Performance Streaks In Sports

Christopher Bradway
Mentor: Dr. Thomas Beatty

April 2019

Abstract

The purpose of our paper is to use the negative binomial experiment, and its ability to calculate probability in order to explore the phenomenon of streaks in sports. A binomial experiment is a series of Bernoulli trials. The Bernoulli trial is a method used to test the probability of an occurrence that has a "success or failure" outcome. We will apply our method to hitting streaks in Major League Baseball, where we will explore the level of consistency needed to accomplish one of the longest held records in sports; Joe DiMaggio's 56 consecutive game hitting streak which has stood for 78 years. DiMaggio's feat remains the pinnacle for exceptional consistency at the plate, however, using the negative binomial experiment we have discovered other players who have overcome much greater odds in achieving their own hitting streaks.

1 Introduction

In order to achieve a notable performance streak in any sport, an athlete must play at an unexpected level of consistency. We can calculate the probability of such consistency by applying a binomial experiment to the given data. A binomial experiment is a series of Bernoulli trials. A Bernoulli trial is a method used to test the probability of an occurrence that has a "success or failure" outcome. That means there is only one of two possible results when running this test.Sports are a great platform to run such an experiment, after all every attempted feat in sports is either a success or a failure.

We choose our streak based on the fact that, if it is too likely to occur, then our results are uninteresting. This will also apply to the other extreme, if it is too unlikely, or even impossible to achieve then it too is uninteresting. For example the number of consecutive free throws made by one basketball player. The average NBA basketball has a 75 percent success rate. ('For Free Throws, 50 years of Practice is No Help" by John Branch, March 3, 2009, New York Times) The other extreme would be in professional soccer. The best strikers in the world only score at a rate of less than one goal per game with an average of fifteen shots per game. That gives us a scoring percentage of less than six percent, which is so infrequent it too becomes uninteresting.

What I would like to look at is consecutive game hitting streaks in American Baseball. When it comes to the batting averages of most Major League Baseball players, you will see percents between 25 and 40. This shows that getting a hit at an at bat isn't guaranteed, but also not impossible. Another aspect that makes this an interesting choice is the record for consecutive games with a hit, held by Joe DiMaggio. DiMaggio's record is 56 consecutive games with a hit, which he set back in 1941. This is one of the longest records that stills holds up in any sport. DiMaggio's 78 year old record has not just stood the tests of time, but has not really been threatened by anyone since. We will model DiMaggio's record setting performance streak and calculate it's probabilistic nature. If we compare the unlikelihood of such an event, we will discover that much less heralded players have overcome much greater odds in their pursuit of Joe DiMaggio's crown.

So how do we get the most efficient calculation of the probability of a streak in sports? For our chosen streak, I feel it is important to define some of the terms and statistics we will be using. First you will hear me refer to "at bats" per game. At bats refers to the number of times a player steps up to the plate with an unbiased chance of getting a hit. That means the opportunity at the plate is not counted if the batter is hit by a pitch, walked(unintentionally or intentionally), sacrifice fly, or a fielder's choice or error. This prevents a streak from being broken by lack of opportunity to get a hit. This value varies in comparison to plate appearances. Plate appearances are used to calculate a players hitting average, and is a valued based strictly on the number of times a player steps up to the plate no matter what the outcome. A quick example, a player steps up to the plate four times in a game and is walked once. The player's at bat would be three, but the player's plate appearance would be four. Again we will be using the at bats, which the MLB uses to calculate a player's batting average. The batting average will be a key part in our calculations.

2 Theory

The model we would like to use is the negative binomial model, also known as the "wait time" model. This will be a sequence of independent Bernoulli trials where p = the constant probability of success for particular trial. We will partition the sequence into blocks of length λ. We can create our probability mass function $f(n; p\lambda)$ where n is the number of blocks of success. For simplicity we will allow q=1-p. So our probability of a block failure becomes $(1-p)^\lambda = q^\lambda$. Then we can say that the probability of block success is therefore $1 - q^\lambda$. Since each trial is independent, then the blocks are independent as well. The probability of n consecutive block successes followed by a block failure at (n+1) block follows a geometric distribution. So to calculate the probability of an n length streak is $q^\lambda(1-q^\lambda)^n$. If we then convert back to terms of p we get the probability mass function of $f(n; p, \lambda) = (1-p)^\lambda(1-(1-p)^\lambda)^n$. This gives us a cumulative distribution function of $F(n; p, \lambda) = \sum_{k=0}^{n} f(k; p, \lambda)$. Again for simplicity we can rewrite this in terms of q giving us

$F(n;p,\lambda) = \sum_{k=0}^{n} q^{\lambda}(1-q^{\lambda})^{k} = q^{\lambda}\sum_{k=0}^{n}(1-q^{\lambda})^{k}$. We can rewrite $\sum_{k=0}^{n}(1-q^{\lambda})^{k} = \dfrac{1-(1-q^{\lambda})^{n+1}}{1-(1-q^{\lambda})}$, which gives us $q^{\lambda}\dfrac{1-(1-q^{\lambda})^{n+1}}{1-(1-q^{\lambda})} = 1-(1-q^{\lambda})^{n+1}$. Again switching back to p we get $F(n;p,\lambda) = 1-(1-(1-p)^{\lambda})^{n+1}$.

Therefore the probability of a streak of exactly length of n is $f(n;p,\lambda)$, and the probability of a streak up to n and including n is $F(n;p,\lambda)$.

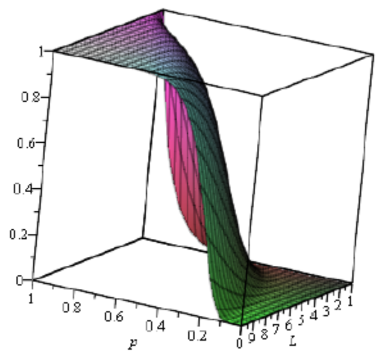
In summary:
P = constant probability of success (batting average)
λ = block length (avg. number of at bats per game)
n = number of block successes (games with at least one safe hit)
$(1-p)^{\lambda}$ = probability of failure
$1-(1-p)^{\lambda}$ = probability of success
$(1-(1-p)^{\lambda})^{n}$ = n consecutive block successes
Failure occurs at the $(n+1)^{\text{st}}$ block, follows a geometric distribution.
Our probability mass function (pmf) is $f(n;p,\lambda)$ where $f(n;p,\lambda) = (1-p)^{\lambda}(1-(1-p)^{\lambda})^{n}$
Our cumulative distribution function (cdf) is $F(n;p,\lambda)$, where $F(n;p,\lambda) = \Sigma_{k=0}^{n} f(k;p,\lambda)$ or after calculating the sum we get $F(n;p,\lambda) = 1-(1-(1-p)^{\lambda})^{n}$.
Now let $S(n;p,\lambda)$ be a streak of at least length n then we get $1-F(n-1;p,\lambda) = (1-(1-p)^{\lambda})^{n}$.

3 Exploring the Effects

As stated in the previous section, the probability of a streak of at least length n is $1-F(n-1;p,\lambda) = (1-(1-p)^{\lambda})^{n}$. Let us label this cumulative distribution function as $S(n;p,\lambda)$.

To show how p and λ effect S I give you the graph in Figure 1. In Figure 1 we plot $S(n;p,\ 4)$ for $0 \leq p \leq 1$ and $1 \leq\ n \leq 10$. Keep in mind that n is an integer but in order to create a good visual we will let be continuous values on [1,10]. We have fixed $\lambda = 4$ in order to create the following 3-D plot.

Figure 1



If we take the 2-D plot of $S(10; p, 4)$ we find that S becomes very sensitive to changes in p when p is in it's mid-range.

Fig. 2

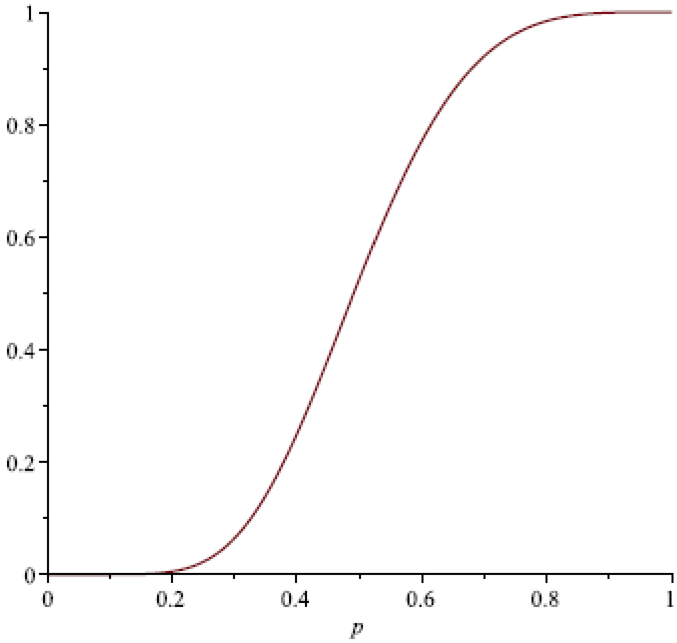
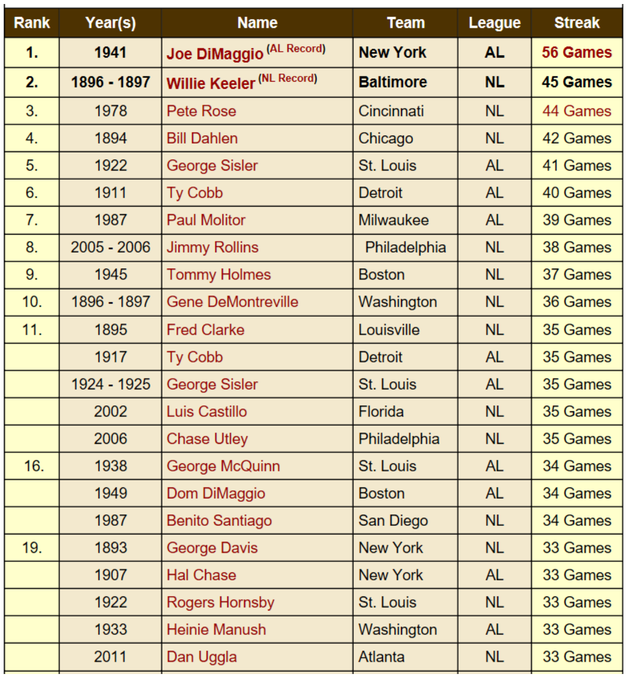


Figure 2 gives us a 2-D plot of S(10;p,4), it represents a section of Figure 1 cut parallel to the p-axis at $\lambda = 4$. This allows us to see the effect on S when p and λ are both set to a constant value.

4 Consecutive Game Hitting Streaks

Let us now look at our chosen subject matter, hitting streaks. First we must amass data to be analyzed. By going to "The Baseball Almanac" we are able to retrieve a list of the more notable hitting streaks in Major League history, setting our cut-off at 33 games.

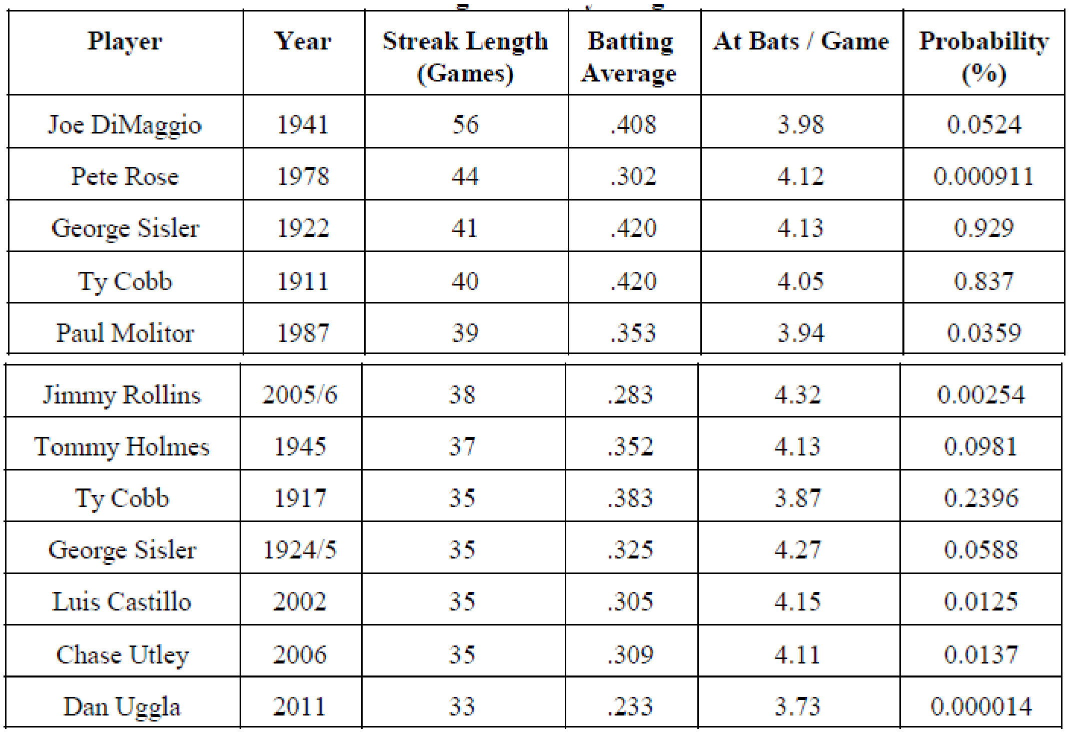


Rank	Year(s)	Name	Team	League	Streak
1.	**1941**	**Joe DiMaggio (AL Record)**	**New York**	**AL**	**56 Games**
2.	**1896 - 1897**	**Willie Keeler (NL Record)**	**Baltimore**	**NL**	**45 Games**
3.	1978	Pete Rose	Cincinnati	NL	44 Games
4.	1894	Bill Dahlen	Chicago	NL	42 Games
5.	1922	George Sisler	St. Louis	AL	41 Games
6.	1911	Ty Cobb	Detroit	AL	40 Games
7.	1987	Paul Molitor	Milwaukee	AL	39 Games
8.	2005 - 2006	Jimmy Rollins	Philadelphia	NL	38 Games
9.	1945	Tommy Holmes	Boston	NL	37 Games
10.	1896 - 1897	Gene DeMontreville	Washington	NL	36 Games
11.	1895	Fred Clarke	Louisville	NL	35 Games
	1917	Ty Cobb	Detroit	AL	35 Games
	1924 - 1925	George Sisler	St. Louis	AL	35 Games
	2002	Luis Castillo	Florida	NL	35 Games
	2006	Chase Utley	Philadelphia	NL	35 Games
16.	1938	George McQuinn	St. Louis	AL	34 Games
	1949	Dom DiMaggio	Boston	AL	34 Games
	1987	Benito Santiago	San Diego	NL	34 Games
19.	1893	George Davis	New York	NL	33 Games
	1907	Hal Chase	New York	AL	33 Games
	1922	Rogers Hornsby	St. Louis	NL	33 Games
	1933	Heinie Manush	Washington	AL	33 Games
	2011	Dan Uggla	Atlanta	NL	33 Games

For those of you who are familiar with American baseball, you will notice that there are some very notable names on this list. Topping this list is our current record holder, who again has held that record for 78 years, Joe DiMaggio. Couple of other things we need to take note of, and it is not stated in the above table. Our players must be regular playing players. This means that the must be consistently in the lineup barring any injuries. We are also considering batters who on average see 3.5 to 5.5 at bats per game. For a regular player this is not a difficult accomplishment.

5 Results

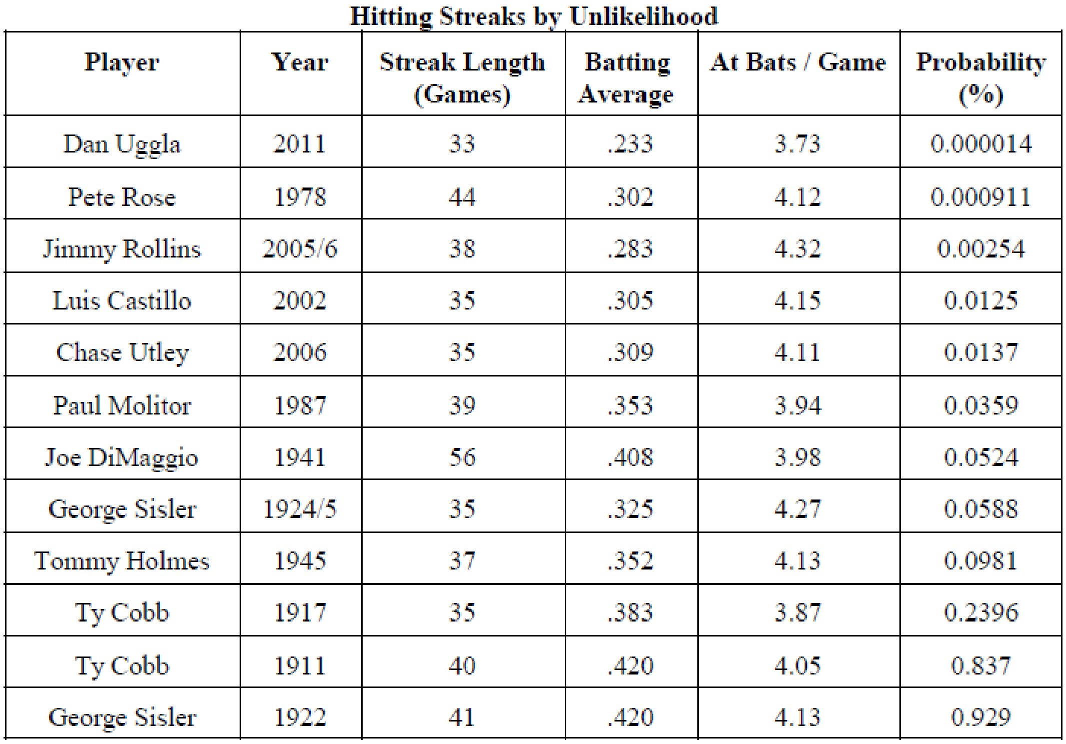
Using the players' batting averages as their likelihood of success per at bat, we can then use our $F(n; p, \lambda)$ and Maple to calculate the probability of each players accomplished consecutive game hitting streak. The following table gives us the results of those calculations.



Player	Year	Streak Length (Games)	Batting Average	At Bats / Game	Probability (%)
Joe DiMaggio	1941	56	.408	3.98	0.0524
Pete Rose	1978	44	.302	4.12	0.000911
George Sisler	1922	41	.420	4.13	0.929
Ty Cobb	1911	40	.420	4.05	0.837
Paul Molitor	1987	39	.353	3.94	0.0359
Jimmy Rollins	2005/6	38	.283	4.32	0.00254
Tommy Holmes	1945	37	.352	4.13	0.0981
Ty Cobb	1917	35	.383	3.87	0.2396
George Sisler	1924/5	35	.325	4.27	0.0588
Luis Castillo	2002	35	.305	4.15	0.0125
Chase Utley	2006	35	.309	4.11	0.0137
Dan Uggla	2011	33	.233	3.73	0.000014

What we notice is that the odds of accomplishing any of these hitting streaks is very close to impossible. Every one of them is less than one percent, and the most probable is George Sisler with a .929 percent chance of reaching his given streak. This is largely due to his impressive batting average of .420.

If we rearrange our table based on least probable to most probable we get the following results.



Hitting Streaks by Unlikelihood

Player	Year	Streak Length (Games)	Batting Average	At Bats / Game	Probability (%)
Dan Uggla	2011	33	.233	3.73	0.000014
Pete Rose	1978	44	.302	4.12	0.000911
Jimmy Rollins	2005/6	38	.283	4.32	0.00254
Luis Castillo	2002	35	.305	4.15	0.0125
Chase Utley	2006	35	.309	4.11	0.0137
Paul Molitor	1987	39	.353	3.94	0.0359
Joe DiMaggio	1941	56	.408	3.98	0.0524
George Sisler	1924/5	35	.325	4.27	0.0588
Tommy Holmes	1945	37	.352	4.13	0.0981
Ty Cobb	1917	35	.383	3.87	0.2396
Ty Cobb	1911	40	.420	4.05	0.837
George Sisler	1922	41	.420	4.13	0.929

What this shows us is that looking at Joe DiMaggio's 56 consecutive game hitting streak, while impressive, falls in that mid range in comparison to the other hitting streaks that were accomplished. It was still incredibly unlikely, but by no means the least likely streak achieved. Again this is due to the batting average of the players, which plays a very large role in the calculations.

6 Conclusion

The purpose of our paper was to explore the level of consistency needed to achieve a performance streak in sports. We were looking into calculating the probability of given streaks. We needed to focus in on a particular streak in order to test our methods for calculation. We chose consecutive game hitting streaks in American Baseball. We felt this was an interesting test subject due to its "interesting" level of consistency. Meaning that the typical players ability to get a hit is not too likely, but not impossible. We also chose this as our test subject because it is one of the longest standing records in all sports. Joe DiMaggio has held the record with a 56 consecutive game hitting streak for over 78 years. Looking at our results, we see that the probability of DiMaggio achieving such a hitting streak is incredibly unlikely. It is definitely an amazing feat. However diving deeper into our calculations and studying our results, we see that Dan Uggla's 33 consecutive game hitting streak was much more unlikely to occur. I think it is safe to say that based on the probability of Uggla achieving his streak, we could say practically impossible. As a matter of fact based on Dan Uggla's .173 batting average at the start of his hitting streak, we calculated that it would be more probable that he would go the entire season without a hit than to achieve a 33 consecutive game hitting streak.

7 Bibliography

7.1 Text

- Branch, John. (2009), *For Free Throws, 50 Years of Practice is No Help*, New York Times.

7.2 Online

- http://www.baseball-almanac.com
- Wearden, Andrew. 2016. *Probability Why Hitting Streaks Are Impressive and Why They're Not* https://www.saberballblog.com/2016/06/19/probability-why-hitting-streaks-are-impressive-and-why-theyre-not/
- Beltrami, Edward and Mendelsohn, Jay. *More Thoughts, DiMaggio's 56 Game Hitting Streak* https://sabr.org/research/more-thoughts-dimaggio-s-56-game-hitting-streak