USING MICROSOFT EXCEL TO MODEL A TENNIS MATCH

Tristan J. Barnett and Stephen R. Clarke
School of Mathematical Sciences
Swinburne University
PO Box 218
Hawthorn
Victoria 3122, Australia

tbarnett@groupwise.swin.edu.au

sclarke@groupwise.swin.edu.au

Abstract

This paper demonstrates the use of Microsoft Excel to generate the probability of winning and mean length of the remainder of the match conditional on the state of the match. Previous models treat games, sets and matches independently. We show how a series of interconnected sheets can be used to repeat these results. We also set up a sheet which can give the required statistics using the full match, set and game score as the state. The exercise could form an interesting and useful teaching example, and allow students to investigate the properties of tennis scoring systems.

1 Introduction

Many papers have investigated the characteristics of various scoring systems, particularly in racquet sports. The game of Lawn Tennis has a history of changing formats. Although it uses a basic nested system of points, games, sets and matches, there is much variation within that basic structure. Matches can be one set, or best of three or five sets, sets can be first to six, eight or ten games, advantage or non advantage, and games can be the traditional four point advantage, or seven or ten point tiebreakers. We even have the recent innovation at the 2001 Australian Open mixed, where the third set becomes a single tiebreaker game. This innovation is extending into the Victorian Tennis Association pennant this year.

The basic principles involved in modelling a tennis match are well known, and a Markov chain model with a constant probability of winning a point was set up by Schutz [4] . While such a model is acceptable within a game, a model which allows a player a different probability of winning depending on whether they are serving or receiving is essential for tennis. Statistics of interest are usually the chance of each player winning, and the expected length of the match. Croucher [1] looks at the conditional probabilities for either player winning a single game from any position. Pollard [3] uses a more analytic approach to calculate the probability for either player winning a game or set along with the expected number of points or games to be played and their corresponding variance.

Most of the previous work uses analytical methods, and treats each scoring unit independently. This results in limited tables of statistics. Thus the chance of winning a game and the expected number of points remaining in the game is calculated at the various scores within a game. The chance of winning a set and the expected number of games remaining in the set is calculated only after a completed game and would not show for example the probability of a player's chance of winning from three games to two, 15-30.

This paper discusses the use of Microsoft Excel to repeat these applications using a set of interrelated spreadsheets. This allows any probabilities to be entered and the resultant statistics automatically calculated or tabulated. In addition, more complicated workbooks can be set up which allow the calculation of the chance of winning a match and the expected number of points in the remainder of the match at any stage of the match given by the match, set and game score. These allow the dynamic updating of player's chances as a match progresses.

2 Simple Model

We explain the method by first looking at the simple model where we have two player's, A and B, and player A has a constant probability p of winning a point. We set up a Markov chain model of a game where the state of the game is the current game score in points (thus 40-30 is 3-2). With probability p the state changes from a, b to $a+1, b$ and with probability $1-p$ it changes from a, b to $a, b+1$. Thus if $P(a, b)$ is the probability that player A wins when the score is (a,b), we have:

$$P(a, b) = pP(a+1, b) + (1-p)P(a, b+1)$$

The boundary values are $P(a, b) = 1$ if $a = 4$, $b \leq 2$, $P(a, b) = 0$ if $b = 4$, $a \leq 2$. The boundary values and formula can be entered on a simple spreadsheet. The problem of deuce can be handled in two ways. Since deuce is logically equivalent to 30-30, a formula for this can be entered in the deuce cell. This creates a circular cell reference, but the iterative function of Excel can be turned on, and Excel will iterate to a solution. Alternatively, it is easily shown that the chance of winning from deuce is:

$$\frac{p^2}{p^2 + (1-p)^2}$$

Entering this formula removes the need for iteration. Table 1 shows the results obtained, using a value of $p = 0.54$. It indicates that a player with a 54% chance of winning a point has a 60% chance of winning the game. Note that since advantage server is logically equivalent to 40-30, and advantage receiver is logically equivalent to 30-40, the required statistics can be found from these cells.

The mean number of points in the remainder of the game is handled similarly as shown in Table 2. All boundary points become 0, the recurrence relation becomes:

$$M(a, b) = 1 + pM(a+1, b) + (1-p)M(a, b+1)$$

The formula for deuce is:

$$\frac{2}{p^2 + (1-p)^2}$$

		A score				
		0	15	30	40	game
	0	0.60	0.74	0.87	0.96	1
	15	0.44	0.59	0.76	0.91	1
B score	30	0.25	0.39	0.58	0.81	1
	40	0.09	0.17	0.31	0.58	
	game	0	0	0		

Table 1: The conditional probabilities of A winning the game from various scorelines

A similar spreadsheet can be set up for a set. The constant probability of winning a game can be obtained from the (0,0) cell of the previous sheet. Similarly for a match. Thus we obtain six connected

		A score				
		0	15	30	40	game
	0	6.7	5.6	4.0	2.1	0
	15	5.9	5.2	4.1	2.3	0
B score	30	4.5	4.4	4.0	2.8	0
	40	2.5	2.7	3.1	4.0	
	game	0	0	0		

Table 2: The expected number of points remaining in a game from various scorelines

sheets - entering the chance of winning a point results in the chance of winning a game and expected number of points in the remainder of the game conditional on the game score, the chance of winning a set and expected number of games in the remainder of the set conditional on the set score, and the chance of winning a match and expected number of sets in the remainder of the match conditional on the match score.

In this case, the sheet suggests a player with a 54% chance of winning a point, has a 59.9% chance of winning a game, a 76.3% chance of winning a tiebreak set, a 85.9% chance of winning a 3 set match and a 91% chance of winning a five-set match.

Excel has excellent facilities for building up one or two way tables. Table 3 shows some of the match statistics for chances and expected lengths of winning games, sets and matches with the probability of winning a point ranging from 0.5 to 1. We will use the following notation:

$p(Game)$ = probability of winning a game
$M(Game)$ = expected number of points in a game
$p(Set)$ = probability of winning a tiebreak set
$M(Set)$ = expected number of games in a tiebreak set
$p(Match)$ = probability of winning a 5 set match
$M(Match)$ = expected number of sets in a 5 set match

Magnus and Klassen [2] tested some often-heard hypotheses relating to the service in tennis. They collected data on 481 matches played in the men's singles and women's singles championships at Wimbledon from 1992 to 1995. We will compare some of their statistics with our model. In men's singles with two seeded player's, on service they won 67% of the points and 86% of the games. In women's singles with two seeded player's, on service they won 57% of the points and 67% of the games. Both these results agree with our model as highlighted in Table 3.

3 Two parameter model

For more realistic models the process is essentially the same, just a little more complicated. We now have 2 parameters where:

p_A = probability of A winning a point if A is serving
p_B = probability of B winning a point if B is serving

Since the chance of a player winning now depends on who is serving, we need to introduce this into the state description. The state of a game becomes the score and whether the player is serving or receiving. Thus we need two sheets for a game, one for player A serving and one for player B serving. Similarly we need two sheets for the set score, and one for the match score. The formula and boundary conditions are similar. The main difference is that adjacent cells in terms of the Markov chain are not necessarily adjacent on the spreadsheet, since players alternate service games. We will also use the following notation:

p'_A = probability of A winning a game if A is serving
p'_B = probability of B winning a game if B is serving

p	$p(Game)$	$M(Game)$	$p(Set)$	$M(Set)$	$p(Match)$	$M(Match)$
0.5	0.50	6.8	0.50	9.7	0.50	4.1
0.52	0.55	6.7	0.56	9.6	0.75	4.0
0.54	0.60	6.7	0.62	9.3	0.91	3.7
0.56	0.65	6.7	0.68	9.0	0.98	3.4
0.57	0.67	6.6	0.73	8.8	1	3.3
0.58	0.69	6.6	0.74	8.6	1	3.2
0.6	0.74	6.5	0.79	8.1	1	3.1
0.62	0.78	6.4	0.83	7.7	1	3
0.64	0.81	6.3	0.87	7.4	1	3
0.66	0.85	6.1	0.9	7.1	1	3
0.67	0.86	6.1	0.92	7	1	3
0.68	0.88	6.0	0.93	6.9	1	3
0.7	0.90	5.8	0.95	6.7	1	3
0.8	0.98	5.1	1	6.1	1	3
0.9	1	4.5	1	6	1	3
1	1	4	1	6	1	3

Table 3: Match statistics depending on the probability of winning a point

The following probabilities are applied: $p_A = 0.6$, $p_B = 0.58$

The conditional probabilities for an advantage set are given in Tables 4 and 5 for each player serving. The probability of A winning an advantage set from 6 games all - A or B serving is:

$$\frac{p'_A p'_B}{1 - p'_A p'_B + (1 - p'_A)(1 - p'_B)}$$

By substituting $p_A = 0.6$ in our simple model, $p'_A = 0.74$. Similarly by letting $p_B = 0.58$, $p'_B = 0.69$. Substituting these figures into the equation above gives 0.55.

		A score						
		0	1	2	3	4	5	6
B score	0	0.57	0.75	0.82	0.93	0.97	1	1
	1	0.49	0.57	0.77	0.84	0.96	0.99	1
	2	0.29	0.48	0.56	0.79	0.87	0.98	1
	3	0.20	0.25	0.46	0.58	0.82	0.92	1
	4	0.06	0.15	0.20	0.44	0.55	0.88	1
	5	0.02	0.03	0.09	0.13	0.41	0.55	0.88
	6	0	0	0	0	0	0.41	0.55

Table 4: The conditional probabilities of A winning an advantage set from various games scores if A is serving

As expected 6-6 = 5-5 = 4-4 regardless of who is serving. Also worth noting is that P(winning set | a score with an even number of games with A serving) = P(winning set | a score with an even number of games with B serving).

4 Six parameter model

Instead of building up independent sheets linked through the probability of winning a point, game or set, we can build a model in which the state of the game is match score, set score , game score and whether

		A score						
		0	1	2	3	4	5	6
	0	0.57	0.64	0.82	0.88	0.97	0.99	1
	1	0.37	0.57	0.65	0.84	0.91	0.99	1
	2	0.29	0.35	0.56	0.65	0.87	0.94	1
B score	3	0.12	0.25	0.31	0.56	0.67	0.92	1
	4	0.06	0.08	0.20	0.26	0.55	0.69	1
	5	0.01	0.03	0.04	0.13	0.17	0.55	0.69
	6	0	0	0	0	0	0.17	0.55

Table 5: The conditional probabilities of A winning an advantage set from various game scores if B is serving

the player is serving or receiving. Also included is first and second serves. With this model, a typical question could be: What is the probability of player A winning a five set match with the current score at 2 sets to 1, 3 games to 4, 30-15 and fault one?

We now have 6 parameters: 1st serve % for each player, winning % on first serve for each player and winning % on second serve for each player.

The same techniques of deriving a formula for the last point of a game or set and allowing recurrence relations to work out the remaining points are applied.

We will take the statistics from the 2002 Australian Open men's singles final, where Thomas Johannson defeated Marat Safin 3-6 6-4 6-4 7-6 (7-4)

Let Johannson = player A, Safin = player B

1st serve % for player A = 56%
1st serve % for player B = 63%
Winning % on 1st Serve for player A = 86%
Winning % on 1st Serve for player B = 67%
Winning % on 2nd Serve for player A = 53%
Winning % on 2nd Serve for player B = 54%

Table 6 represents the conditional probabilities of player A winning a 5 set match at a set all, 4-4 in the 3rd set with player A serving. It highlights the importance of a 1st serve on certain points. At 30-40 a 1st serve in would give a 77% chance of winning the match, whereas 30-40 fault drops to 74%. Particularly in tiebreaks, this margin has greater emphasis to the outcome of a match.

		A score								
		0	0/2nd	15	15/2nd	30	30/2nd	40	40/2nd	game
	0	0.84	0.83	0.84	0.84	0.85	0.85	0.85	0.85	1
	15	0.81	0.8	0.83	0.82	0.84	0.84	0.85	0.85	1
B score	30	0.77	0.76	0.8	0.78	0.82	0.81	0.84	0.84	1
	40	0.71	0.7	0.74	0.71	0.77	0.74	0.82	0.81	
	game	0	0	0	0	0	0	0		

Table 6: Conditional probabilities of winning a game from various scorelines

5 Conclusions

Through the use of Excel we have shown how to produce the conditional probabilities of either player winning the match and the expected number of points, games and sets remaining from any position within the match.

The use of "if statements" often used in Excel would allow us to enter what type of match it is (tiebreaker of advantage, 3 or 5 sets) and the current position in the match, to give us the corresponding probabilities and the expected length remaining. This idea would make accessible the relevant statistics for a match in progress. Microsoft Excel now incorporates Visual Basic for Applications (VBA), enabling the modelling of a tennis match to be powerful and still relatively easy to implement.

References

[1] J.S. Croucher, "The conditional probability of winning games of tennis", *Res. Quart. for Exercise and Sport*, **57**(1) (1986), 23–26.

[2] J.R. Magnus and F.J.G.M Klaassen, "On the advantage of serving first in a tennis set: four years at Wimbledon". *The Statist.*, **48**(2) (1999), 247–256.

[3] G.H. Pollard, "An analysis of classical and tie-breaker tennis". *Austral. J. Statist.*, **25** (1983), 496–505.

[4] R.W. Schutz, "A mathematical model for evaluating scoring systems with specific reference to tennis". *Res. Quart. for Exercise and Sport*, **41** (1970), 552–561.