Developing a tennis calculator to teach probability and statistics

Tristan Barnett

1. Introduction

In an article by Reza Noubary "Teaching Mathematics and Statistics Using Tennis" (Noubary, 2010), a tennis match is analyzed with a view towards its use as an aid for teaching mathematical and statistical concepts. It shows that through sports students can be exposed to the basics of mathematical modelling and statistical reasoning using material that interests them. This idea of using tennis to teach mathematics was also given in Barnett et al (2011) where it was demonstrated how students could build their own spreadsheets to generate the probability of winning a tennis match conditional on the state of the match. This would allow students to investigate the properties of tennis scoring systems, obtain knowledge in using spreadsheet software such as Excel and learn elementary probability theory and statistics. This article will extend on this idea by obtaining equations to the distribution of the total number of points played in a game as well as commonly recognized parameters (mean and variance) of this distribution. Similar calculations can be obtained for the total number of points played in a tiebreak game, total number of games played in a tiebreak or advantage set, and the total number of sets played in a best-of-3 or best-of-5 set match. It is then shown how these calculations can be presented as an interactive tennis calculator through sports multimedia with reference to the 'long' men's singles match between John Isner and Nicholas Mahut at the 2010 Wimbledon Championships.

2. Method

2.1 Winning a game

We explain the method by looking at a single game where we have two players, A and B, and player A has a constant probability p_A of winning a point on serve. We set up a Markov chain model of a game where the state of the game is the current point score in the game (thus 40-30 is 3-2). With probability p_A the state changes from a,b to a+1,b and with probability $q_A=1-p_A$ it changes from a,b to a,b+1; where a and b represent the current point score for player A and player B respectively. Thus if $P_A(a,b)$ is the probability that player A wins the game on serve when the score is (a,b), we have the following backward recursion formula:

$P_A(a,b)=p_AP_A(a+1,b)+q_AP_A(a,b+1)$

The boundary values are:

$P_A(a,b)=1$, if $a=4$ and $b\leq 2$

$P_A(a,b)=0$, if $b=4$ and $a\leq 2$

The boundary values and recursion formula can be entered on a simple spreadsheet (such as Excel). 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. In preference, an explicit formula is obtained by recognizing that the chance of winning from deuce is in the form of a geometric series

$P_A(3,3)=p_A^2+p_A^2 2p_Aq_A+p_A^2(2p_Aq_A)^2+p_A^2(2p_Aq_A)^3+\ldots\ldots\ldots$

where the first term is p_A^2 and the common ratio is $2p_Aq_A$

The sum is given by $p_A^2/(1-2p_Aq_A)$ provided that $-1<2p_Aq_A<1$. We know that $0<2p_Aq_A<1$, since $p_A>0$, $q_A>0$ and $1-2p_Aq_A=p_A^2+q_A^2>0$.

Therefore the probability of winning from deuce is $p_A^2/(1-2p_Aq_A)$. Since $p_A+q_A=1$, this can be expressed as:
$P_A(3,3)=p_A^2/(p_A^2+q_A^2)$

Excel spreadsheet code to obtain the conditional probabilities of player A winning a game on serve is as follows:
Enter the text **p_A** in cell D1
Enter the text **q_A** in cell D2
Enter **0.6** in cell E1
Enter **=1-E1** in cell E2
Enter **1** in cells C11, D11 and E11
Enter **0** in cells G7, G8 and G9
Enter **= E1^2/(E1^2+E2^2)** in cell F10
Enter **=$E$1*C8+$E$2*D7** in cell C7
Copy and Paste cell **C7** in cells D7, E7, F7, C8, D8, E8, F8, C9, D9, E9, F9, C10, D10 and E10

Notice the absolute and relative referencing used in the formula =$E$1*C8+$E$2*D7. By setting up an equation in this recursive format, the remaining conditional probabilities can easily and quickly be obtained by copying and pasting.

Table 1 represents the conditional probabilities of player A winning the game from various score lines for p_A = 0.6. It indicates that a player with a 60% chance of winning a point has a 73.6% 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. Also worth noting is that the chances of winning from deuce and 30-30 are the same.

				B score		
		0	15	30	40	game
	0	0.736	0.576	0.369	0.150	0
	15	0.842	0.714	0.515	0.249	0
A score	30	0.927	0.847	0.692	0.415	0
	40	0.980	0.951	0.877	0.692	
	game	1	1	1		

Table 1: The conditional probabilities of player A winning the game from various score lines for p_A = 0.6

2.2 Distribution of the total number of points played in a game

Let $N_A(g,h|a,b)$ be the probability of reaching a point score (g,h) in a game from point score (a,b) for player A serving, where a and b represent the current point score for player A and player B respectively,

and g and h represent the projected point score for player A and player B respectively. The forward recursion formulas are:

If a=g and b=h, then $N_A(g,h|a,b)=1$, otherwise
$N_A(g,h|a,b)=p_AN_A(g-1,h|a,b)$, if g=4 and 0≤h≤2; h=0 and 1≤g≤4; g≥3, h≥3 and g=h+1; g≥3, h≥3 and g=h+2
$N_A(g,h|a,b)=q_AN_A(g, h-1|a,b)$, if h=4 and 0≤g≤2; g=0 and 1≤h≤4; g≥3, h≥3 and h=g+1; g≥3, h≥3 and h=g+2
$N_A(g,h|a,b)=p_AN_A(g-1,h|a,b)+q_AN_A(g,h-1|a,b)$, if 1≤g≤3 and 1≤h≤3; g≥4, h≥4 and g=h
$N_A(0,0|0,0)=0$

Table 2 lists the probability of reaching various score lines in a game from the outset (a=0, b=0) with $p_A=0.6$. It indicates that the probability of reaching deuce in such a game is 0.276.

The probability of player A winning the game on serve from the outset when $p_A=0.6$ can be obtained from:
$N_A(4,0|0,0)+N_A(4,1|0,0)+N_A(4,2|0,0)+N_A(3,3|0,0)P_A(3,3)=0.130+0.207+0.207+0.276*0.692=0.736$

				B score		
		0	15	30	40	game
	0	1	0.400	0.160	0.064	0.026
	15	0.600	0.480	0.288	0.154	0.061
A score	30	0.360	0.432	0.346	0.230	0.092
	40	0.216	0.346	0.346	0.276	
	game	0.130	0.207	0.207		

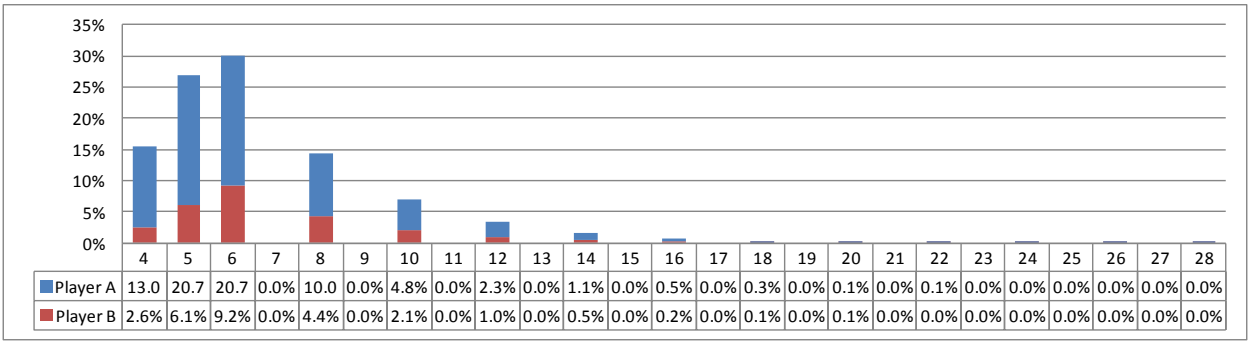
Table 2: The probability of reaching various score lines in a game from the outset with $p_A=0.6$

Let $X_A(a,b)$ be a random variable of the total number of points played in a game at point score (a,b) for player A serving. Let $f(X_A(a,b)=x_A(a,b))$ represent the distribution of the total number of points played in the game at point score (a,b) for player A serving. Then:

$f(X_A(a,b)=4)=N_A(4,0|a,b)+N_A(0,4|a,b)$
$f(X_A(a,b)=5)=N_A(4,1|a,b)+N_A(1,4|a,b)$
$f(X_A(a,b)=6)=N_A(4,2|a,b)+N_A(2,4|a,b)$
$f(X_A(a,b)= x_A(a,b))= N_A((x_A(a,b)+2)/2,(x_A(a,b)-2)/2|a,b)+ N_A((x_A(a,b)-2)/2,(x_A(a,b)+2)/2|a,b)$, for $x_A(a,b)=8,10,12,\ldots$.

Figure 1 represents the distribution graphically of the total number of points played in a game from the outset for $p_A=0.6$. Notice how the blue colour is the chances of player A winning the game and the maroon colour is the chances of player B winning the game. For example, the chances of player A winning the game to 15 is given by the frequency distribution of blue for 5 total points played. This numerical value is 20.74%. Similarly, the chances of player B winning the game to 15 is given by the frequency distribution of maroon for 5 total points played. This numerical value is 6.14%. Therefore, the

game finishing with either player winning to 15 (or 5 total points played) is given by 20.74%+6.14%=26.9%.



	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
Player A	13.0	20.7	20.7	0.0%	10.0	0.0%	4.8%	0.0%	2.3%	0.0%	1.1%	0.0%	0.5%	0.0%	0.3%	0.0%	0.1%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Player B	2.6%	6.1%	9.2%	0.0%	4.4%	0.0%	2.1%	0.0%	1.0%	0.0%	0.5%	0.0%	0.2%	0.0%	0.1%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

Figure 1: The distribution of the total number of points played in a game from the outset for $p_A=0.6$

2.3 Parameters of distribution of the total number of points played in a game

Consider the random variable $X_A(a,b)$ of the total number of points played in a game at point score (a,b) for player A serving. If the game has not reached its completion then the next point must be contested, and one of two results must occur. If player A wins the point then the score progresses to (a+1,b); otherwise the score progresses to (a, b+1). It follows that:
$X_A(a,b)=X_A(a+1,b)$ with probability p_A, and
$X_A(a,b)=X_A(a,b+1)$ with probability q_A.

Taking expectations we obtain a backwards recurrence formula
$E(X_A(a,b))=p_AE(X_A(a+1,b))+q_AE(X_A(a,b+1))$

The mixture law above applies at each and every score before the game is completed.

Consider the random variable $Y_A(a,b)$ of the number of points remaining in a game at point score (a,b) for player A serving. Then
$X_A(a,b)=a+b+Y_A(a,b)$ for all $a\geq0$, $b\geq0$

As the score progresses we now have
$Y_A(a,b)=1+Y_A(a+1,b)$ with probability p_A, and
$Y_A(a,b)=1+Y_A(a,b+1)$ with probability q_A.

Taking expectations we obtain a backwards recurrence formula
$E(Y_A(a,b))=1+p_AE(Y_A(a+1,b))+ q_AE(Y_A(a,b+1))$

The mixture law above applies at each and every score before the game is completed.

Let $\mu(Y_A(a,b))$ represent the mean number of points remaining in a game at point score (a,b) for player A serving. Since the expectation of a random variable is equivalent to the mean, it follows that the backward recursion formula for the mean number of points remaining in a game at point score (a,b) becomes:

$\mu(Y_A(a,b))=1+p_A\,\mu(Y_A(a+1,b))+q_A\,\mu(Y_A(a,b+1))$

The boundary value $\mu(Y_A(3,3))$ is obtained as follows.

Let $M_{Y_A(a,b)}(t)$ represent the moment generating function for the number of points remaining in a game from point score (a,b) with player A serving. Therefore: $M_{Y_A(3,3)}(t)=(p_A^2+q_A^2)e^{2t}/(1-2p_Aq_Ae^{2t})$

The first moment $E(Y_A(3,3))$ is obtained as $M^{(1)}_{Y_A(3,3)}(0)=2(p_A^2+q_A^2)/(1-2p_Aq_A)^2=2/(p_A^2+q_A^2)$. Therefore $\mu(Y_A(3,3))=E(Y_A(3,3))=2/(p_A^2+q_A^2)$

Therefore the boundary values are obtained as
$\mu(Y_A(a,b))=0$, if b=4 and $a \le 2$; a=4 and $b \le 2$
$\mu(Y_A(3,3))=2/(p_A^2+q_A^2)$

Table 3 lists the mean number of points remaining in a game from point score (a,b) with $p_A=0.6$. It indicates that the expected number of points to be played in such a game is 6.5.

				B score		
		0	15	30	40	game
	0	6.5	6.0	4.8	2.8	0
	15	5.2	5.0	4.5	3.0	0
A score	30	3.6	3.7	3.8	3.3	0
	40	1.8	2.0	2.5	3.8	
	game	0	0	0		

Table 3: The mean number of points remaining in a game from various score lines with $p_A=0.6$

Let $\mu(X_A(a,b))$ represent the mean number of total points played in a game at point score (a,b) for player A serving.

It can be shown that:
$\mu(X_A(a,b)) = \mu(Y_A(a,b))+a+b$, for all $a\ge0, b\ge0$

The following analysis is used to obtain $\sigma^2(X_A(a,b))$ and $\sigma^2(Y_A(a,b))$, the variance of the total number of points played and the variance of the number of points remaining respectively in the game at point score (a,b) for player A serving.

Clarke and Norman (1979) used recurrence relations to calculate probabilities of winning, mean and variance of lengths to squash. In particular they showed for a random variable Z which takes the value Z_1 with probability Π and the value Z_2 with probability 1-Π, that

$E(Z)=\Pi E(Z_1)+(1-\Pi)E(Z_2)$

$\sigma^2(Z)=\Pi\sigma^2(Z_1)+(1-\Pi)\sigma^2(Z_2)+\Pi(1-\Pi)(E(Z_1)-E(Z_2))^2$

Since the equation representing E(Z) is in the same format as $E(X_A(a,b))$, then it follows that $\sigma^2(X_A(a,b))$, the variance of the total number of points played in the game at point score (a,b) for player A serving is given by:

$$\sigma^2(X_A(a,b)) =p_A\sigma^2(X_A(a+1,b))+q_A\sigma^2(X_A(a,b+1))+p_Aq_A(\mu(X_A(a+1,b))-\mu(X_A(a,b+1)))^2$$

Since $E(X_A(a,b))=\mu(X_A(a,b))$, it follows that:

$E(X_A(a,b))=a+b+E(Y_A(a,b))$

$E(X^2_A(a,b))=(a+b)^2+2(a+b)E(Y_A(a,b))+E(Y^2_A(a,b))$

$\sigma^2(X_A(a,b))=E(X^2_A(a,b))-E(X_A(a,b))^2=\sigma^2(Y_A(a,b))$, for all $a\geq0,b\geq0$

The boundary value $\sigma^2(Y_A(3,3))$ is obtained as follows.

Using the analysis above to obtain the moment generating function for the number of points remaining in a game from point score (a,b) with player A serving, the second moment $E(Y^2_A(3,3))$ is obtained as $M^{(2)}_{Y_A(3,3)}(0)=4(1+2p_Aq_A)/(p_A^2+q_A^2)^2$.

Therefore $\sigma^2(Y_A(3,3))= E(Y^2_A(3,3))-E(Y_A(3,3))^2=8p_Aq_A/(p_A^2+q_A^2)^2$

Therefore the boundary values are obtained as

$\sigma^2(Y_A(a,b))=0$, if $b=4$ and $a\leq2$; $a=4$ and $b\leq2$

$\sigma^2(Y_A(3,3))=8p_Aq_A/(p_A^2+q_A^2)^2$

Table 4 lists the variance of the number of points remaining in a game from point score (a,b) with $p_A=0.6$. It indicates that the variance of the number of points to be played in such a game is 6.7.

				B score		
		0	15	30	40	game
	0	6.7	7.2	7.7	6.5	0
	15	6.2	6.7	7.4	7.3	0
A score	30	4.9	6.1	7.1	7.8	0
	40	2.6	4.1	6.4	7.1	
	game	0	0	0		

Table 4: The variance of the number of points remaining in a game from various score lines with $p_A=0.6$

3. Sports Multimedia

Strategic Games specializes in delivering online sports content and currently has an interactive tennis calculator freely available in a Java applet (www.strategicgames.com.au). The user interacts by firstly entering the probabilities of each player winning a point on serve followed by the current server and score line; and the calculator outputs the chances of winning the game, set and match. A more extensive version of the calculator could include the chances of reaching a future score line from a particular score line, distributions of the total number of points, games and sets played at different levels within a tennis match, along with the parameters of distribution (such as the mean and variance). Calculations to obtain these results were outlined in sections 2.2 and 2.3 by focusing on points within a single game. Further, a predictive feature could be included that gives serving probabilities for any two given players on the men's and women's main tour on a particular surface.

Figure 2 represents a tennis calculator at the outset of the Isner versus Mahut match played at the 2010 Wimbledon Championships – a match lasting 11 hours and 5 mins with Isner winning 70-68 in the advantage fifth set. Forecasting methods (Barnett et al 2005, 2011) were used to estimate the serving chances of Isner and Mahut winning 69.3% and 70.7% of points on serve respectively. It shows that Mahut has a 90.9% chance of winning a game on serve, 54.3% chance of winning the set and a 58.4% chance of winning the match. It also shows that there is 17.8% chance of reaching deuce on Mahut's serve, 37.9% chance of reaching a tiebreak game and a 36.9% chance of reaching a deciding advantage 5th set. From the graph of the total number of sets played in the match, it shows that Isner has a 9.5%, 15.5% and 16.6% chance of winning the match in 3, 4 and 5 sets respectively, and that Mahut has a 16.0%, 22.0% and 20.4% chance of winning the match in 3, 4 and 5 sets respectively. The parameters of distribution of the total number of sets played in the match are also displayed, and show that the mean number of sets to be played is 4.11 with a corresponding standard deviation of 0.78. An interesting piece of analysis can be used to show that even though this 'long' match was difficult to predict from the outset, there is evidence based on the serving performance throughout the match that a 'long' advantage fifth set could somewhat be predicted at the start of the fifth set. For example, based on the serving statistics in the 3rd set where Isner and Mahut won 81.1% and 75.5% of points on serve; shows that the chances of reaching 68-68 in an advantage fifth set is 3.0%. This is turn can be used in teaching to allow students to investigate properties of scoring systems, where by altering p_A and p_B will give a guide as to the likelihood of the length of the match.

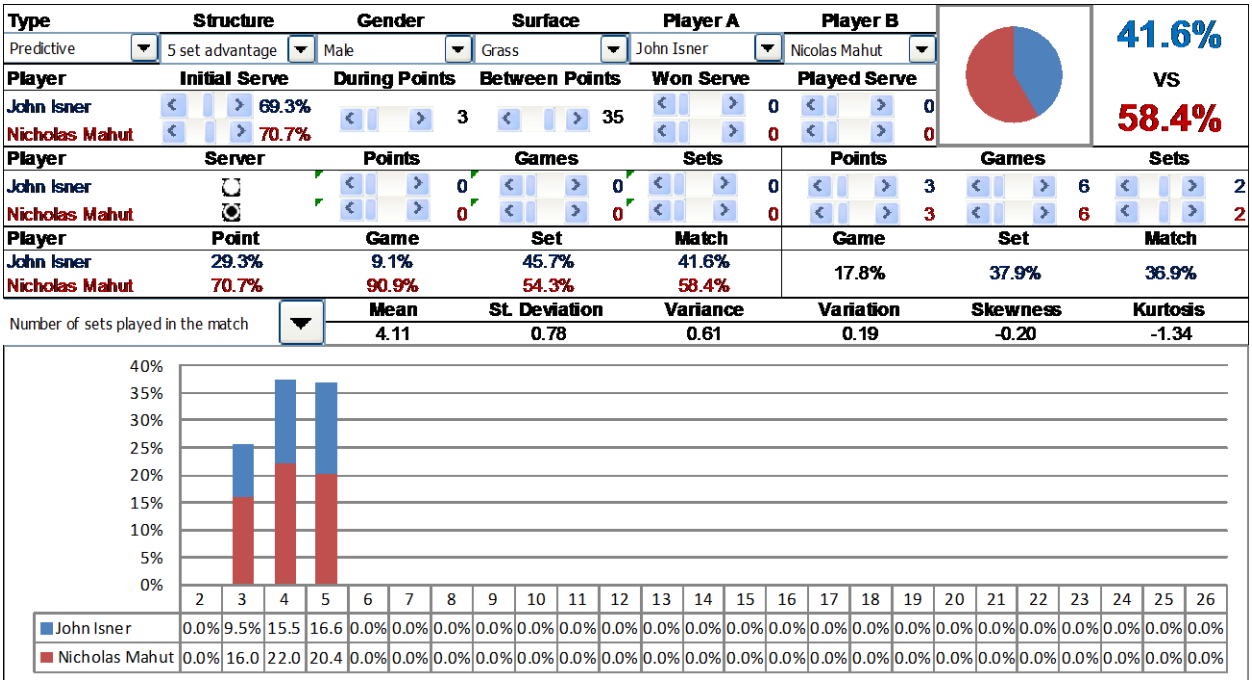


Figure 2: Screenshot of a tennis calculator for the Isner versus Mahut match at the 2010 Wimbledon Championships

References

Barnett T and Clarke SR (2005). Combining player statistics to predict outcomes of tennis matches. IMA Journal of Management Mathematics. 16 (2), 113-120.

Barnett T, O'Shaughnessy D and Bedford A (2011). Predicting a tennis match in progress for sports multimedia. OR Insight 24(3), 190-204.

Clarke S and Norman J (1979). Comparison of North American and international squash scoring systems: analytical results, Research Quarterly 50(4) (1979), 723-728.

Noubary R (2010). Teaching mathematics and statistics using tennis. Mathematics and Sports. Mathematical Association of America.