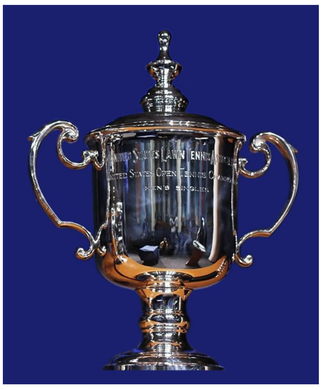
Do Summary Statistics Really Mean Anything in Tennis?

By Gautham Pasupathy • 01 Dec 2018 • 5 min read

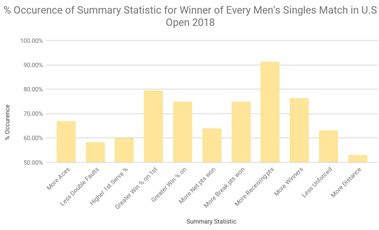
Four times a year, the entire tennis community gathers to witness one of the most spectacular events: The Grand Slams. Fans young and old crowd around laptops and televisions, host watch parties with friends, and some even witness the magic in person if lucky enough. From watching new upcomers like Denis Shapovalov and Borna Coric, to legends such as Roger Federer and Novak Djokovic, each match is always fun to watch. Of course, just like with every other sport, debating and betting about who will win each match in casual settings as well as in fantasy draws is a crucial part of the entire experience.

And in determining these winners, many people use blind intuition, flip a coin or pick the players they know. We all have that one friend who sees the 26th seed with a record for the most double faults and puts them in the finals of their fantasy draw. But for every ten people using this technique, there is one savvy person looking for data metrics and statistics to help them pick a winner. Tennis data can be obtained from various places, with little to no information of the accuracy of the data. Still, there are clearly reliable, verified sites such at the IBM SlamTracker data, which is the site from which this data was obtained. However, there is no real indication for the person trying to predict the winner of a match whether these statistics actually matter, or which statistics matter more than others.

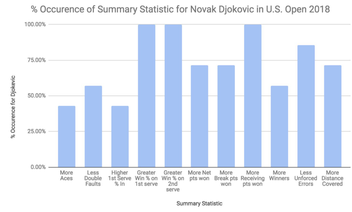


The 2018 Men's Tennis U.S. Open Trophy (https://www.pinterest.com/pin/350865418283972606/)

In this article, we look at match data for every single match in U.S. Open 2018 and analyze the summary statistics of the winner of every match to see if they are good predictors of winning matches. If they are, we can further analyze these summary statistics to see if we can find the most accurate predictor of winning matches.



This first graph is a bar graph of the % occurrence of each summary statistic in a winning match. The first important thing to take note of is the y-axis, where the percentages are displayed. What is very important to note is that it starts at 50%, meaning that there is always a greater percentage of matches where the winner is leading in every single summary statistic. This is very important, as it tells us that all of these summary stats are relevant and do actually factor in predicting a winner. Beyond that, we can take a look at each of the summary statistics, and see which one is the most accurate predictor in determining the winner of a match. Looking at the bar graph, it is clear that "more receiving points won" is the most important category (91.34%). Obviously this makes sense; most professional tennis players hold serve with relative ease, whereas "breaking serve", or winning the serve games of opponents, is what changes the actual outcome of the match. Another important measurement to note is the percent occurrence of the aggressive versus the defensive measurements, and what that says about the data. For example, when comparing the aggressive measurements such as the "more aces" and "more winners" categories and the defensive measurements such as "less double faults" and "less unforced errors", it is clear that a greater percentage of the winning matches contain aggressive statistics than defensive statistics. This is very important, because it contradicts the "defense is the best Offense" theory. It suggests the importance of being a more aggressive player, and gives way to a range of possible topics to explore, in the context of the benefits of an offensive or a defensive playing style. We can analyze these summary statistics further, as we take a look at the filtered version showing results for only Novak Djokovic, the winner of the 2018 U.S. Open.



Of course, Novak Djokovic, as the winner of the tournament, would have won all of his matches: because of this, his data would follow the data for the summary statistics (which we have determined as a good predictors to some extent). Djokovic's data presents an interesting contrast to the previous observations: his defensive qualities show up more. The 3 most offensive statistics (more aces, higher 1st serve % in, and more winners) show up the least in all the metrics, which confounds with the data accumulated from the winner of every single match. We can take the future research options we were looking into before (the usefulness of offense vs. defense) and narrow our focus, seeing if there is a difference between general victors and the champions of each tournament in terms of playing style.

Novak Djokovic, champion of the 2018 U.S Open (https://www.timesdive.co.za/sport/2018-06-25-as-not-one-of-wimbledon-favourites-insists-novak-djokovic/)

Overall, it is clear that the data being collected for the matches, rather than just a measure of a given player in a given match, allows us to make some sort of an insightful decision on the winner of a match. Using the sample size of one tournament, of course, is too small to see how accurate of a predictor the summary statistics are, but it does provide an indication that these summary statistics can be useful when drawing inferences. In the future, we have many new avenues of research that open up to us. We could expand the observations to include every single tournament of 2017, and see if those summary statistics played a factor in shaping the results of 2018. We could also compare general winners to the tournament champion, and see if winning the tournaments comes from a different playing style or outperforming the other players in the same playing style. At least we know there is one thing to take note of: listening to that one friend who has a gut feeling that the player who has the most double faults will win the tournament isn't such a great idea.
