Math HL Portfolio Type II:
Modelling Probabilities in Games of Tennis

May 2009

In this portfolio we will look at the probability involved in playing tennis. Our calculations will be based on the estimated probability a player has of scoring a point. We will develop models for different kinds of tennis games and use Excel to explore up to what extent we can exploit the two probabilities with which we start. Furthermore, we will differentiate between probability and odds, comparing them and analyzing how they can affect the way we look at the same numbers. In my conclusion I will mention the possibility of involving other kinds of distribution in this portfolio, such as Poisson.

Part 1: Club Practice.

1. Games to 10 points.

a) Since we know that Adam wins about twice as many points as Ben does, we can say that the probability of Adam winning a point is $\frac{2}{3}$, and the probability of Ben winning a point is $\frac{1}{3}$. So, given that P(A) is the probability of Adam winning a point and P(B) is the probability of Ben winning a point, we have that:

$P(A)= \frac{2}{3}$

$P(B)= \frac{1}{3}$

Clearly, this is a binomial distribution. Hence, we will use the formula $_nC_x p^x q^{n-x}$, where n is the total number of trials, x the number of successes, p the probability of success and q the probability of failure. Because we want the variable x to represent the number of points won my Adam, we will substitute P(A) for p and P(B) for q. Regardless of what is considered a success or a failure, n will be 10. Following these guidelines our model will be the following:

$$_{10}C_x\left(\frac{2}{3}\right)^x\left(\frac{1}{3}\right)^{10-x}$$

Where x is the number of points won by Adam.

The only concerns I have about the validity of such a model come from the assumptions made in the question. The words "two players have played against each other often enough to know […]" make me wonder how many times is often enough.

Theoretically, if we want to develop a truly pure mathematical formula, both players would have to play against each other an infinite number of times. This would eliminate any impurity brought about by an external influence. However, this is obviously impossible, so "often enough" will be considered to be a sufficient amount of times to include chance and possible alterations into P(A) and P(B). Possible alterations could be the conditions the players play under. Adam could play a lot worst when the temperature increases, whereas Ben might play better on hard courts than on grass courts. Chance simply represents either Ben or Adam having a good day. These uncertainties render a mathematical formula inefficient, as there is no way of including every single external factor into the equation. Therefore, we will work under the assumption that the probabilities include any possible alterations and the frequency with which they might happen.

b) With the model developed in 1.a) and through the use of Excel we can find all the probabilities Adam has of scoring X points with a spreadsheet like the following:

X, number of points Adam wins	$P(x=X)$, probability that Adam wins X points
1	0.000339
2	0.003048
3	0.016258
4	0.056902
5	0.136565
6	0.227608
7	0.260123
8	0.195092
9	0.086708
10	0.017342

In order to calculate each probability Adam has of winning X points we insert the following command into Excel: =(COMBINAT(10,A)*(2/3)^(A)*(1/3)^(10-A), where the column A stands for X, the number of points won by Adam. When we graph the data we get a histogram like the one on the next page:

Probabilities of Adam Winning X Points
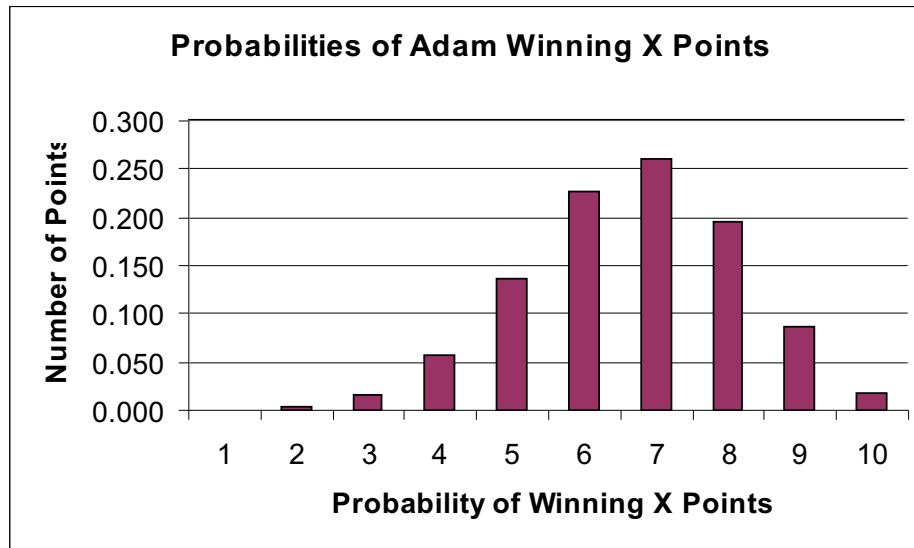
c) To find the expected value and standard deviation we must use the following equations:

$$E(X) = np$$

$$\sigma = \sqrt{npq}$$

We have that $n=10$ and $p = \left(\dfrac{2}{3}\right)$, so

$$E(X) = 10 \times \left(\frac{2}{3}\right) = \left(\frac{20}{3}\right) = 6.667$$

$$\sigma = \sqrt{10 \times \left(\frac{2}{3}\right) \times \left(\frac{1}{3}\right)} = \sqrt{\left(\frac{20}{9}\right)} = 1.491$$

This information tells us that Adam usually wins 7 points, although sometimes he might score a couple of points below or above the mean. It is very unlikely for him to get more than 8 points or less than 4.

Part 2: Non-extended play games.

2. Different ways in which a game might be played.

The first thing we must do if we want to know all the different ways a game can be played is find the domain of *Y*, the number of points played. Because a player must score at least 4 points and maintain a difference of at least 2 points to win, we can already begin to list the possible values of *Y* and all the possible score combinations in which a match can end:

Y= 4,5,6,7 →

4-0

4-1

4-2

4-3

The combinations are so limited because the game cannot go over 7 points, and if a superiority is not established before 6 points deuce is called. Another observation we must make before writing out the different combinations is that, in order for the game to end, the last point must be scored by the winner. If the last point is not scored by the winning player, then the game goes on until either one wins. For example, if the match is at *3-0*, the winner must score the next point or else the game goes to *3-1* and does not end. Again at *3-1*, the winner must score the last point or the game goes on.

Given all of this information we can begin to form our combinations. Because the winner has to score the last point, the possible combinations will look like this: $_{Y-1}C_3$. This means that the winner can score 3 out of the 4 points in any order he wants, but he must always score one last. So out of the remaining *Y-1* points, he can score 3 in any order. This is a little different for the first combination (*4-0*), where there is only one possible way of scoring 4 points in a row. Now we can calculate the different ways one player can win:

$$(_4C_4) + (_{5-1}C_3) + (_{6-1}C_3) + (_{7-1}C_3)$$

$$1+4+10+20$$

$$35$$

Because these are all the ways one player can win we have to double it, which gives us 70.

3. Probability vs. odds that Adam wins.

The model used for this part of the investigation will be the following:

$$p^4 \times q^{Y-4} \times {}_{Y-1}C_3$$

This model only changes when the score ends in *4-0,* in which case it would look like this:

$$p^4 \times q^0 \times {}_YC_4$$

The probability that Adam wins can then be found by adding the probabilities that he wins for every different combination.

$$4\text{-}0 \rightarrow \left(\frac{2}{3}\right)^4 \times \left(\frac{1}{3}\right)^0 \times {}_4C_4 = \left(\frac{16}{81}\right)$$

$$4\text{-}1 \rightarrow \left(\frac{2}{3}\right)^4 \times \left(\frac{1}{3}\right)^1 \times {}_4C_3 = \left(\frac{64}{243}\right)$$

$$4\text{-}2 \rightarrow \left(\frac{2}{3}\right)^4 \times \left(\frac{1}{3}\right)^2 \times {}_5C_3 = \left(\frac{160}{729}\right)$$

$$4\text{-}3 \rightarrow \left(\frac{2}{3}\right)^4 \times \left(\frac{1}{3}\right)^3 \times {}_6C_3 = \left(\frac{320}{2187}\right)$$

$$\left(\frac{1808}{2187}\right) = \text{Probability that Adam wins}$$

Ben will be exactly the same but switching $\left(\frac{2}{3}\right)$ and $\left(\frac{1}{3}\right)$ around:

$$4\text{-}0 \rightarrow \left(\frac{1}{3}\right)^4 \times \left(\frac{2}{3}\right)^0 \times {}_4C_4 = \left(\frac{1}{81}\right)$$

$$4\text{-}1 \rightarrow \left(\frac{1}{3}\right)^4 \times \left(\frac{2}{3}\right)^1 \times {}_4C_3 = \left(\frac{8}{243}\right)$$

$$4\text{-}2 \rightarrow \left(\frac{1}{3}\right)^4 \times \left(\frac{2}{3}\right)^2 \times {}_5C_3 = \left(\frac{40}{729}\right)$$

$$4\text{-}3 \rightarrow \left(\frac{1}{3}\right)^4 \times \left(\frac{2}{3}\right)^3 \times {}_6C_3 = \left(\frac{160}{2187}\right)$$

$$\left(\frac{379}{2187}\right) = \text{Probability that Ben wins}$$

We can perform a quick check on ourselves by adding both probabilities and finding that they give $\left(\frac{2187}{2187}\right)$. Thus, the odds that Adam wins the game are $\left(\frac{1808}{2187}\right) : \left(\frac{379}{2187}\right)$ or *4.77:1*. Note that this is not the only way to solve this problem; there are many different approaches. The answer, though, will always be the same.

4. Generalizing the model to fit random probabilities.

If we are going to generalize the model to fit any random probabilities c and d we should first simplify it with summations. The first case is a little different so it will go separately:

$$p^4 \times q^0 \times {}_4C_4 + \sum_{Y=5}^{7} p^4 \times q^{Y-4} \times {}_{Y-1}C_3$$

Where Y still represents the number of points played.

Because ${}_4C_4$ and q^0 will always be 1, they can be omitted from the final model. That said, we can write out the model for the probability that Player C wins:

$$c^4 + \sum_{Y=5}^{7} c^4 \times d^{Y-4} \times {}_{Y-1}C_3$$

Note: The Combination is represented with a capital C, whilst the variable is a lowercase c.

Part 3: Extended play games.

5. Theoretically endless games.

This point in the investigation gets very tricky, as it becomes very easy to make a carless mistake that can provide us with the wrong numbers. The main difference here with our previous model is that now $Y \geq 4$, whereas before it was $4 \leq Y \leq 7$. In other words, the total number of points played now does not need to be less than 7. However, because 4 is still the minimum amount of points required to win a match, the cases of $4 \leq Y \leq 6$ still apply as the "non-deuce cases." With this in mind we can already separate three probabilities which will be unaffected by whether the game goes to deuce or not:

Y, number of points played:	P(A), Probability that Adam wins:	P(B), Probability that Ben wins:
$Y=4$	$\left(\dfrac{2}{3}\right)^4 \times \left(\dfrac{1}{3}\right)^0 \times {}_4C_4 = \left(\dfrac{16}{81}\right)$	$\left(\dfrac{1}{3}\right)^4 \times \left(\dfrac{2}{3}\right)^0 \times {}_4C_4 = \left(\dfrac{1}{81}\right)$
$Y=5$	$\left(\dfrac{2}{3}\right)^4 \times \left(\dfrac{1}{3}\right)^1 \times {}_4C_3 = \left(\dfrac{64}{243}\right)$	$\left(\dfrac{1}{3}\right)^4 \times \left(\dfrac{2}{3}\right)^1 \times {}_4C_3 = \left(\dfrac{8}{243}\right)$
$Y=6$	$\left(\dfrac{2}{3}\right)^4 \times \left(\dfrac{1}{3}\right)^2 \times {}_5C_3 = \left(\dfrac{160}{729}\right)$	$\left(\dfrac{1}{3}\right)^4 \times \left(\dfrac{2}{3}\right)^2 \times {}_5C_3 = \left(\dfrac{40}{729}\right)$
	Sum: $\left(\dfrac{496}{729}\right)$	Sum: $\left(\dfrac{73}{729}\right)$

From this point on, if the game goes past 6 points it means that a deuce has been called. However, the game cannot be settled at 7 points because a superiority of 2 points is required. This means that if deuce is called at 6 points, and both Adam and Ben win one point again, deuce will be called once more. If A is a point won by Adam and B a point won by Ben, it would look like this:

AAABBB (+ all other possible combinations for the first deuce) AB

AAABBB (+ all other possible combinations for the first deuce) BA

For the game to end at 8 points the same player must score two points consecutively:

AAABBB (+ all other possible combinations for the first deuce), AA

AAABBB (+ all other possible combinations for the first deuce), BB

This basically states that once the first deuce is called the games can only end on an even number of points. Either way, the probability of the game going to the first deuce should be calculated if we want to know what the probability of the consecutive deuces is. The probability of going to the first deuce, $P(d)$, is a simple binomial distribution:

$$P(d) = \ _6C_3 \ \times \left(\frac{2}{3}\right)^3 \times \left(\frac{1}{3}\right)^3$$

So the probability of Adam winning or Ben winning after the first deuce is simply:

$$P(A) = \ _6C_3 \ \times \left(\frac{2}{3}\right)^3 \times \left(\frac{1}{3}\right)^3 \times \left(\frac{2}{3}\right)^2$$

or

$$P(B) = \ _6C_3 \ \times \left(\frac{2}{3}\right)^3 \times \left(\frac{1}{3}\right)^3 \times \left(\frac{1}{3}\right)^2$$

Because every deuce is dependant on its predecessor, we cannot simply find the probability of going to third deuce by increasing the values in the previous model. This means that if we want to find the probability of the third deuce taking place we will have to add the probability of the first, second and third deuces. Therefore, the probability of a deuce taking place is the sum of a geometric series. On top of this lies the fact that the first deuce is different from the rest since it requires 6 points and not just 2. Now we know that the first deuce should be considered separately from the rest of the deuces. Once we have defined this part of the investigation we can find what the probability of any deuce after the first is.

The deuce number that the players are on will thus be represented as $Y - 8$. This is so because the point they are on must be subtracted by 8 to find how many points from the first deuce they are (as this includes the two points after the first deuce). Because every deuce is composed of 2 points, multiplying $\left(\frac{2}{3}\right)^{Y-8} \times \left(\frac{1}{3}\right)^{Y-8}$

9

will double Y and hence form a deuce. Now we can find the probability either player has of winning after going to deuce:

$$\sum_{Y=8}^{\infty} {}_6C_3 \times \left(\frac{2}{3}\right)^3 \times \left(\frac{1}{3}\right)^3 \times 2^{Y-8} \times \left(\frac{2}{3}\right)^{Y-8} \times \left(\frac{1}{3}\right)^{Y-8} \times p^2$$

Where P can be P(A) or P(B).

However, we must add to this the probabilities that either payer wins without reaching deuce, when $Y=4,5,6$. So the final model will look like this:

For P(A)$\rightarrow \left(\frac{496}{729}\right) + \sum_{Y=8}^{\infty} {}_6C_3 \times \left(\frac{2}{3}\right)^3 \times \left(\frac{1}{3}\right)^3 \times 2^{Y-8} \times \left(\frac{2}{3}\right)^{Y-8} \times \left(\frac{1}{3}\right)^{Y-8} \times \left(\frac{2}{3}\right)^2$

For P(B)$\rightarrow \left(\frac{73}{729}\right) + \sum_{Y=8}^{\infty} {}_6C_3 \times \left(\frac{2}{3}\right)^3 \times \left(\frac{1}{3}\right)^3 \times 2^{Y-8} \times \left(\frac{2}{3}\right)^{Y-8} \times \left(\frac{1}{3}\right)^{Y-8} \times \left(\frac{1}{3}\right)^2$

For now we will only look into solving for P(A). This will be done through the use of a simple spreadsheet like the following:

Y	P(A)
8	0.097546106
9	0.043353825
10	0.019268367
11	0.008563718
12	0.003806097
18	2.93349E-05
20	5.79455E-06
22	1.1446E-06
24	2.26094E-07
26	4.46606E-08
28	8.82185E-09
30	1.74259E-09
32	3.44215E-10
34	6.7993E-11
36	1.34307E-11
38	2.65298E-12
40	5.24046E-13
42	1.03515E-13
44	2.04475E-14
46	4.039E-15
48	7.97828E-16
50	1.57596E-16
52	3.113E-17
Sum:	0.172574669
Adam's probability of winning:	0.852958756

The probability of winning is the same as the Sum of the probabilities of winning deuce and non-deuce games. We can find P(B) through the use of a similar spreadsheet but changing the values to fit in for P(B):

Y	P(B)
8	0.024386526
9	0.010838456
10	0.004817092
11	0.00214093
12	0.000951524
18	7.33373E-06
20	1.44864E-06
22	2.86151E-07
24	5.65236E-08
26	1.11652E-08
28	2.20546E-09
30	4.35647E-10
32	8.60537E-11
34	1.69983E-11
36	3.35768E-12
38	6.63246E-13
40	1.31011E-13
42	2.58788E-14
44	5.11186E-15
46	1.00975E-15
48	1.99457E-16
50	3.93989E-17
52	7.7825E-18
Sum:	0.043143667
Ben's probability of winning:	0.143280841

Now that we know P(B) and P(A) we can find the odds Adam has of winning against Ben:

$$P(A):P(B)$$

$$0.8530 : 0.1433$$

$$5.95: 1.00$$

This proves that Adam's game odds are almost 6:1 compared to Ben's

6. Player C plays extended play games.

The formulas for Players C and D are very simple to write as they are the same as the ones that we used in question 5. The formula for the probability that Player C wins is the following:

$$ |c|^4 + \sum_{Y=5}^{6} |c|^4 \times |d|^{Y-4} \times {}_{Y-1}C_3 \; + $$

$$ \sum_{Y=8}^{\infty} {}_6C_3 \times |c|^3 \times |d|^3 \times 2^{Y-8} \times |c|^{Y-8} \times |d|^{Y-8} \times |c|^2 $$

With this equation we can form the spreadsheet:

Y	P(c)				
	0.5	0.55	0.6	0.7	0.9
8	0.078125	0.091723577	0.0995328	0.0907578	0.0118098
9	0.0390625	0.045403171	0.047775744	0.038118276	0.002125764
10	0.01953125	0.02247457	0.022932357	0.016009676	0.000382638
11	0.009765625	0.011124912	0.011007531	0.006724064	6.88748E-05
12	0.004882813	0.005506831	0.005283615	0.002824107	1.23975E-05
18	7.62939E-05	8.10089E-05	6.46217E-05	1.55016E-05	4.21665E-10
20	1.90735E-05	1.98492E-05	1.48888E-05	2.73448E-06	1.36619E-11
22	4.76837E-06	4.86355E-06	3.43039E-06	4.82363E-07	4.42647E-13
24	1.19209E-06	1.19169E-06	7.90362E-07	8.50888E-08	1.43418E-14
26	2.98023E-07	2.91994E-07	1.82099E-07	1.50097E-08	4.64673E-16
28	7.45058E-08	7.15459E-08	4.19557E-08	2.64771E-09	1.50554E-17
30	1.86265E-08	1.75305E-08	9.66659E-09	4.67055E-10	4.87795E-19
32	4.65661E-09	4.29542E-09	2.22718E-09	8.23886E-11	1.58046E-20
34	1.16415E-09	1.05249E-09	5.13143E-10	1.45333E-11	5.12068E-22
36	2.91038E-10	2.57885E-10	1.18228E-10	2.56368E-12	1.6591E-23
38	7.27596E-11	6.31883E-11	2.72398E-11	4.52233E-13	5.37549E-25
40	1.81899E-11	1.54827E-11	6.27604E-12	7.9774E-14	1.74166E-26
42	4.54747E-12	3.79365E-12	1.446E-12	1.40721E-14	5.64297E-28
44	1.13687E-12	9.2954E-13	3.33158E-13	2.48232E-15	1.82832E-29
46	2.84217E-13	2.27761E-13	7.67597E-14	4.37882E-16	5.92376E-31
48	7.10543E-14	5.5807E-14	1.76854E-14	7.72424E-17	1.9193E-32
50	1.77636E-14	1.36741E-14	4.07472E-15	1.36256E-17	6.21853E-34
Sum:	0.151468913	0.176340361	0.186616016	0.154452744	0.014399474
P(C):	0.495218913	0.617858017	0.730936016	0.898762744	0.998549474
P(D):	0.504781087	0.382141983	0.269063984	0.101237256	0.001450526
Odds:	0.981056789	1.616828419	2.71658809	8.877786534	688.4051632

7. Expressions that represent the odds.

The expression that represents the odds is simply $\left(\dfrac{P(C)}{P(D)}\right)$. Analyzing the spreadsheet we came up with for question 6 helps us understand what happens when point-winning probabilities are close together or very different. When point-winning probabilities are close together, the odds one player has against the other tend to be closer to 1:1. For example, when P(c) is 0.5, P(d) will also be 0.5, so P(C) and P(D) will inevitably be very similar. The odds I calculated here were simply 0.98:1.00, which means that in this scenario players C and D have a similar level of skill. Other cases, like when P(c)= 0.55 also help to prove this point. However, only a 5% difference can prove to be pretty significant. By looking at how the odds increase with respect to the increase in the point-winning probability we can conclude that there is an exponential relationship. When P(c) increases from 0.6 to 0.7, the odds only increase by about 6:1. In contrast, when P(c) increases from 0.7 to 0.9, the odds increase by over 600:1. The increment in odds clearly does not show a linear relationship with the point-winning probabilities.

If we use our Graphic Display Calculators we can easily obtain the correlation coefficient (r) of the exponential regression for the different odds. The value for the correlation coefficient for an exponential regression is 0.99. This indicates that the relationship is strongly exponential as it has a very strong correlation with all the points. The following is a graph of odds versus point-winning probabilities:
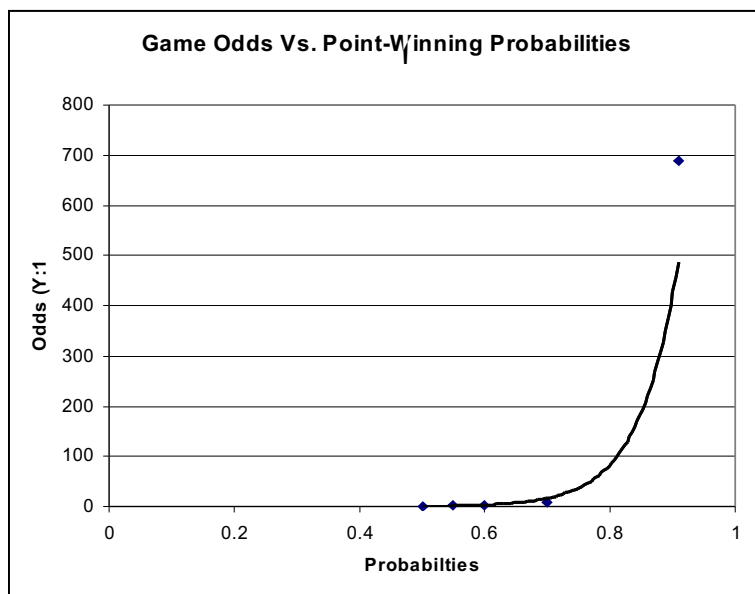
8. Usefulness, limitations and conclusion.

The most obvious limitation to our model is that after all our calculations and finding that Adam's odds are almost 6:1 against Ben's, Adam may still loose. I talked about this briefly in my answer to question 1, where the validity of our first model was debated. I find the whole investigation to revolve around one undeniable idea: It is all hypothetical. I reiterate that the accuracy of our models lies only in the meaning of "often enough" ("Two players have played against each other often enough to know that […]"). In the real world, there would be no such thing as "often enough," since there would always be a margin for error. All this tells us is that when we use our models we must be cautious and always be aware of the fact that they are only predictions. Unfortunately, there is always a small chance that these predictions may be completely wrong. I do not believe that this means the models are useless though, as they can still prove useful for large spans of time. In other words, our models will work better when we deal with averages than when we deal with specific events.

An interesting field of further investigation for this portfolio would be to compare how different distributions can apply to the tennis games. For example, Poisson distribution. In question 1 we found that the expected value for X was about $\left(\dfrac{20}{3}\right)$. If we use this as λ we can try to find the probability that Adam scores X number of points in a match. We can use Excel to construct another simple spreadsheet, this time showing the Poisson distribution:

X, number of points won by Adam	P(X=x)
1	0.008483985
2	0.028280091
3	0.062844961
4	0.104742126
5	0.139656867
6	0.155175072
7	0.147786522
8	0.123156051
9	0.09122716
10	0.060818411

The following spreadsheet is the same, only that it shows the Binomial distribution:

14

X, number of points Adam wins	P(x=X), probability that Adam wins X points
1	0.000339
2	0.003048
3	0.016258
4	0.056902
5	0.136565
6	0.227608
7	0.260123
8	0.195092
9	0.086708
10	0.017342

The following is a comparison between both probabilities:

P(X=x) for Poisson	P(x=X) for Binomial
0.008483985	0.000339
0.028280091	0.003048
0.062844961	0.016258
0.104742126	0.056902
0.139656867	0.136565
0.155175072	0.227608
0.147786522	0.260123
0.123156051	0.195092
0.09122716	0.086708
0.060818411	0.017342

It is interesting to observe that when using different distributions the probability values can vary. Regardless, the correct method is binomial, as a Poisson distribution could technically still work for x=11. What I mean by this is that since the binomial distribution is restricted by the presence of *n*, it will most likely be more accurate as its values will add up to one. The Poisson distribution, on the other hand, will not always add up to exactly one for the value of *n* we are using.