Mini-project: Stochastic Tennis

Adams Zequi Mohammed 20131887

18th December,2020.

Section 1 Introduction: Stochastic processes and stochastic tennis

1.1 Stochastic Processes

Generally speaking,stochastic processes use mathematical models to describe random events and how they change through time.We model stochastic process of random events not only as random occurences but model them as they move with time and in space. Stochastic processes are used extensively to model various phenomena in different fields of academia.A few of these are finance,economics and operations analysis.

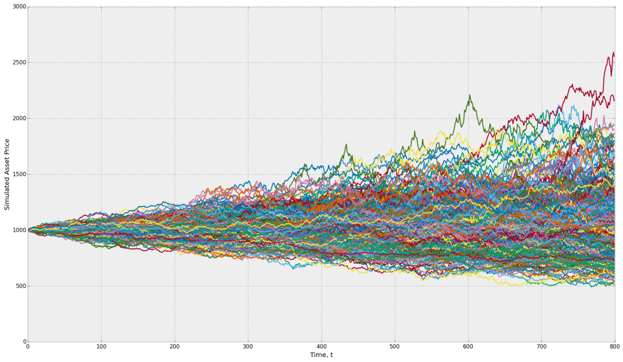


Figure 1: An example of application of Stochastic process to finance

Mathematically, stochastic processes can be defined by :

$$X(t) : t \in T$$

where a random event X(t) occurs at time t in the set T.There are several types of stochastic process(Discrete time process,Continuous time process,Martingale or fair game process,Point Process etc.) but we only define the most fundamental types of stochastic processes which are Discrete and Continuous time processes.

A stochastic process is referred to as:

$$Discrete\ time\ if\ the\ set\ T\ is\ finite\ , where\ T = [0, 1, 2, 3, ...]\ and$$

$$X(t) : t \in T\ are\ independent\ random\ variables.$$

A stochastic process is referred to

$$continuous\ time\ if\ the\ set\ T\ is\ infinite\ , where\ T = [0, \infty],$$

and

$$X(t) : t \in T\ are\ independent\ random\ variables.$$

$$In\ a\ continuous\ stochastic\ process\, X(t)\ changes\ at\ all\ times$$

$$including\ but\ not\ limited\ to\ t = 0, 1, 2, ...$$

1.2 Stochastic Tennis Model

The stochastic tennis model was developed by Paul K. Newton and Joseph B. Keller in their paper 'Probability of Winning at Tennis I' in 2005.Assuming that ina 2 player game,a particular player's probability of winning a point are identical and independently distributed,the probability of winning a game during a specific match are based on the probability of winning that point.

Let probability of winning a point for player 1 be p ie.

$$Pr(P1\ wins) = p$$

and probability of winning a point for player 2 be 1-p or q ie.

$$Pr(P2wins) = 1 - p = q$$

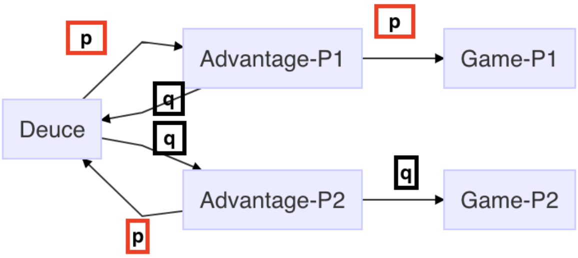


Figure 2: Possible movements of tennis games from state to state

By analysing Figure 2 we can see that from Deuce ,P1 can move to Advantage-P1 with probability of p and from Advantage-P1 ,P1 can move back to Deuce or win the game therefore P1 has a probability of

$$p * p = p^2$$

of winning the game and probability of not winning with $p * q$ and because player 2 has the same state movements player 2 has a probability of

$$q * q = q^2$$

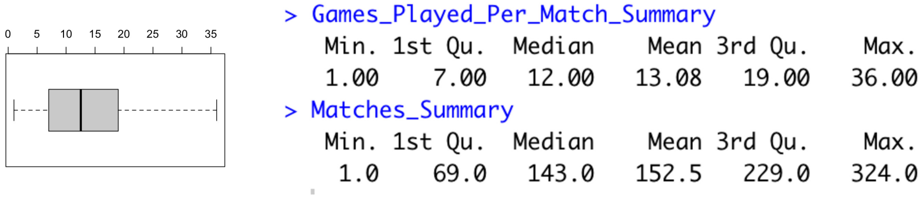
of winning the game and $q * p$ of not winning the game.The probability that neither P1 or P2 wins the game is

$$2pq$$

.

Section 2 Exploratory Data Analysis

2.1 Summary Statistics



```
> Games_Played_Per_Match_Summary
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    7.00   12.00   13.08   19.00   36.00
> Matches_Summary
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    1.0    69.0   143.0   152.5   229.0   324.0
```

Figure 3: Boxplot of Quartile distribution and correponding Summary Statistics of our dataset

From Figure 3 ,We can see from the dataset using summary statistics that a maximum of 324 matches were played.On average Venus has played an average of 13.08 games.Interestingly,from the boxplot of quartiles,we notice that the games that venus played are skewed to the lower half of the data with only a few games exceeding the 3rd quartile which is 19.

2.2 Analysis of Venus Williams vs Serena Williams

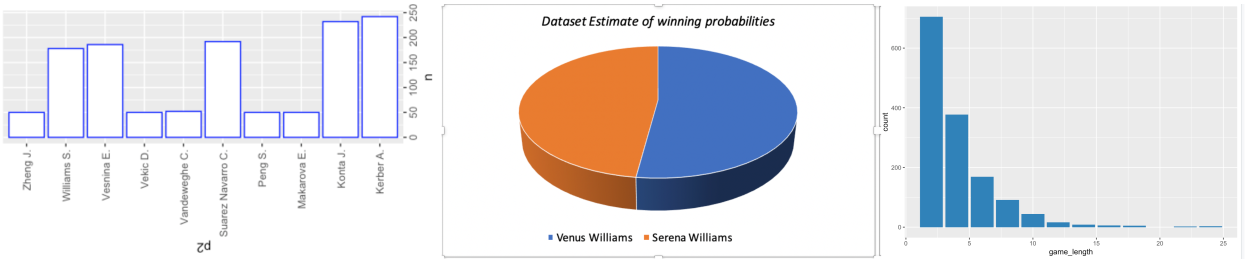


Figure 4: Top 5 highest and Lowest points played against a player P2

Our Analysis is going to be based on analysing how Venus Williams fares against other players so we set Venus Williams as player1 (P1) and other players,especially Serena Williams as P2.

From Figure 4, we see the top 5 highest and lowest number of points Venus has played against a particular player P2.Graphically we can see that the lowest points revolve around a particular point,50 ,while the highest points vary.We can also see from the corresponding pie chart the probability of winning or losing against her sister which are

$$p = 0.521$$

and

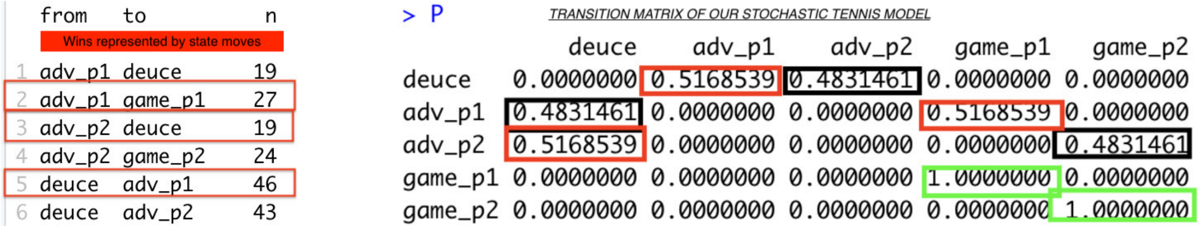
$$q = 1 - p = 0.479$$

.The probabilities are calculated by counting the number of games that ended with Venus winning and number of games that ended with Serena winning,and finding the respecting average of those by multiplying the two.

We also observe from the rightmost picture in Figure 4 that games generally last for a short period of time with the most number of games lasting between 0 and 10 and least number of games lasting between 10 and 25.The duration of the game was found by grouping by unique_id which was found by selecting the maximum number of points played before someone was ultimately found victorious ,selecting the maximum point among those and adding 1.

Section 3 Estimating the transition probabilities

3.1 Model Parameters



	from	to	n
	Wins represented by state moves		
1	adv_p1	deuce	19
2	adv_p1	game_p1	27
3	adv_p2	deuce	19
4	adv_p2	game_p2	24
5	deuce	adv_p1	46
6	deuce	adv_p2	43

TRANSITION MATRIX OF OUR STOCHASTIC TENNIS MODEL

```
> P
            deuce    adv_p1    adv_p2   game_p1   game_p2
deuce   0.0000000 0.5168539 0.4831461 0.0000000 0.0000000
adv_p1  0.4831461 0.0000000 0.0000000 0.5168539 0.0000000
adv_p2  0.5168539 0.0000000 0.0000000 0.0000000 0.4831461
game_p1 0.0000000 0.0000000 0.0000000 1.0000000 0.0000000
game_p2 0.0000000 0.0000000 0.0000000 0.0000000 1.0000000
```

Figure 5: Sum of State movemnts and Transition matrix

The probability of winning or losing a point for either Venus or Serena is determined by calculating sum of movement between each states for either player which is represented by "Wins represented by each move".The sums have been highlighted red in the figure ie.27,19 and 46.

These sums are then used to calculate the probabilities of transition by dividing them by the total sums.These probabilities are shown in our probabiility transition matrices with red being Venus wins(p) and black being Serena wins(q).

3.2 Outcome and Notes of Model

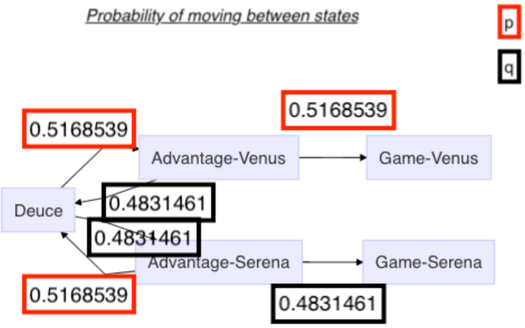


Figure 6: Probability of moving between states

This diagram represents the state transition probabilities and we can clearly see that the probability that venus wins against serena is higher.

From our introduction of the stochastic tennis model, we can then identify the probability that any player P1 with similar probabilities in the figure would be victorious with the probability :

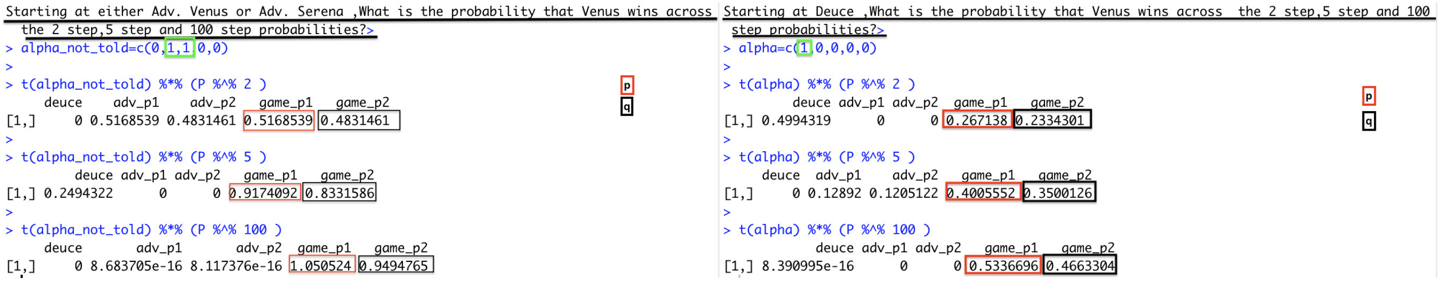
$$0.5168539 * 0.5168539 = 0.2671379539$$

and any player P2 with similar probabilities would be victorious with probability

$$0.4831461 * 0.4831461 = 0.2334301539$$

Section 4 Stochastic tennis results

4.1 Estimating Win probabilities



```
Starting at either Adv. Venus or Adv. Serena ,What is the probability that Venus wins across
  the 2 step,5 step and 100 step probabilities?>
> alpha_not_told=c(0,1,1,0,0)
>
> t(alpha_not_told) %*% (P %^% 2 )
      deuce     adv_p1    adv_p2   game_p1   game_p2
[1,]      0 0.5168539 0.4831461 0.5168539 0.4831461
>
> t(alpha_not_told) %*% (P %^% 5 )
          deuce adv_p1 adv_p2   game_p1   game_p2
[1,] 0.2494322      0      0 0.9174092 0.8331586
>
> t(alpha_not_told) %*% (P %^% 100 )
      deuce       adv_p1       adv_p2  game_p1   game_p2
[1,]      0 8.683705e-16 8.117376e-16 1.050524 0.9494765
```

```
Starting at Deuce ,What is the probability that Venus wins across  the 2 step,5 step and 100
 step probabilities?>
> alpha=c(1,0,0,0,0)
>
> t(alpha) %*% (P %^% 2 )
         deuce adv_p1 adv_p2  game_p1   game_p2
[1,] 0.4994319      0      0 0.267138 0.2334301
>
> t(alpha) %*% (P %^% 5 )
     deuce  adv_p1    adv_p2   game_p1   game_p2
[1,]     0 0.12892 0.1205122 0.4005552 0.3500126
>
> t(alpha) %*% (P %^% 100 )
            deuce adv_p1 adv_p2   game_p1   game_p2
[1,] 8.390995e-16      0      0 0.5336696 0.4663304
```

Figure 7: Probabilities that Venus Wins calculated from 2 step ,5 step and 100 step transition matrices

From Figure 7 above,the alpha and alpha_not_told variables define the initial distribution that we have to begin from.

$$1 = Represented\ by\ states\ to\ begin\ from.$$

$$0 = Represented\ by\ states\ to\ that\ are\ not\ being\ accounted\ for.$$

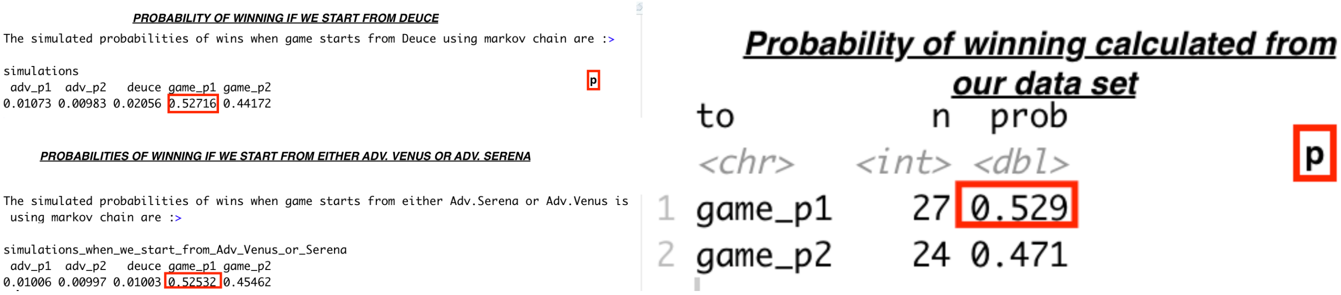
We calculate three probability sets using the 2nd step , 5th Step and 100th step transition probabilities.The t-step probabilities are calculated by multiplying the transition matrix (P) by our alpha values.

$$t - step\ state\ transition\ probabilities\ = alpha\ *\ Probability - matrix^n$$

From these probabilites we realise that the probability that Venus wins is always slightly higher than the probabilities that Serena wins regardless of which state we start in.

We also notice that when we start from Advantage Venus or Advantage Serena,the probability that Venus wins gets closer to calculated probability from our dataset as the steps get lower while the probability that Venus wins gets closer to calculated probability from our dataset as the steps get higher when we start from deuce.

4.2 Simulating our results



PROBABILITY OF WINNING IF WE START FROM DEUCE

```
The simulated probabilities of wins when game starts from Deuce using markov chain are :>

simulations
 adv_p1  adv_p2   deuce game_p1 game_p2
0.01073 0.00983 0.02056 0.52716 0.44172
```

PROBABILITIES OF WINNING IF WE START FROM EITHER ADV. VENUS OR ADV. SERENA

```
The simulated probabilities of wins when game starts from either Adv.Serena or Adv.Venus is
 using markov chain are :>

simulations_when_we_start_from_Adv_Venus_or_Serena
 adv_p1  adv_p2   deuce game_p1 game_p2
0.01006 0.00997 0.01003 0.52532 0.45462
```

Probability of winning calculated from our data set

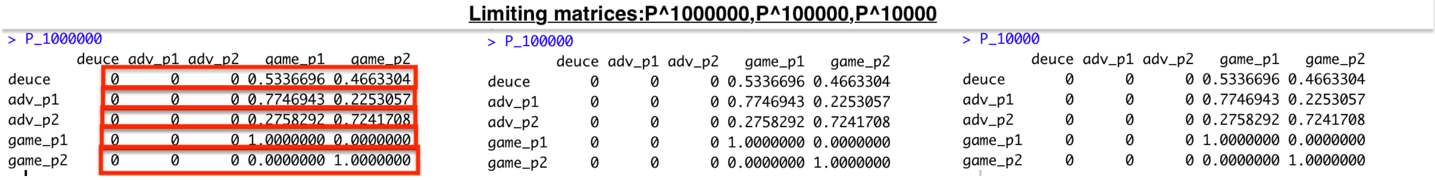
```
  to          n  prob
  <chr>   <int> <dbl>
1 game_p1    27 0.529
2 game_p2    24 0.471
```

Figure 8: Simulation probabilities when we begin at deuce and when we begin at either Advantage Venus or Advantage Serena

The right side of Figure 8 represents two markov simulations that were run to determine the probabilities of winning for either Serena or Venus.When we begin from Deuce ,the probability that Venus wins is similar to the probability that she wins calculated from our dataset(0.52716 versus 0.529) .When we being from Advantage Venus or Advantage Serena , we get similar results compared to the probability we got fromour datatse(0.52535 versus 0.529)

Section 5 Extend the results of our Stochastic tennis model

5.1 Long term behaviour of the model



Limiting matrices:P^1000000,P^100000,P^10000

```
> P_1000000
        deuce adv_p1 adv_p2   game_p1   game_p2
deuce       0      0      0 0.5336696 0.4663304
adv_p1      0      0      0 0.7746943 0.2253057
adv_p2      0      0      0 0.2758292 0.7241708
game_p1     0      0      0 1.0000000 0.0000000
game_p2     0      0      0 0.0000000 1.0000000
```

```
> P_100000
        deuce adv_p1 adv_p2   game_p1   game_p2
deuce       0      0      0 0.5336696 0.4663304
adv_p1      0      0      0 0.7746943 0.2253057
adv_p2      0      0      0 0.2758292 0.7241708
game_p1     0      0      0 1.0000000 0.0000000
game_p2     0      0      0 0.0000000 1.0000000
```

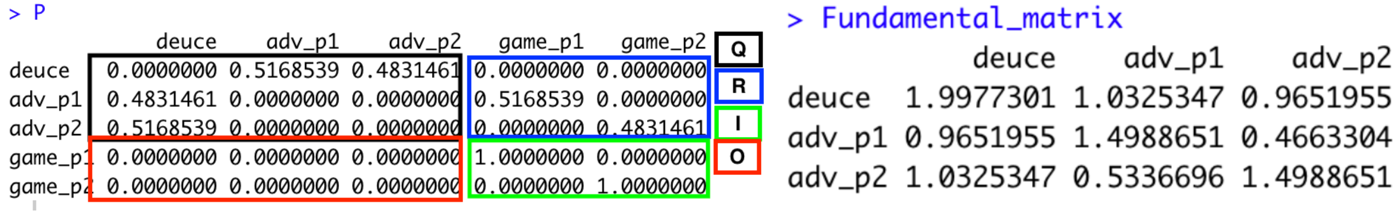
```
> P_10000
        deuce adv_p1 adv_p2   game_p1   game_p2
deuce       0      0      0 0.5336696 0.4663304
adv_p1      0      0      0 0.7746943 0.2253057
adv_p2      0      0      0 0.2758292 0.7241708
game_p1     0      0      0 1.0000000 0.0000000
game_p2     0      0      0 0.0000000 1.0000000
```

I analyse the long term behaviour of the model.With respect to the limiting distribution, we can test whether or not a transition matrixhas a limiting distribution by taking higher powers of the transition matrix as it approaches infinity.

$$\lim_{i->inf} P_n^{ij} = \lambda_j$$

From the limiting matrices, we see that the long-term behavior of our stochastic tennis model depends on the initial state, therefore cannot not have a limiting distribution. We can however have a limiting matrix as it does not require the rows to be the same as this is the requirement for a limiting distribution.We can also describe this distribution as not being unique.

5.2 Absorption Probabilities



```
> P
            deuce    adv_p1    adv_p2   game_p1   game_p2
deuce   0.0000000 0.5168539 0.4831461 0.0000000 0.0000000
adv_p1  0.4831461 0.0000000 0.0000000 0.5168539 0.0000000
adv_p2  0.5168539 0.0000000 0.0000000 0.0000000 0.4831461
game_p1 0.0000000 0.0000000 0.0000000 1.0000000 0.0000000
game_p2 0.0000000 0.0000000 0.0000000 0.0000000 1.0000000
```

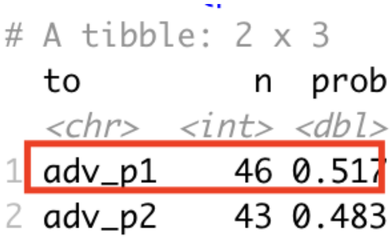
```
> Fundamental_matrix
           deuce    adv_p1    adv_p2
deuce  1.9977301 1.0325347 0.9651955
adv_p1 0.9651955 1.4988651 0.4663304
adv_p2 1.0325347 0.5336696 1.4988651
```

Figure 9: Canonial form of Transition Matrix and Fundamental matrix

We find the canonical form of the transition matrix P which is equal to the matrix P.From the canonical form we deduce the submatrices as highlighted by different colours.I find the fundamental matrix which gives us absorption probabilities after moving spending an expected number of time in a particular transient state.

From our probabilities ,we notice that if we start from Deuce,the expected number of times in deuce,adv_p1 and adv_p2 before entering an absorption state is 1.9977301 1.0325347 and 0.9651955 respectively

We can also make similar observations for expected number of times a player will remain in a transient state from adv_p1 and advp2 to absorption states.There are 0.9651955,1.4988651,0.4663304 and 1.0325347,0.5336696,1.4988651 respectively.



```
# A tibble: 2 x 3
  to         n  prob
  <chr>  <int> <dbl>
1 adv_p1    46 0.517
2 adv_p2    43 0.483
```

Figure 10: Venus and Serena's probability of hitting first point

I also discover that Venus has a higher probability of hitting the first point with a p=0.571 leaving Serena a probability of 0.483 of hitting the first point.

Section 6 Conclusions

In both simulation and data analysis ,Venus wins the game against her sister with the probability of 0.52 to 0.53 where using our data Venus wins by 0.521 and by simulation Venus wins by 0.52716

I began by exploring the stochastic tennis data and finding some descriptive statistics about the data.I found that a maximum of 324 matches of Venus versus another player were played, and out of those matches Venus played an average of 13.08 games.I also observed that Venus rarely exceeded 19 games during a match which shows that her match with the highest number of games is an outlier.

I aslo discuss the probability of Venus winning against her sister,Serena, from our data ,which is $p = 0.521$.I also make the observation that the duration of majority of games are between 0-10 with the highest games being 25.

I find the probability of winning a point by Venus against her sister which is p= 0.5168539 and probability of winning a point by Serena which is q= 0.4831461.I use this information to populate my transition matrix P.

Using t-step transition matrices ,we observe that Venus wins is always slightly higher than the probabilities that Serena wins regardless of which state we start in.When we start from Advantage Venus or Advantage Serena,the probability that Venus wins gets closer to calculated probability from our dataset as the steps get lower while the probability that Venus wins gets closer to calculated probability from our dataset as the steps get higher when we start from deuce.

I later run 2 simulations to compare to the results found directly from the dataset.I find that when we begin from Deuce ,the probability that Venus wins is similar to the probability that she wins calculated from our dataset and when we being from Advantage Venus or Advantage Serena , we get similar results compared to the probability we got from our datatset.

I observe that the transition matrix has no limiting distribution and therefore not unique but has a limiting matrix.

We found the absorption probabilities using the fundamental matrix and also found that Venus has a higher probability of hitting the first point.