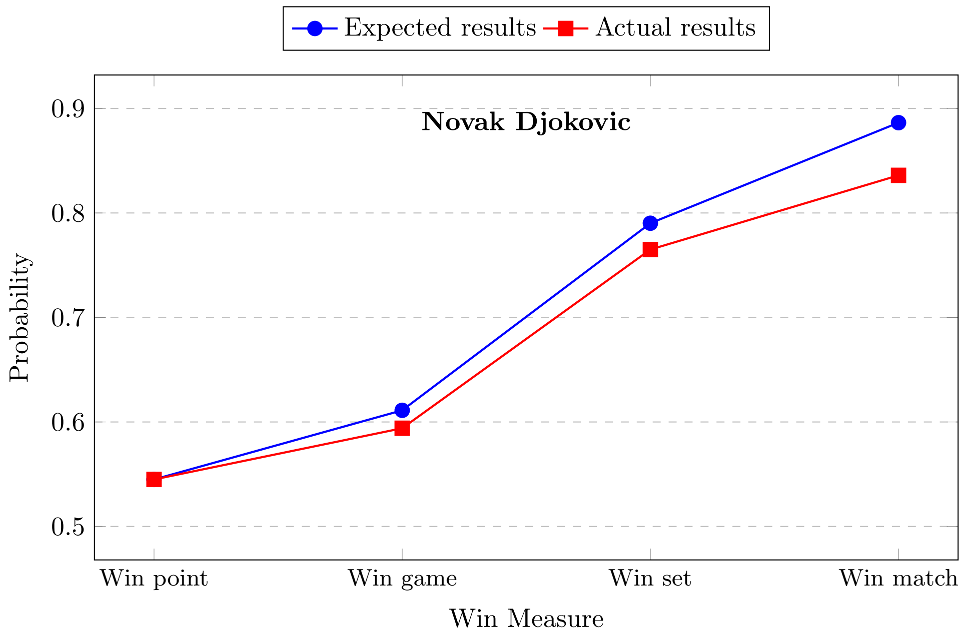

Figure 5: Comparison of Djokovic's expected and actual probabilities.

Why? This prompted some more in-depth look into some statistics over these player's careers - why are they not able to convert this more advantageous set win probability, to an equally as advantageous match win probability? The answer came in an underlying statistic called **Deciding Sets Won %**.

Deciding Sets: The recipe of downfall

A deciding set is the final (3rd) set, and is played when a player only wins one of the first 2 sets.

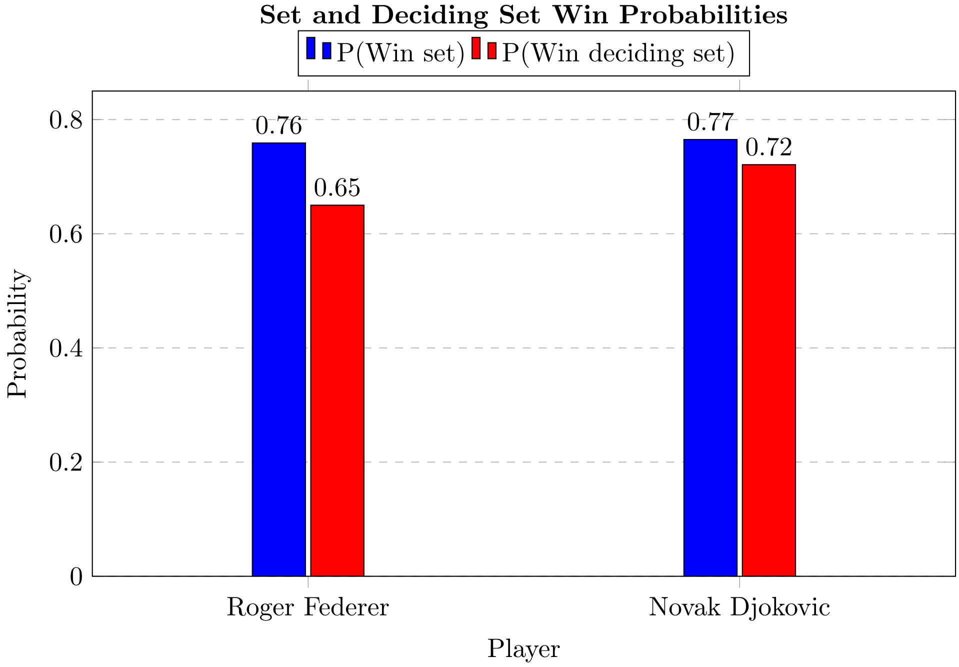

Figure 6: Comparison of Federer's and Djokovic's probabilities of winning a set and a deciding set.

Figure 6 shows both players perform significantly lower in deciding sets than they do in other sets. This can certainly be a part of the reason as to why our model has overestimated the amount of wins these players should have, but why does this occur?

I. They perform less well under pressure, or, their opponents perform better under pressure. While this seems unlikely, it is possible that worse opponents feel a higher drive due to their "nothing to lose" mindset, whereas top players experience the pressure of losing to a worse opponent.

II. Being the final set of the match, the deciding set is often played 3/4 hours after the match has begun. Therefore, especially later on in Federer/Djokovic's careers, younger players may hold an advantage due to their better fitness.

III. As per conditional probability, the chance of Federer/Djokovic winning a set given they have lost one of the previous 2, is a lot lower than simply the probability of them winning any set. Deciding sets also occur more often against better opponents, with higher win-rates themselves, so it does seem natural for top players to perform worse than usual under these circumstances.

Any overperformers?

Our model has shown to overestimate significantly the win-rate of top players. Do any players overperform our calculated expectations?

John Isner
Point Win Probability: 0.506
Calculated P(Win match): 0.5641
Matches Played: 809
Expected Record: 456-353
Actual Record: 489–320

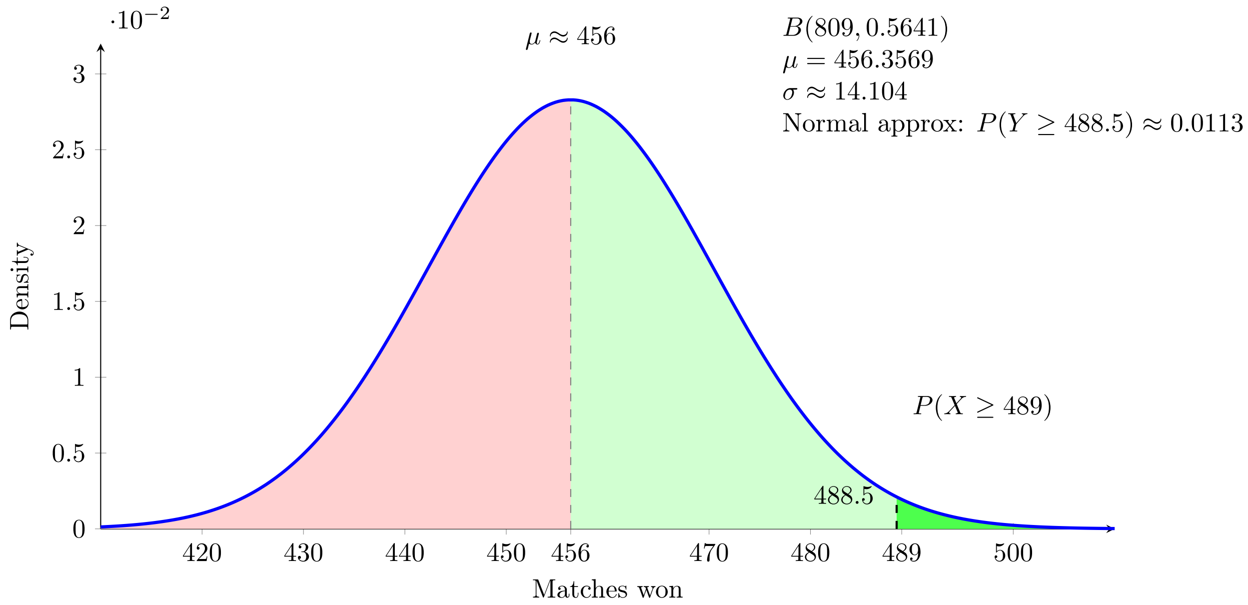

Figure 7: Normal approximation showing Isner's expected number of wins

Figure 7 shows that Isner has performed exceptionally well compared to the estimation made by our model. Isner is known for his service-based play style - often winning service games with ease, but struggling to break the opponent's serve. Doing this, he has defied our expectations. So what does this mean? Is Isner the GOAT?

Maybe. Maybe not. Underrated for sure.

A more sensible conclusion is that point win probabilities close to 0.50 will result in more accurate models.

Conclusion

We have explored how the probability of winning a single point affects the overall success of tennis players. Here are my conclusions:

I. Federer's 54% point win-rate is deceptive - it seems ordinary, but as backed up by the maths, leads to exceptionally high overall win percentages. So no, he indeed did not defy mathematics and statistics as it initially seemed. In fact, his 82% win-rate was an outlier on the wrong tail of the distribution!

II. Yes, 1% really does make all the difference in tennis. A point win probability difference of 1% can lead to swings of up to 10% in win percentages, potentially making the difference between making a decent living or making hundreds of millions from the sport.

III. The inaccuracies in our model are mainly caused by our assumption P(Win point) being an independent event, when, in reality conditional probability plays a huge role within tennis. Our model works better for average players. After a lot of thought, I believe the reason for this is that: top players play against a wider range of opponents within a given tournament, from weak opponents in the first round to extremely strong opponents as the rounds progress. This leads to a greater variance in P(Win point) across different matches, which, reduces the accuracy of our model as we model P(Win point) as a constant. Another way to put this would be that, strong players' very high point-win rates (eg. 0.70) against weaker opponents skews their average point-win rate, beyond what their match-win rate represents.

IV. Our model, portraying GOATs of tennis as underachievers, could not handle John Isner and his massive serves. Whether that is due to our model, or John Isner's brilliance, is up to you.

I would like to thank you, the reader, for your time and hope that you have enjoyed this tour we have taken through tennis and the power of probability behind it.