Streaking: Finding the Probability for a Batting Streak

Stanley Rothman and Quoc Le

Introduction

In baseball, a player can gain instant fame by duplicating or exceeding one of the fabled types of batting streaks. The most well known is Joe DiMaggio's (1941) streak of hitting in 56 consecutive games. There are also other batting streaks such as Ted Williams' (1949) 84-game consecutive on-base streak, Joe Sewell's (1929) 115-game streak of not striking out in a game, and the 8-game streak of hitting at least one home run in each game, held by three players. Other streaks include most consecutive plate appearances with a hit (the record is 12 held by Walt Dropo (1952)), most consecutive plate appearances getting on-base (the record is 16 held by Ted Williams (1957)), and the most consecutive plate appearances with a walk (the record is seven held by five players).

In this paper, we present two functions to calculate the probability of a player duplicating a hitting streak. One is recursive; the other is a new piecewise function that calculates the probability directly.

We use them to compare the difficulty of duplicating different streaks, in particular, the 56-game consecutive hitting streak and the 84-game consecutive on-base streak.

Our results can be applied to streaks outside of baseball.

A recursive function to calculate the probability of a player having a 56-game hitting streak at some point in a season

Michael Freiman[1] gives a recursive function, *R(n)*, to calculate the probability of a player having a 56-game hitting streak at some point in the season.

Let
H = the number of hits in the season
PA = the number of plate appearances in the season.
(A plate appearance will include an at-bat (AB), a walk (BB), a hit-by-pitch (HBP), a sacrifice fly (SF), and a sacrifice bunt (SH). Normally, a sacrifice bunt is not considered a plate appearance.)
G = the number of games a player had at least one plate appearance in the season,
p = H / PA,
j.kl = PA / G (for j.kl = 5.18, j is 5 and .kl is .18),
$$U = [1 - (1 - p)^{j}]*(1 \bullet .kl) + [1 - (1 - p)^{(j+1)}]*.kl,$$
$$R(0) = R(1) = R(2) = \ldots = R(55) = 0,$$
$$R(56) = U^{56},$$
$$R(n) = R(n - 1) + (1 - R(n - 57))*(1 - U)*U^{56} \text{ ; when } n > 56.$$

The average number of plate appearances per game is **j.kl,** the probability of getting at least one hit in a game for a given season is **U**, and the probability of getting a hit in any plate appearance is **p**. $\mathbf{U^{56}}$ is the probability of having a 56-game hitting streak in any particular 56-game span. The value **R(n)** is the probability of having at least one 56-game hitting streak in the first n games.

The function R(n), assumes the following:

(1) The probability **p** is assigned to each plate appearance in the season.

(2) The outcomes of plate appearances are independent events.

(3) The player has the same number of plate appearances in each game for the season, **j.kl**.

(4) A player's batting performance in each game is independent of his performance in any other game.

Some baseball researchers have questioned the independence assumptions and have attempted to show the number of streaks predicted from them underestimate the actual number of streaks. For this paper, these assumptions will be used.

R(n) is the probability of a player having at least one 56-game hitting streak at some point in the first n-games in a season.

At this point, we explain the logic used to define R(n).

For a player to have a 56-game hitting streak, at some point, in the first n-games, he must either have a 56-game hitting streak in the first (n – 1) games or have his first hitting streak in the last 56 games. The probability of having a 56-game hitting streak at some point in the first (n – 1) games is R(n – 1). To have his first hitting streak in the last 56 games, three events must all occur. First, he must not have a 56-game hitting streak in the first (n – 57) games, then he must not have a hit in game number (n – 56), finally, he must get a hit in each of the games (n – 55) through n. Call the three events, E_1, E_2, and E_3:

E_1 = Not having a 56-game hitting streak in the first (n – 57) games

E_2 = Not having a hit in game number (n – 56)

E_3 = Getting a hit in each of the games from game number (n – 55) through game number n.

Since the three events are independent, we have

$$\mathbf{Pr(E_1 \text{ and } E_2 \text{ and } E_3) = Pr(E_1)*Pr(E_2)*Pr(E_3) =(1 - R(n - 57))*(1- U)*U^{56}}$$

The event of having a 56-game hitting streak at some point in the first (n – 1) games and the compound event of (E_1 and E_2 and E_3) are mutually exclusive (they both cannot occur at the same time). Therefore,

$$\mathbf{R(n) = R(n - 1) + (1 - R(n - 57))*(1- U)*U^{56}}$$

Next, a non-recursive function, that we developed to evaluate R(n) is given.

A non-recursive piecewise function, NR(n), to calculate the probability of a player having a 56-game hitting Streak at some point in a season

NR(n) is the probability of at least one hitting streak of length 56 in the first n games. We will call **NR(n)** the **NR Function** to distinguish it from the **Recursive Function R(n)**.

The **NR Function** for calculating the probability of a player achieving a 56-game hitting streak at some point in the first n-games of a season is defined as follows.

For 0 • n • 55, NR(n) = 0.
For 56 • n • 112, NR(n) = U^{56} + (n – 56)*(1 – U)*U^{56}.
For 113 • n • 162,
NR(n) = U^{56} + (n – 56)*(1 – U)*U^{56} • .5*(n – 112)*(1 - U)*U^{112}*[(n –112)*(1 – U)+1+ U].
The meaning of U is the same as in the recursive function.

To see the advantage of the NR Function over the recursive function, consider the following example.

Example 1: Suppose that during a season player A played in 66 games with AB = 322, H = 138, BB = 20, HBP = 0, SF = 0, and SH = 0. What is the probability that at some point in his season he had a 56 game hitting streak?

Solution 1: Using the recursive function

Step 1: p = H / PA = 138 / (322+20+0+0+0) =.404 (p is the probability of a hit in a plate appearance).

Step 2: The average number of plate appearances per game expressed as a decimal is j.kl = PA / G = 342/66 = 5.18. (j = 5, .kl = .18)

Step 3: The weighted average, U, for the probability of a player getting at least one hit in j.kl plate appearances per game is

$$U = [1• \ (1 – p)^{j}]*(1• \ .kl) + [1 – (1 – p)^{(j+1)}]*.kl \ =$$
$$[1• \ (1 – .404)^{5}]*(1• \ .18) + [1 – (1 – .404)^{(5+1)}]*.18 =$$
$$[1- (.596)^{5}]*(.82) + [1 – (.596)^{6}]*(.18) =$$

$$.9248*.82 + .9552*.18 = .9302$$

Step 4: The probability of a player having a 56-game hitting streak in any particular 56-game span in his season of 66 games is :

$$U^{56} = .9302^{56} = .0174$$

Step 5: The use of a recursive function begins with a known beginning term.

The known beginning term is $R(56) = U^{56}$. Our goal is to reach R(66). To reach it we must find R(57), R(58), …, R(66).

R(0) = R(1) = R(2) =…= R(55) = 0,

$R(56) = U^{56} = .9302^{56} = .0174$,

$R(57) = R(56) + (1 – R(n - 57))*(1- U)*U^{56} = R(56) + (1 – R(0))*(1 - .9302)*(.9302)^{56} = .0174 + .0012 = .0186$,

$R(58) = R(57) + (1 – R(n – 57))*(1- U)*U^{56} = R(57) + (1 – R(1))*(1- U)*U^{56} = .0186 + .0012 = .0198$,

$R(59) = R(58) + (1 – R(n - 57))*(1- U)*U^{56} = R(58) + (1 – R(2))*(1 – U)*U^{56}$
$.0198 + .0012 = .0210$,

$R(60) = R(59) + (1- R(n – 57))*(1- U)*U^{56} = .0210 + .0012 = .0222$.

Continuing, we obtain

R(61) = .0234, R(62) = .0246, R(63) = .0258, R(64) = .0270, R(65) = .0282, and R(66) = .0292.

Solution 2: Using the NR Function

Since 56 • n • 112, $NR(n) = U^{56} + (n – 56)*(1 – U)*U^{56}$.

$NR(66) = (.9302)^{56}+(66-56)*(1 - .9302)*(.9302)^{56}$.

NR(66) = .0292

The functions R(n) and NR(n) can be adjusted for any s-game hitting streak as follows:

For the **recursive function R(n)**, we

define $R(0) = R(1) = R(2) = … = R(s – 1) = 0, R(s) = U^s$ and
$R(n) = R(n – 1) + (1 – R(n – (s + 1)))*(1 – U)*U^s$; where n > s.

For the **piecewise function NR(n)**, we

let n be the number of games played by a player in a season.
NR(n) is the probability of at least one s-game hitting streak at some point in the first n games.
For 0 • n • (s – 1), NR(n) = 0.
For s • n • 2s, NR(n) = U^s + (n – s)*(1 – U)*U^s.
For (2s+1) • n • (3s+1),
NR(n) = U^s + (n – s)*(1 – U)*U^s • .5*(n – 2s)*(1 –U)*U^{2s}*[(n – 2s)*(1 – U)+1+ U]

The generalized <u>NR Function, NR(n),</u> is valid only when the number of the first n–games is less than or equal to 3s+1.

If s is replaced in the above function by 56, we obtain the function for the probability of having a 56-game hitting streak.

Since the number of games in a current season is equal to 162 and prior to 1961 was 154 games, the NR Function can be applied to both Joe DiMaggio's 56-game hitting streak and to Ted Williams' 84-game on-base streak or to any streak such that (3s+1) • 162. The NR Function can also be used if the number of the first n-games is less than or equal to 3s+1.

For streaks with smaller lengths s, if the number of games n exceeds 3s+1, NR(n) must be defined differently for each of the intervals [3s+2,4s+2], [4s+3,5s+3], [5s+4, 6s+4],….

For n > (3s+1), the definition of NR(n) for n is dependent on which of the above intervals contain n. Instead of defining a new function for each of them, it turns out that if we use the function NR(n) defined for n in the interval [**2s+1,3s+1]** in most cases the error will be very small.

We will now examine the error in using NR(n), defined for n in the interval [2s+1,3s+1], when, in fact, n > (3s+1). The letter n represents the first n-games a player played in a given season. We define the error by the absolute value of the difference between R(n) and NR(n) (Error = |R(n) – NR(n)|).

The Error = |R(n) – NR(n)|

The **Error Table** provides, for a given streak length from s=5 to s=52, the maximum value for U that would result in an error less than $1*10^{-4}$. For a given streak length, s, any U value less than or equal to the U value in Table 1 would also result in an error less than $1*10^{-4}$. Before 1961 a season consisted of 154 games. Starting in 1961, a season consisted of 162 games. Table 1 shows how changing from a season with 154 games to one with 162 games affects the value of U.The interpretation of this is if a player plays in fewer games a larger U can be tolerated. The U that works for n =162 games will also work for n less than 162 games. Also, as s increases, a larger U can be tolerated.

Table 1 – The Error Table

Streak Length	n=162	n=154	Streak Length	n=162	n=154
S	U	U	S	U	U
5	0.2374	0.2403	29	0.8412	0.8449
6	0.3076	0.3108	30	0.8484	0.8522
7	0.3701	0.3734	31	0.8552	0.8590
8	0.4249	0.4284	32	0.8616	0.8655
9	0.4729	0.4764	33	0.8677	0.8717
10	0.5150	0.5186	34	0.8735	0.8776
11	0.5521	0.5557	35	0.8791	0.8833
12	0.5848	0.5884	36	0.8844	0.8888
13	0.6139	0.6175	37	0.8896	0.8941
14	0.6399	0.6435	38	0.8945	0.8992
15	0.6632	0.6667	39	0.8993	0.9043
16	0.6842	0.6877	40	0.9040	0.9092
17	0.7032	0.7067	41	0.9086	0.9140
18	0.7205	0.7240	42	0.9130	0.9189
19	0.7363	0.7398	43	0.9175	0.9237
20	0.7507	0.7542	44	0.9219	0.9287
21	0.7640	0.7675	45	0.9263	0.9338
22	0.7762	0.7797	46	0.9307	0.9392
23	0.7876	0.7911	47	0.9353	0.9452
24	0.7981	0.8016	48	0.9401	0.9521
25	0.8079	0.8114	49	0.9453	0.9611
26	0.8170	0.8206	50	0.9511	0.9772
27	0.8256	0.8292	51	0.9581	
28	0.8337	0.8373	52	0.9679	

Therefore, if a player misses a few games, in a season, consisting of 162 games, the table value for U would still be valid.

If we choose a particular player who plays more than 3s+1 games and his value for U is less than or equal to the U in the table, we know that the error in using NR(n) will be less than $1*10^{-4}$.

In the next example, we look at the record streak of hitting a home run in eight consecutive games. This has been accomplished by three players. Dale Long did it in 1956, Don Mattingly in 1987, and Ken Griffey Jr. in 1993. For this streak, Step 1 in Example 1 becomes p = HR / PA and Step 4 in Example 1 becomes U^8.
For Dale Long (1956) we have:

p = 27 / 582 = .046; **j.kl** = 582 / 148 = 3.93; **s** = 8; **U** = .1702.
$NR(n) = U^s + (n - s)*(1 - U)*U^s \bullet \ .5*(n - 2s)*(1 - U)*U^{2s}*[(n - 2s)*(1 - U) + 1 + U]$.
NR(148) = $(.1702)^8 + (148 - 8)*(1 - .1702)*(.1702)^8 - .5*(148 - 16)*(1 - .1702)*(.1702)^{16}$ * [(148 – 16)*(1 - .1702) + 1 + .1702] = .0000826.

Dale Long had a 1 in 12,048 chance of achieving his streak in 1956.

Using both the NR(n) and R(n) functions, Table 2 provides the probability of each player achieving this streak in the year that they did.Table1 shows for s = 8 and U < 0.425 we would expect these two probabilities to differ by no more than $1*10^{-4}$. Clearly, the three players who hold the record all have a small enough value for U to satisfy the error condition. From Table 2, the two probabilities agree to seven decimal places. Table 2 shows that, because Griffey was a better home run hitter than Long, he was 26 times more likely to hit a home run in eight consecutive games.

Table 2

Year	Player	#G	#*PA*	#HR	U	NR(n) Function	R(n) Function
1956	Dale Long	148	582	27	0.1702	0.0000826	0.0000826
1987	Don Mattingly	141	629	30	0.1956	0.0002313	0.0002313
1993	Ken Griffey Jr	156	691	45	0.2575	0.0021428	0.0021428

Generalizing the concept of a streak

In this section, we expand the concept of a streak to other streaks in and outside of baseball.

Definitions

A **binomial experiment** consists of **k trials**, t_1, t_2, t_3, …, t_k. Each trial results in one of two mutually exclusive outcomes (one outcome is called success; the other outcome is called failure). The trials are independent and the probability of success is the same for each trial.

Success is associated with one of the two outcomes of a trial; **failure** is associated with the other outcome.

A **season** consists of a fixed number, N, of trials. The first trial is called trial number 1 and the last trial is trial number N.

Let **W** equal the number of successes in the season.

A **game**, G, consists of a certain number of trials in a season. Game #1 begins with trial number 1 and ends at a trial less than or equal to N. Game #2 begins with the first trial after the last trial of game #1. The games continue in this manner with the last game ending after the last trial. Every trial belongs to exactly one game. The number of trials in each game can be the same or can be different. A game can consist of exactly one trial.

The **number of games in a season** is **m**.

The length of the season is **N trials** and consists of **m** games. The games are sequential with the first trial appearing in game #1 and the last trial appearing in game number m.

Season = $t_1, t_2, t_3, \ldots, t_N$. ($t_i$ is either a success or failure)
Game 1 = $t_1, t_2, \ldots, t_k$
Game 2 = $t_{(k+1)}, t_{(k+2)}, \ldots, t_{(k+z)}$
Game 3 would start at $t_{(k+z+1)}$.
↓
Game m contains trial t_N as the last trial.

The **probability of sucess** for each trial is **p = W / N**.
The **number of trials used for each game** is **j.kl = N / m**. (j is the integer part; .kl is the decimal part)
The probability of at least one success in a game is

U = $\mathbf{(1-(1-p)^{j})*(1-.kl)+(1-(1-p)^{(j+1)})*.kl}$.

A streak of length s means having at least s consecutive games with at least one success in each game.

We assume independence between individual trials as well as between games. Even though individual games need not have the same number of trials, for our model, we will use, j.kl, the average number of trials per game for a season, for each game. Also, the same p will be used for each trial in each game in the season.

The following inputs and calculations are needed to evaluate the probability of a player duplicating a given streak.

Inputs and calculations

Inputs:
N = number of trials in a season
W = number of successes in a season
m = number of games in a season
s = length of the streak

Calculations:
p = W / N [probability of a success in a typical trial]
N / m = j.kl [number of trials in a typical game]
$\mathbf{U = (1-(1-p)^{j})*(1-.kl)+(1-(1-p)^{(j+1)})*.kl}$ [probability of at least one success in a typical game]

NR(n) is equal to the probability of at least one streak of length s in the first n-games. The piecewise function, NR(n), was defined on page 5.

When each plate appearance is considered a game (that is, N = m), we can look at streaks involving individual plate appearances. Since N = m and j.kl = N / m, we have j.kl = 1.00, j = 1,.kl = .00, and U = p. The next two examples look at this case.

Each Individual plate appearance is a game

Example 2: The current record of 12 consecutive plate appearances with a hit is held by Walt Dropo (1952). Pinky Higgins (1938) had 12 consecutive hits in 12 consecutive at-bats. Higgins actually needed 14 plate appearances to achieve his 12 consecutive hits. One can say that if we looked at 12 consecutive hits in 12 consecutive at-bats these two players would share the record. Example 4 examines this streak.

Both these players are unlikely candidates to hold the record. Higgins, a career .292 hitter, batted .303 in 1938. Dropo, a career .270 hitter, batted .276 in 1952. We will use the NR function, NR(n), to find the probability for Walt Dropo achieving this streak. A plate appearance is a trial; each plate appearance is a game; success is getting a hit in a plate appearance. Since m > 3s +1 for these two players we provide in Table 3 an error table for s = 12 and m = 162, m = 500, m = 600, and m = 700. Since the error |NR(n) – R(n)| decreases as s increases, Table 3 can be used for any s • 12.

Walt Dropo (1952 season):
N = PA = 633; W = H = 163; m = G = 633; s = 12
p = W / N = 163 / 633; j.kl = N / m = 633 /633 = 1.00;
$U = (1-(1-p)^{j})*(1-.kl)+(1-(1-p)^{(j+1)})*.kl = p = .257504$
$NR(n) = U^{s} + (n – s)*(1 – U)*U^{s}$ • .5*(n – 2s)*(1 – U)*U^{2s} *[(n – 2s)*(1 – U) + 1 + U]
NR(633) and R(633) give the probability of Dropo having 12 hits in 12 consecutive plate appearances.
NR(633) = .000039276
R(633) = .000039275
Error = |R(633) – NR(633)| = .000000001.

Since m < 700 and U < 0.50, from Table 3, we would expect the error, |R(633) – NR(633)| to be less than 0.0001.

The NR(n) function shows Walt Dropo's chance of achieving his streak of 12 consecutive hits in 12 consecutive plate appearances in 1952 is 1 in 25,641.

Table 3- Error Table for s = 12

U	m=162	m=500	m=600	m=700
0.50			0.0001	0.0001
0.51		0.0001	0.0001	0.0002
0.52		0.0001	0.0002	0.0003
0.53		0.0002	0.0004	0.0006
0.54		0.0004	0.0007	0.0011
0.55		0.0007	0.0012	0.0019
0.56		0.0012	0.0021	0.0034
0.57		0.0021	0.0037	0.0059
0.58	0.0001	0.0036	0.0063	0.0101
0.59	0.0001	0.0061	0.0106	0.0171
0.60	0.0002	0.0102	0.0178	0.0285
0.61	0.0004	0.0168	0.0294	0.0468
0.62	0.0006	0.0274	0.0477	0.0757
0.63	0.0010	0.0440	0.0764	0.1206
0.64	0.0016	0.0698	0.1205	0.1892
0.65	0.0026	0.1092	0.1872	0.2925
0.66	0.0041	0.1682	0.2868	0.4453
0.67	0.0065	0.2554	0.4326	0.6676
0.68	0.0100	0.3824	0.6430	0.9858
0.69	0.0153	0.5642	0.9414	1.4336
0.70	0.0230	0.8203	1.3581	2.0537

Example 3: The current record (since 1900) for consecutive plate appearances getting on base is 16 held by Ted Williams in 1957. For the 1957 season we have:
N = PA = 546; W = OB = H+ BB + HBP = 163+119+5 = 287; m = G = 546; S=16
p = W / N = .525641; j.kl = N / m = 546 / 546 = 1.00
U = p = .525641
NR(546) = .0085385
R(546) = .0085386
Error = |R(546) – NR(546)| = .0000001

The NR(n) function shows that Ted Williams' chance of getting on base in 16 consecutive plate appearances is 1 in 117.

Each Individual At-Bat is a Game

Example 4: As mentioned in Example 2, the record for consecutive hits in consecutive at-bats is 12 hits in 12 at-bats. This record is shared by Walt Dropo (1952) and Pinky Higgins (1938). For this streak, a plate appearance which is not an at-bat is considered a game in which the player did not play. For this reason N = AB is used instead of N = PA.

Walt Dropo (1952 season):

N = AB = 591; W = H = 163; m = G = 591; s = 12
p = W / N = 163 / 591; j.kl = N / m = 591 / 591 = 1.00;
$U = (1-(1-p)^{j})*(1-.kl)+(1-(1-p)^{(j+1)})*.kl = p = .275804$
$NR(n) = U^{s} + (n – s)*(1 – U)*U^{s}$ • .5*(n – 2s)*(1 – U)*U^{2s} *[(n – 2s)*(1 – U) + 1 + U]
NR(591) and R(591) give the probability of Dropo having 12 hits in 12 consecutive at-bats.
NR(591) = .0000814240
R(591) = .0000814243
Error = |R(591) – NR(591)| = .0000000003.
This translates into a 1 in 12,281 chance of Dropo achieving this streak in 1952.

Pinky Higgins (1938 season)
N = AB = 524; W = H = 159; m = G = 524; s = 12
p = W / N = 159 / 524; j.kl = N / m = 524 / 524 = 1.00;
$U = (1-(1-p)^{j})*(1-.kl)+(1-(1-p)^{(j+1)})*.kl = p = .303435$
$NR(n) = U^{s} + (n – s)*(1 – U)*U^{s}$ • .5*(n – 2s)*(1 – U)*U^{2s} *[(n – 2s)*(1 – U) + 1 + U]
NR(524) and R(524) give the probability of Higgins having 12 hits in 12 consecutive at-bats.
NR(524) = .00021787
R(524) = .00021786
Error = |R(591) – NR(591)| = .00000001.
This translates into a 1 in 4,590 chance of Higgins achieving this streak in 1938.

Comparing Ted Williams' 84-game consecutive on base streak to Joe DiMaggio's 56-game consecutive hitting streak

In 1941, two great players performed fabulous batting feats. Joe DiMaggio had his 56-game hitting streak and Ted Williams was the last major league player to bat over .400 for a season. During his 56-game hitting streak, Joe DiMaggio had 91 hits in 223 at-bats for an average of .408. In 1941, Ted Williams had a 23-game hitting streak, the only year that Ted had a hitting streak of at least 20 games. In 1949, Williams established the record of getting on base in 84 consecutive games. In 1941, wrapped around Joe DiMaggio's 56-game hitting streak, was a streak of 74 consecutive games getting on base. Getting on base means either reaching base by a hit, or a walk, or being hit by a pitch. Reaching base on an error or a fielder's choice does not count. Table 4 shows the batting statistics for DiMaggio (1941) and Williams (for the 1941 season and 1949 season). A success in a given plate appearance is different for the two streaks. For the 56-game streak, success is getting a hit; for the 84-game streak, success is getting on base. For the 56-game streak, p = (H / PA). For the 84-game streak, p = (OB / PA); where OB represents the total number of plate appearances in the season which resulted in the player getting on-base.

One question that is often asked is: Which of these two streaks would be hardest to duplicate? If the two streaks involved the same number of games the answer would be obvious. Since a player's season on base total is greater than or equal to his season hit total, the probability of a success in a plate appearance would be higher with the on-base streak. This would lead to a higher value for U and a higher NR(n) for the on base streak. However, since the streak of 84 consecutive games is 28 more games than

DiMaggio's 56-game hitting streak, this might lead to the belief that the 84-game streak would be harder to duplicate.

From Table 4, DiMaggio had a probability of 0.0001 of achieving his streak, whereas, Williams in 1949 had a probability of 0.09444 of achieving his streak. **For every 10,000 seasons, we would have expected DiMaggio in 1941 to accomplish his streak once. For every 10,000 seasons, we would have expected Williams in 1949 to accomplish his streak 944 times.** In 1941, DiMaggio's ratio of the probability of the 84-game streak to the probability of the 56-game streak was 56.70. This indicates that, for the year DiMaggio achieved his 56-game hitting streak, he was actually 56 times likelier to have the 84-game consecutive on base streak. In fact, in 1941, Williams' probability of achieving his 84-game consecutive on-base streak was 0.15700

Comparing Williams' two streaks of 16 consecutive plate appearances getting on base (Example 3) and his 84-game consecutive game on base streak, his probability of .09444 of achieving his 84-game streak in 1949 was 11 times likelier than his probability of .00854 of achieving his 16 consecutive plate appearance getting on-base streak in 1957. **Williams had a 1 in 11 chance of achieving his 84-game streak in 1949 and a 1 in 117 chance of achieving his 16 consecutive on base streak.**

Table 4 supports DiMaggio's streak as the harder one to duplicate. It is apparent that the real difference between the two players is their number of walks. A walk in a game extends the consecutive on base streak but has a negative effect on the consecutive hitting streak. If a player walks it uses up a plate appearance**.**

Table 4 - The 1941 Season

Player	Year	G	PA	AB	H	BB	BA	OBP	HBP	SH	SF	NR(n) Prob. 56 Streak	NR(n) Prob. 84 Streak	Ratio 84 Streak 56 Streak
DiMaggio	1941	139	621	541	193	76	0.357	0.440	4	0		0.00010	0.00565	56.50
Williams	1941	143	606	456	185	147	0.406	0.553	3	0		0.00002	0.15700	7850.00
Williams	1949	155	730	566	194	162	0.343	0.490	2	0		0.00001	0.09444	13491.43

We will now look at other great hitters to compare their probabilities of achieving these two streaks.

These two streaks evaluated for other great hitters

In the history of baseball, one of the greatest 5-year spans of hitting belongs to Rogers Hornsby. For the years from 1921 to 1925 his cumulative batting average was .402. Table 5 shows Hornsby's statistics for these five years. Hornsby had a 33-game hitting streak in 1922. From Table 5, we can see that in 1922 Hornsby had his highest probability of achieving the 56-game hitting streak.

Using the ratio of the probability of Hornsby having an 84-game on base streak to the probability of him having a 56-game hitting streak, we see the ratio went from a low of 4 to a high of 47. For each year, Hornsby had a higher probability of achieving the 84-

game consecutive on-base streak. Hornsby's statistics provide further evidence that DiMaggio's 56-game streak would be harder to duplicate.

Table 5 – Rogers Hornsby

Player	Year	G	PA	AB	H	BB	BA	OBP	HBP	SH	SF	Prob. 56 Streak	Prob. 84 Streak	Ratio 84 Streak 56 Streak
Hornsby	1921	154	674	592	235	60	0.397	0.458	7	15		0.00121	0.00763	6.31
Hornsby	1922	154	704	623	250	65	0.401	0.459	1	15		0.00356	0.01475	4.14
Hornsby	1923	107	487	424	163	55	0.384	0.459	3	5		0.00053	0.00762	14.38
Hornsby	1924	143	640	536	227	89	0.424	0.507	2	13		0.00225	0.05804	25.80
Hornsby	1925	138	605	504	203	83	0.403	0.489	2	16		0.00043	0.02052	47.72

Tables 6 and 7 present the batting statistics for two players recognized by many baseball people as the two best hitters since 2000. These players are Ichiro Suzuki and Albert Pujols.

Ichiro Suzuki throughout his baseball career in Japan was recognized as a superstar. He led the Japanese League in batting average for each year he played in Japan. In 2001, at the age of 27, he began his career in the United States. In each of his first eight years, he had over 200 hits. Ichiro is a lead-off hitter who had over 725 plate appearances for each year. He received relatively few walks. He batted from the left side. Even though Suzuki hit home runs, he was known more as a singles hitter with great speed who could beat out many infield singles. Before 2009, Ichiro had logged five hitting streaks of at least 20 games, which tied him with Ty Cobb who also had five 20+ hitting streaks. The record is seven 20+ hitting streaks held by Pete Rose. In 2001, Ichiro had a 23-game hitting streak and a 21-game hitting streak. In 2004, he had a 21-game streak; in 2006, he had a 20-game streak; and in 2007, he had a 25-game streak. From Table 6, we see that for the years 2001, 2004, 2006, and 2007, Ichiro had his highest probabilities of achieving a 56-game hitting streak. For the year 2004, Ichiro had his highest probability of achieving both streaks. What we find interesting is that for the years 2001 and 2003 his probability of achieving the 56-game hitting streak was higher than his probability of achieving the 84-game on base streak. Looking at the ratio of the probability of the 84-game streak to the probability of the 56-game streak, for all but one year it was at or below three.

Albert Pujols started his career in 2001 at the age of 21. He was nicknamed "the machine" due to his great hitting ability. In contrast to Suzuki, Pujols batted from the right side and is known as a power hitter. Albert was feared by pitchers, which led to his receiving many walks each season. He batted either in the third spot or fourth spot in the batting order. Except for two years, his number of plate appearances for a season was in the upper 600s to 700. The last column of Table 7 shows that the ratio of his probability of achieving an 84-game on-base streak to his probability of achieving a 56-game hitting streak ranged from 70 to over 1000.

Both players provide us with more evidence favoring the 56-game hitting streak as the streak that is the more difficult to duplicate.

Table 6 – Ichiro Suzuki

Player	Year	G	PA	AB	H	BB	BA	OBP	HBP	SH	SF	Prob. 56 Streak	Prob. 84 Streak	Ratio 84 Streak 56 Streak
Suzuki	2001	157	738	692	242	30	0.350	0.381	8	4	4	0.00116	0.00055	0.47
Suzuki	2002	157	728	647	208	68	0.321	0.388	5	3	5	0.00003	0.00065	21.67
Suzuki	2003	159	725	679	212	36	0.312	0.352	6	3	1	0.00004	0.00003	0.75
Suzuki	2004	161	762	704	262	49	0.372	0.414	4	2	3	0.00360	0.00523	1.45
Suzuki	2005	162	739	679	206	48	0.303	0.350	4	2	6	0.00001	0.00003	3.00
Suzuki	2006	161	752	695	224	49	0.322	0.370	5	1	2	0.00011	0.00024	2.18
Suzuki	2007	161	736	678	238	49	0.351	0.396	3	4	2	0.00052	0.00085	1.63
Suzuki	2008	162	749	686	213	51	0.310	0.361	5	3	4	0.00003	0.00008	2.67

Table 7 – Albert Pujols

Player	Year	G	PA	AB	H	BB	BA	OBP	HBP	SH	SF	Prob. 56 Streak	Prob. 84 Streak	Ratio 84 Streak 56 Streak
Pujols	2001	161	676	590	194	69	0.329	0.403	9	1	7	0.000004	0.00028	70.00
Pujols	2002	157	675	590	185	72	0.314	0.394	9	0	4	0.000002	0.00023	115.00
Pujols	2003	157	685	591	212	79	0.359	0.439	10	0	5	0.000060	0.00482	80.33
Pujols	2004	154	692	592	196	84	0.331	0.415	7	0	9	0.000010	0.00198	198.00
Pujols	2005	161	700	591	195	97	0.330	0.430	9	0	3	0.000004	0.00295	738.50
Pujols	2006	143	634	535	177	92	0.331	0.431	4	0	3	0.000005	0.00328	656.00
Pujols	2007	158	679	565	185	99	0.327	0.429	7	0	8	0.000002	0.00218	1090.00
Pujols	2008	148	641	524	187	104	0.357	0.462	5	0	8	0.000011	0.01129	1026.45

The next two tables will examine all players, since 1900, with the longest consecutive hitting streak and the longest consecutive on-base streak.

Table 8 – Longest Consecutive Game Hitting Streaks

Streak Length	Player	Year	G	PA	AB	H	BB	BA	OBP	HBP	SH	SF	Prob. 56 Streak	Prob. 84 Streak	Ratio 84 Streak 56 Streak
56	DiMaggio	1941	139	620	541	193	76	0.357	0.440	4	0		0.000105	0.005650	53.81
44	Rose	1978	159	729	655	198	62	0.302	0.362	3	2	7	0.000010	0.000074	7.40
41	Sisler	1922	142	654	586	246	49	0.420	0.467	3	16		0.010570	0.019085	1.81
40	Cobb	1911	146	654	591	248	44	0.420	0.467	8	11		0.008430	0.015280	1.81
39	Molitor	1987	118	542	465	164	69	0.353	0.44	2	5	1	0.000072	0.004460	61.94
37	Holmes	1945	154	713	636	224	70	0.352	0.420	4	3		0.000326	0.004140	12.70
36	Rollins	2005	158	732	677	196	47	0.290	0.338	4	2	2	0.000005	0.000012	2.40

Table 9 – Longest Consecutive Game On Base Streaks

Streak Length	Player	Year	G	PA	AB	H	BB	BA	OBP	HBP	SH	SF	Prob. 56 Streak	Prob. 84 Streak	Ratio 84 Streak 56 Streak
84	Williams	1949	155	730	566	194	162	0.343	0.490	2	0		0.000007	0.094446	13492.29
74	DiMaggio	1941	139	620	541	193	76	0.357	0.440	3	0		0.000105	0.005650	53.81
69	Williams	1941	143	606	456	185	147	0.406	0.553	3	0		0.000024	0.157000	6541.67
65	Williams	1948	137	638	509	188	126	0.369	0.490	3	0		0.000067	0.081327	1213.84
63	Cabrera	2006	153	675	607	171	51	0.282	0.335	3	3	11	0.000000	0.000002	6.33
58	Snider	1954	149	679	584	199	84	0.341	0.423	4	1	6	0.000041	0.003813	92.55
58	Bonds	2003	130	550	390	133	148	0.341	0.529	10	0	2	0.000000	0.073201	3660050
57	Boggs	1985	161	758	653	240	96	0.368	0.450	4	3	2	0.000587	0.024690	42.06

In Table 8, every ratio of the probability of an 84-game on base streak to the probability of a 56-game hitting streak is greater than 1.80. This shows that even for those players with the longest consecutive hitting streaks, there was a higher probability of achieving the 84-game consecutive on-base streak.

In Table 9, every ratio of the probability of an 84-game on base streak to the probability of a 56-game hitting streak is greater than 6.33. This shows that for the players with the longest consecutive on-base streaks, their probability of achieving the 84-game consecutive on-base streak was considerably higher than their probability of achieving the 56-game consecutive hitting streak.

Comparing Table 8 to Table 9 we observe

- Joe DiMaggio is the only player in both tables.
- Ted Williams appears three times in Table 9.
- Excluding Joe DiMaggio, every player in Table 8 had fewer than 71 walks.
- Excluding Joe DiMaggio and Orlando Cabrera, every player in Table 9 had at least 84 walks.
- Orlando Cabrera's appearance in Table 9 is unusual considering he had only 51 walks in 2006. Cabrera had a 1 in 500,000 chance of achieving the 84-game on-base streak.

Tables 8 and 9 provide further evidence that the 56-game hitting streak is more difficult to duplicate than the 84-game consecutive on-base streak.

We now turn to some streaks that involve more than one success in each game. We give two examples next.

Consecutive game streaks <u>with more than one success</u> in each game:

Most consecutive games with two or more hits in each game – The record is 13 held by Rogers Hornsby (1923).
Most consecutive games with three or more hits in each game – The record is 6 held by two players (since 1900). The two players were George Brett (1976) and Jimmy Johnston (1923).

Define:
p = H / PA
j.kl = PA / G [j.kl is the average number of plate appearances (trials) used for each game]
x = the minimum number of successes in each game of the streak
U = [Pr(X• x), for j trials]*(1 • .kl) + [Pr(X• x), for (j+1) trials]*.kl [since the assumptions for this model satisfy the Binomial Model, the two probabilities can be calculated using the Binomial Formula]
$_mC_k$ = the number of combinations of m distinct objects taken k at a time.

We now look at two cases for a minimum number of x successes in a game, based on a fixed number of trials. A trial corresponds to a plate appearance.

Case 1: If x = 1:
Pr(X • 1, for j trials per game) = 1 – Pr(X • (x – 1))
= 1 – Pr(X • (1 – 1); for j trials) = 1 – Pr(X=0)
= 1 – $_jC_0$* $p^0(1 – p)^j$ = (1 – (1 – p)^j)
Pr(X • 1, for (j+1) trials per game) = 1 – Pr(X • (x – 1))
= 1 – Pr(X • (1 – 1); for (j+1) trials) = 1 – Pr(X=0)
= 1 – $_{(j+1)}C_0$* $p^0(1 – p)^{(j+1)}$ = (1 – (1 – p)$^{(j+1)}$) [$_jC_0$ = 1 and $_{(j+1)}C_0$ = 1]

Up to now, the only streaks studied required x =1 (at least one success in each game of the streak). This is what is used in the calculation for U in the 56-game hitting streak and all other streaks discussed before in this paper.

Case 2: If x>1:
Pr(X • x, for j trials per game) = 1 – Pr(X • (x – 1), for j trials)
Pr(X • x, for (j+1) trials per game) = 1 – Pr(X • (x – 1), for (j+1) trials)
You can either use a Binomial Table or the Binomial Formula in Excel to calculate Pr(X • (x – 1)).

We will look at the example of Rogers Hornsby's 13-game consecutive streak of at least two hits in each game which he accomplished in 1923. From Table 5, the statistics for Hornsby in 1923 were

G = 107, PA = 487, H = 163
p = 163/487 = .335, j.kl = 487/107 = 4.55
U = [Pr(X • 2), for j = 4 trials]*(1 - .55) + [Pr(X• 2), for j =5 trials]*(.55)
U = [1 - Pr(X • 1), for j = 4 trials]*(1 - .55) + [1 - Pr(X• 1), for j =5 trials]*(.55)

Using the binomial formula in Excel, we calculated the above probabilities. The resulting value for U was .482416. Using .482416 for U, the probability of Hornsby achieving the 13-game streak of at least two hits in each game for the 1923 season was R(107) = .0038003 and NR(107) = .0038003.

From Table 5, the probability of Hornsby achieving the 56-game hitting streak in 1923 was NR(107) = .00053. **Hornsby had a 1 in 263 chance of achieving the 13-game streak and a 1 in 1,886 of achieving the 56-game streak.**

The ratio of .0038003 to .00053 is over 7. The interpretation is, in 1923, Hornsby had seven times the chance of duplicating his 13-game streak than duplicating the 56-game hitting streak.

Conclusion

Two functions were provided to calculate the probability of various players duplicating various streaks in baseball. The two functions are a recursive function named R(n) and a direct piecewise function called NR(n). Both functions have for their domains the first n games of a player's season. The piecewise function NR(n) is easier to use. However, the NR(n) function is correct only when n is less than or equal to 3s +1 where s is the length of the streak. Since a player's season consists of 162 games or less, for such streaks as Joe DiMaggio's 56-game hitting streak and Ted Williams' 84-game consecutive on-base streak, the NR(n) function will equal the R(n) function.

For streaks with a small length s, even though the NR(n) function has an error when n > 3s +1, for many streaks this error is very small.

Two categories of streaks were discussed. The first category was streaks based on games; the second category was streaks based on plate appearances. Ted Williams had record streaks of both types. He holds the record for consecutive games getting on base at least one time (84 games in 1949). He also holds the modern day record of consecutive plate appearances getting on base (16 times in 1957). Using the NR(n) and R(n) functions, it was shown that Williams' consecutive plate appearance on base streak would be harder to duplicate than the 84-game streak, based on the years in which Williams accomplished these two streaks.

Much of the paper compares DiMaggio's 56-game hitting streak to Williams' 84-game consecutive on base streak. The probability of achieving each of these two streaks is calculated for several great hitters of the past and present. The results indicated that of these two streaks DiMaggio's 56-game hitting streak would be the hardest to duplicate.

The paper also looks at game streaks involving more than one success in each game. The probability of achieving such streaks as Rogers Hornsby's consecutive 13-game streak of getting at least two hits in each game is calculated.

We end this paper with these questions. As mentioned in this paper, for streaks with small lengths and n > 3s+1, each of the intervals [3s+2,4s+2], [4s+3,5s+3], [5s+4,

6s+4],… requires its own function NR(n). Can you produce a piecewise function that will work for each of these intervals? Another question is in comparing the 56-game consecutive hitting streak to the 84-game consecutive on base streak, can we find a function to calculate for any player, how many games must be added to 84 games or subtracted from 84 games so that the probability of that player achieving both streaks is the same? Finally, we ask for which independent variables is the rate of change greatest for the function $NR(n, U, s) = U^s + (n – s)*(1 – U)*U^s \bullet .5*(n – 2s)*(1 – U)*U^{2s} *[(n – 2s)*(1 – U) +1 + U]$; $0 \bullet U \bullet 1$, $2s+1 \bullet n \bullet 3s+1$, and $s \bullet 5$?

As mentioned in the introduction, we are looking for applications, outside the world of baseball, which can apply the models for streaks developed in this paper.

Reference

1. Michael Freiman, *56-Game Hitting Streaks Revisited*, The Baseball Journal, SABR, 31 (2003) 11.15.

STANLEY ROTHMAN received his BS from Montclair State University in 1966 and his MS and Ph.D. from the University of Wisconsin in 1967 and 1970. His entire teaching and research career has been spent at Quinnipiac University. From 1970 to 1977, he was an Assistant Professor of Mathematics, from 1978 to 1984, he was an Associate Professor of Mathematics. Since 1985, he has been a full professor of mathematics. From 1992 he has served as Chairman of the Mathematics Department. In 1973, he founded a computer consulting firm called Acc-U-data. He served as President from 1973 to 2005. Since 2006, his major research interest has been in Sabermetrics. He spoke at both the San Diego and Washington D.C. National Mathematics Conferences in 2008 and 2009. The theme of both talks was the teaching of statistics from baseball records. He is now finishing his new book *Basic Statistics: Using Baseball to Bring Statistics to Life* which will be published by Johns Hopkins Press in 2010.

QUOC LE is a senior joint major in mathematics and finance at Quinnipiac University. He is in the process of applying to various graduate schools. His future plans include obtaining a Ph.D. and teaching at the university level.

Stanley Rothman can be reached by email at Stanley.rothman@quinnipiac.edu and by phone at 203-582-8751.