
ISyE 2028 – Basic Statistical Methods - Fall 2015

Bonus Project: “Big” Data Analytics

Final Report

As an avid tennis fan, many of the discussions I have about tennis are about who is better between Rafael Nadal and Roger Federer. Most of the time somebody will decide who is better based on who has the style they can relate to more. Rafael Nadal is known for his speed and tremendous forehand while Roger Federer is known more for his fluid playing style and calm demeanor. For the record I’m a huge Rafael Nadal fan. Fans of either player will often cite statistics that support the player they prefer. Often times those statistics boil down to their head to head record, which Nadal owns a 23-10 advantage over Federer, or to the number of Grand Slams they have collected, Federer leads 17 Slams to 14. I wanted to take a closer look at something that is baked into both of those stats: the difference that the court surface has on how players of their ilk perform.

One result of their difference in playing styles is how they perform on grass courts and clay courts. Rafael Nadal is unanimously agreed to be the best clay court player of all time. He has won an astonishing 93 percent of his matches in the French Open and at the 3 clay court Masters events. He has won 9 French Open titles including the first time he played the tournament when he was still a teenager. Roger Federer has had similar success on grass courts. He shares the record for the most Wimbledon titles (7) and has been to more finals than anybody else (10). They have each beaten the other on their favorite surface (and even once played a match on a court that was half grass and half clay!), but their performance on hard courts is much closer.

Each surface affects the ball differently upon contact and can provide limitations to how the players can move on the surface. For instance, clay courts are made of crushed bricks, so when the ball makes contact with the surface there is much more friction and it is slowed down significantly. Another effect of the surface is that players are able to slide into their shots, which allows them to smoothly run from one ball to the next. Grass courts have an almost complete opposite effect on players because they cannot slide to their shots (although some players still try to) and the grass causes the balls to skid through the court. However, hard courts are somewhere in the middle. The reason why this is so important is that tournaments are played on hard courts more than any other surface and two of the four Slams are played on hard. Therefore if hard courts are shown to favor players that are successful on grass, then that could mean that Federer should be expected to have won more Slams than Rafa simply because of the way the tournaments are distributed. I think that hard courts are more likely to favor grass players because typically grass court players have strong serves, which translates very well to hard courts. This gave me an idea. If I were to consider Rafael Nadal as a clay courter and Roger Federer as a grass courter, then would I be able to gain some insight into how unprecedented their performance has been on hard courts?

There were a couple ways I wanted to address this question. The first way was to identify players that had been successful at Wimbledon and the French Open. I accomplished this by web scraping data from tennis28.com. I got a list of players that had been to multiple French Open finals and a list of players that had been to multiple Wimbledon finals. My next step was to look at how many times they had been to the semis or finals of the US Open or Australian Open. I

found that players that had been to multiple Wimbledon finals made it to roughly 4.5 hard court finals while players that had been to multiple French Open finals only made it to 1.8 hard court finals. If I looked at medians instead of means I found that players in the Wimbledon category went to 4 finals and the French Open players went to 1 final. When I broadened my view to semifinals I found the same dynamic. Players that preferred Wimbledon were making it to 4.1 semifinals with a median of 3, and players that preferred the French were making it to an average of 1.84 finals with a median of 0. This seemed to support my suspicion that players that were better on grass courts were more likely to succeed on hard courts. However, I still wanted to do some analysis on these values. In order to find if my data was significant I knew I should calculate a p-value. I decided to compare the data utilizing hypothesis testing of two populations with the assumption that the variances were equal and unknown. My null hypothesis was that the difference was zero and my alternative hypothesis was that the average for the grass players was higher than the clay players. My test statistic yielded a p-value of 0.0137, which was lower than my alpha value of 0.05. This meant I could reject my null hypothesis and accept my alternative hypothesis that players that made multiple Wimbledon semifinals were more likely to make hard court semifinals.

Next I wanted to expand my view from counting the times they reached the semifinals to win loss percentage. I felt this could be helpful because it would be a better indicator of how the players performed on a match-to-match basis. Also, this meant I had to come up with a new list of players to collect data on. At first this seemed like a tall order because I struggled to find a data set that would allow me to accurately calculate these numbers. However, I had a stroke of luck when I was reading some articles online and they cited where they got their data. Low and behold I found a folder of csvs that contained match statistics from every match since 1968. I had struck a gold mine. In my list of players I included all time greats from the 70's all the way up to players that may not make the hall of fame, but are competing at a high level right now. After I came up with these names I wrote a program that would sift through the csvs and give me their wins and losses on each surface. Then I had to determine whether or not these players were in fact clay or grass players. The way I did that was by seeing if their highest win percentage was on clay or grass, then seeing if it was above 65%. I calculated both their win percentage in every tournament they played as well as their win percentage in Slams and Masters events. In the end Slams and Masters events are probably more indicative of how good the players are on hard courts because those are the tournaments that all the top players enter. That means that the players are less likely to pad their victory count against inferior opponents. I decided to use that number moving forward. As for 65 percent I saw that as a good indicator that the player was really a top-level guy.

In the end I had 15 players I considered as grass players and 16 players that I considered clay players. The grass players had a win percentage of 65.9% on hard courts while clay players were at 61.8%. I had two ideas for how to determine if the data was significant and both involved using hypothesis testing to compare two populations. My null hypothesis was that the difference in means was 0 and my alternative hypothesis was that the clay players' percentage is lower than the grass players'. The goal here was to determine if this difference in means was significant. My first idea was to pool all of the clay players' results and the grass players' results then compare to get a p-value using the population proportions method. My p-value in this case was 0.0025, which was lower than my alpha value of 0.05. However, I was a little bit suspicious of this result because I had some doubts about the efficacy of pooling the results. This would introduce the possibility of one player having a large effect on the result if he simply played more hard court tournaments than the others. After looking at my data I found that was not really the case, but I wanted to repeat the calculations using a different assumption. Next I

wanted to calculate a p-value without pooling the players and assuming that the variance was equal and unknown. I knew that when I used this method my sample size would decrease drastically, but I was curious if my data would still be significant. When I calculated in this way I had a p-value of 0.1337, which is much higher than my alpha value of 0.01. This made me question if my data was in fact significant. At this point I had exhausted the ideas I had for how to address my question about Rafael Nadal and Roger Federer, but I felt that if I kept trying different things I would likely just be trying to prove my own beliefs.

In the end my venture into big data analytics yielded mixed results. I started off hoping to be able to find some proof that hard courts favored players that had playing styles better suited for grass courts. Then I could hopefully tie that back into the Federer versus Nadal rivalry. My calculation where I found counted the Grand Slam or Masters semifinals yielded results that supported my hypothesis. The p-value in this case was sufficiently low enough to reject my null hypothesis. In order to attempt to tie that result back to Federer and Nadal I looked at how many standard deviations they were beyond the sample means. Federer was around 4 standard deviations higher than the norm while Nadal was a little above 2 (interestingly Nadal's performance of making it to 5 Wimbledon semifinals was 10 standard deviations above the norm!). That data was a good start, but I did not feel it had quite the strength I was looking for. When I tried to look at overall performance at Slams and Masters I had some more difficulty. Even though these tournaments feature the world's best players, it is still possible that in the first two rounds the players in my dataset were facing players they were significantly better than. If they were to win those first two matches and then lose, they would have a win percentage of 66%. That leads me to believe that semifinal performance is probably a better way of answering my question. Although that does support my claim that hard courts favor players with a grass court style, it is still likely too much of a stretch to apply that to the Rafael Nadal and Roger Federer rivalry. I repeated the standard deviations above the norm for Nadal and Federer. Rafael Nadal's win percentages were 1.7, 3.2, and 1.3 standard deviations above the norm on hard, clay and grass respectively (his grass court number is skewed because he has entered a few Wimbledon draws when he was hurt and lost in the first or second round). Roger Federer had results of 1.9, 2.0, and 2.4 above the norm.

I knew going into this that sports analytics are very difficult because while one can have great ideas before going into calculations, often times it is very easy to get bogged down in the details. I started off with some pretty big goals, but I think as I moved on I really tried to narrow my focus. While somebody else may not benefit from that finding, I feel this will certainly help me tackle similar problems in the future. If I were to improve my data I would probably have to incorporate regression. In order to do this I would have to do more data analysis to find what number I could look at that that was most correlated to success at different surfaces. Then I would want to look at how the clay court players ranked in those numbers. As I said earlier that sounds somewhat easy, but that could take me months to sort through. In the end I'm not sure if somebody could use my data to settle the ongoing debate between Rafael Nadal and Federer, but one can probably look at it and conclude that these guys really are pretty good.

I just want to take a second to thank Professor Xie for letting us do this project. It was really a blast to get to apply the things we learned in class (along with things we learned in other classes) to help answer a question we found interesting. I'm a little embarrassed to admit how much time I spent tinkering with this data to really make sure I wasn't making decisions just to support my preconceived notion that Rafael Nadal was better. I think that it is a pretty important lesson in scenarios like this. Thanks again.