Teaching probability theory through tennis

Tristan Barnett

1. Introduction

Suppose we wish to calculate the mean (average value) number of points remaining in a game. Using a standard formula for calculating the mean value of a discrete distribution, this can be calculated by $\mu=E(X)$. Similarly the variance (standard deviation squared) of the number of points remaining in a game can be calculated by $\sigma^2 =E(X^2)-E(X)^2$; which is recognized as a measure of the dispersion of a set of data from its mean. Barnett (2006) applies generating functions to calculate the mean and variance of the number of points remaining in a game from the outset and show that when the server has a 60% chance of winning a point on serve, the mean number of points remaining from the outset is 6.5 with a corresponding standard deviation of 2.6. Barnett (2013) applies backward recurrence formulas to obtain the mean and variance of the number of points remaining in a game from any point score within the game. For example when the server has a 60% chance of winning a point on serve; from 30-0 the mean number of points remaining in the game is 5.6 with a corresponding standard deviation of 2.2.

Both the mean and variance contain important information to describe the shape of the distribution and these characteristics could be used to compare one distribution to another. For example, comparing the mean and variance of a tiebreak set to an advantage set to help identify why 'long' matches can occur. However, if a distribution is not symmetric (as typically occurs in a game and an advantage set) the mean and variance do not 'adequately' describe the shape of the distribution. Two other characteristics that are used to describe the distribution and measure risk are skewness and kurtosis. Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the centre point. Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. Note that excess kurtosis will be used throughout the article such that excess kurtosis=kurtosis-3, so the normal distribution has an excess kurtosis of 0, and therefore a kurtosis of 3.

This article uses techniques of recursion formulas and generating functions to obtain the mean, variance, and coefficients of skewness and excess kurtosis of the number of points remaining in a game conditional on the point score. Similar calculations can be obtained for the number of points remaining in a tiebreak game, number of games remaining in a tiebreak or advantage set, and the number of sets remaining in a best-of-3 or best-of-5 set match. The methods can be readily set up in spreadsheets to obtain numerical results (Barnett 2013), and could form an interesting teaching exercise in probability theory by allowing students to compare distributional characteristics of tennis scoring systems.

2. Method

2.1 Moments of the number of points in a game

Let $X_A(a,b)$ and $Y_A(a,b)$ be random variables of the total number of points played in a game and the number of points remaining in a game respectively at point score (a,b) for player A serving.

Let $E(X_A(a,b))$ and $E(Y_A(a,b))$ represent the first moment (or expectation) of the total number of points played in a game and the number of points remaining in a game respectively at point score (a,b) for player A serving.

It can be shown that
$E(X_A(a,b))=p_AE(X_A(a+1,b))+q_AE(X_A(a,b+1))$
$E(Y_A(a,b))=1+p_AE(Y_A(a+1,b))+q_AE(Y_A(a,b+1))$

Let $X^n_A(a,b)$ represent the n^{th} power of the random variable $X_A(a,b)$ for each $n>0$.

Then $E(X^n_A(a,b))$ represents the n^{th} moment with the following important relation
$X^n_A(a,b)=(a+b+Y^n_A(a,b))$

which, when expanded involves various powers of $Y_A(a,b)$. Thus calculation must proceed recursively, i.e. first moment, second moment, and so on. These higher moments can then be used to calculate other statistics such as variance, skewness and kurtosis.

Taking expectations gives the following recurrence formula:
$E(X^n_A(a,b))=p_AE(X^n_A(a+1,b))+q_AE(X^n_A(a,b+1))$

The boundary values for $X^n_A(a,b)$ are obtained as:
$E(X^n_A(4,0))$ and $E(X^n_A(0,4))=4^n$,
$E(X^n_A(4,1))$ and $E(X^n_A(1,4))=5^n$,
$E(X^n_A(4,2))$ and $E(X^n_A(2,4))=6^n$

The boundary values at $E(X^n_A(3,3))$ are obtained as follows:
The moment generating function for the total number of points played in a game from (3,3) with player A serving is given by:
$M_{XA(3,3)}(t)=(p_A^2+q_A^2)e^{8t}/(1-2p_Aq_Ae^{2t})$

Therefore:
$E(X_A(3,3))=M^{(1)}_{XA(3,3)}(0)=4(3p_Aq_A-2)/(2p_Aq_A-1)$
$E(X^2_A(3,3))=M^{(2)}_{XA(3,3)}(0)=8(18p_A^2q_A^2-23p_Aq_A+8)/(2p_Aq_A-1)^2$
$E(X^3_A(3,3))=M^{(3)}_{XA(3,3)}(0)=16(108p_A^3q_A^3-200p_A^2q_A^2+131p_Aq_A-32)/(2p_Aq_A-1)^3$
$E(X^4_A(3,3))= M^{(4)}_{XA(3,3)}(0)=32(648p_A^4q_A^4-1556p_A^3q_A^3+1462p_A^2q_A^2-655p_Aq_A+128)/(2p_Aq_A-1)^4$

Excel spreadsheet code to obtain the first moment of the number of points played in a game for player A serving is as follows:
Enter the text $\mathbf{p_A}$ in cell D1

Enter the text **q_A** in cell D2
Enter **0.6** in cell E1
Enter **=1-E1** in cell E2
Enter **4,5** and **6** in cells C11, D11 and E11 respectively
Enter **4**,**5** and **6** in cells G7, G8 and G9 respectively
Enter = **4*(3*$E$1*$E$2-2)/(2*$E$1*$E$2-1)** in cell F10
Enter **=$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

Tables 1, 2, 3 and 4 represent the first, second, third and fourth moment of the total number of points played in a game at various score line for player A serving given p_A=0.6.

				B score		
		0	15	30	40	game
	0	6.5	7.0	6.8	5.8	4
	15	6.2	7.0	7.5	7.0	5
A score	30	5.6	6.7	7.8	8.3	6
	40	4.8	6.0	7.5	9.8	
	game	4	5	6		

Table 1: The first moment of the total number of points played in a game at various score lines for player A serving given p_A=0.6

				B score		
		0	15	30	40	game
	0	48.8	55.5	54.2	40.1	16
	15	44.2	56.4	63.6	56.1	25
A score	30	36.1	51.6	68.7	76.8	36
	40	25.7	40.3	63.2	104.0	
	game	16	25	36		

Table 2: The second moment of the total number of points played in a game at various score lines for player A serving given p_A=0.6

				B score		
		0	15	30	40	game
	0	441.3	525.7	515.9	346.9	64
	15	385.1	532.2	628.6	535.5	125
A score	30	286.9	468.0	690.6	809.2	216
	40	166.2	319.6	611.5	1204.7	
	game	64	125	216		

Table 3: The third moment of the total number of points played in a game at various score lines for player A serving given p_A=0.6

				B score		
		0	15	30	40	game
	0	4909.8	6009.7	5945.3	3809.3	256
	15	4176.6	6052.6	7369.3	6178.1	625
A score	30	2925.9	5174.9	8163.4	9880.2	1296
	40	1426.6	3182.5	7018.8	15603.0	
	game	256	625	1296		

Table 4: The fourth moment of the total number of points played in a game at various score lines for player A serving given p_A=0.6

2.2 Parameters of distribution of the number of points in a game

Let $\mu(X_A(a,b))$, $\sigma^2(X_A(a,b))$, $\gamma_1(X_A(a,b))$ and $\gamma_2(X_A(a,b))$ represent the mean, variance, coefficient of skewness and coefficient of excess kurtosis of the total number of points played in a game at point score (a,b) for player A serving.

The following standard results are used to obtain $\mu(X_A(a,b))$, $\sigma^2(X_A(a,b))$, $\gamma_1(X_A(a,b))$ and $\gamma_2(X_A(a,b))$

$E(X_A(a,b))=\mu(X_A(a,b))$

$E(X^2_A(a,b))= \sigma^2(X_A(a,b))+ E(X_A(a,b))^2$

$E(X^3_A(a,b))= \gamma_1(X_A(a,b))\sigma^2(X_A(a,b))^{3/2}+3E(X^2_A(a,b))E(X_A(a,b))-2E(X_A(a,b))^3$

$E(X^4_A(a,b))=\gamma_2(X_A(a,b))\sigma^2(X_A(a,b))^2+4E(X^3_A(a,b))E(X_A(a,b))+3E(X^2_A(a,b))^2-12E(X^2_A(a,b))E(X_A(a,b))^2+6E(X_A(a,b))^4$

Let $\mu(Y_A(a,b))$, $\sigma^2(Y_A(a,b))$, $\gamma_1(Y_A(a,b))$ and $\gamma_2(Y_A(a,b))$ represent the mean, variance, coefficient of skewness and coefficient of excess kurtosis of the number of points remaining in a game at point score (a,b) for player A serving.

Finally, the following relations are used to obtain $\mu(Y_A(a,b))$, $\sigma^2(Y_A(a,b))$, $\gamma_1(Y_A(a,b))$ and $\gamma_2(Y_A(a,b))$

$\mu(Y_A(a,b))= \mu(X_A(a,b))-a-b$

$\sigma^2(Y_A(a,b))=\sigma^2(X_A(a,b))$

$\gamma_1(Y_A(a,b))=\gamma_1(X_A(a,b))$

$\gamma_2(Y_A(a,b))=\gamma_2(X_A(a,b))$

Tables 5, 6, 7 and 8 represent the mean, variance, coefficient of skewness and coefficient of excess kurtosis respectively of the number of points remaining in a game at various score lines for player A serving given p_A=0.6.

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

Table 5: The mean number of points remaining at various score lines for player A serving given p_A=0.6

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

Table 6: The variance of the number of points remaining in a game at various score lines for player A serving given p_A=0.6

			B score		
		0	15	30	40
	0	2.2	2.1	1.9	2.4
A score	15	2.4	2.2	2.0	2.1
	30	2.8	2.5	2.1	1.9
	40	4.1	3.4	2.4	2.1

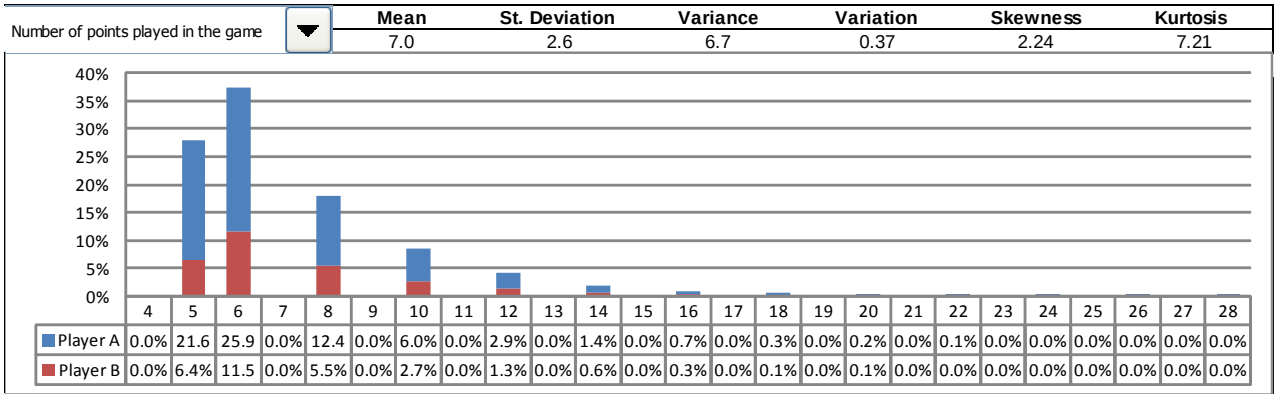
Table 7: The coefficient of skewness of the number of points remaining in a game at various score lines for player A serving given p_A=0.6

			B score		
		0	15	30	40
	0	7.2	6.5	5.6	7.8
A score	15	8.3	7.2	6.1	6.3
	30	11.6	8.6	6.6	5.4
	40	25.4	15.5	8.0	6.6

Table 8: The coefficient of excess kurtosis of the number of points remaining in a game at various score lines for player A serving given p_A=0.6

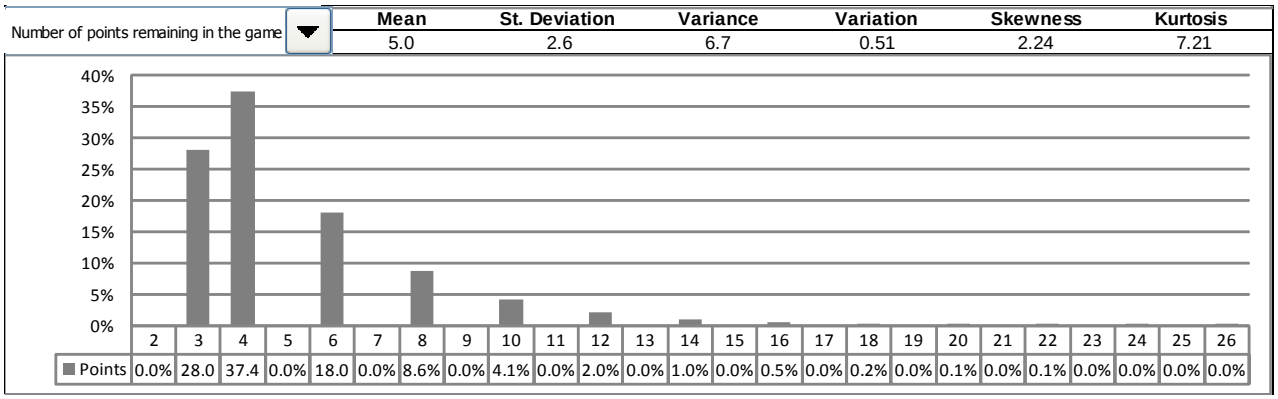
Figure 1 represents the distribution of the total number of points played in a game from 15-15 (a=1, b=1) for player A serving given p_A=0.6 (Barnett 2013). Notice how the blue colour is the chances of player A winning the game and the maroon colour is the chances of player B winning the game. For example, 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%. Figure 2 represents the distribution of the number of points remaining in a game

from 15-15 for player A serving given p_A=0.6. Note that the shapes of both distributions from figures 1 and 2 are the same. In other words the variance and coefficients of skewness and excess kurtosis remain unchanged by adding a constant (c) to all values of the variable. This is widely known as an invariant property in variance such that V(X+c)=V(X). The differences in both distributions are reflected only by shifting the horizontal scale by a constant; as reflected by the mean property M(X+c)=M(X)+c. Note that coefficient of variation = standard deviation/mean.



	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	0.0%	21.6	25.9	0.0%	12.4	0.0%	6.0%	0.0%	2.9%	0.0%	1.4%	0.0%	0.7%	0.0%	0.3%	0.0%	0.2%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Player B	0.0%	6.4%	11.5	0.0%	5.5%	0.0%	2.7%	0.0%	1.3%	0.0%	0.6%	0.0%	0.3%	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: Distribution of the total number of points played in a game from 15-15 for player A serving given p_A=0.6.



	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
Points	0.0%	28.0	37.4	0.0%	18.0	0.0%	8.6%	0.0%	4.1%	0.0%	2.0%	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%

Figure 2: Distribution of the number of points remaining in a game from 15-15 for player A serving given p_A=0.6.

Conclusions

It has been demonstrated in this article how setting up recursion formulas with the appropriate boundary conditions in spreadsheets can generate the first four moments of the total number of points played in a game conditional on the point score. Standard formulas are then used to obtain the distributional characteristics of the mean, variance, and coefficients of skewness and excess kurtosis of the total number of points played in a game and the number of points remaining in a game conditional on the point score. These two distributions are then compared and used to show graphically that the variance and

coefficients of skewness and excess kurtosis remain unchanged by adding a constant to all values of the variable. The methods outlined could form an interesting teaching exercise in probability theory by allowing students to compare distributional characteristics of tennis scoring systems. This in turn allows students to build their own tennis calculator and become familiar with using spreadsheet software such as Excel.

A further application of these four characteristics of distribution is through the Normal Power approximation to estimate frequency distributions (Brown 2012). This allows for computational methods to estimate the probability of a set or match going beyond a specified number of points which could readily be implemented in spreadsheets. Even further, by assigning a constant amount of time to play a point, the probability of a match going beyond a specified amount of time could also be readily implemented in spreadsheets.

References

Barnett T, Brown A and Pollard G (2006). Reducing the likelihood of long tennis matches. Journal of Sports Science & Medicine. 5(4), 567-574.

Barnett T (2013). Developing a tennis calculator to teach probability and statistics. Journal of Medicine and Science in Tennis. 18(1), 30-34.

Brown (2012). Better approximation to the distribution of points played in a tennis match. In proceedings of the 11th Mathematics and Computers in Sport Conference, Melbourne, Australia.