Let’s Go Streaking!

Probability & Hitting Streaks in Baseball

Phil Yates
DePaul University
October 28, 2025

1941 Baseball Season

PLAYER	BA	HR	RBI	AB	H	OBP	SLG	WAR
Player A	0.357	30	125	541	193	0.440	0.643	9.1
Player B	0.406	37	120	456	185	0.553	0.735	10.6

Who are these players?

Joe DiMaggio (Player A – won the AL MVP) and Ted Williams (Player B)

Longest Hitting Streaks in MLB History

YEAR	NAME	TEAM	GAMES
1941	Joe DiMaggio	New York Yankees	56
1896-97	Willie Keeler	Baltimore Orioles	45
1978	Pete Rose	Cincinnati Reds	44
1894	Bill Dahlen	Chicago Colts (Cubs)	42
1922	George Sisler	St. Louis Browns	41
1911	Ty Cobb	Detroit Tigers	40
1987	Paul Molitor	Milwaukee Brewers	39
2005-06	Jimmy Rollins	Philadelphia Phillies	38
1945	Tommy Holmes	Boston Braves	37
1896-97	Gene DeMontreville	Washington Senators	36

Source: Baseball Reference

Hitting Streaks

What is a hitting streak?

- the number of consecutive official games in which a player appears **and** gets at least one base hit
- the streak is ended when a player has **at least 1 plate appearance** and no hits
- the streak is **not** terminated if all official plate appearances result in a base on balls (a walk), hit by pitch, defensive interference, or a sacrifice bunt
- the streak shall terminate if the player has a sacrifice fly and no hit

Source: Official Rules, Major League Baseball, 10.23 Guidelines for Cumulative Performance Records

Probability and Hitting Streaks

Sample Space

All possible outcomes in an experiment. Typically denoted by Ω.

Experiment: A batter has four at bats in a single game. Each at bat can be a hit (H) or an out (O).

Sample Space Ω: 16 total outcomes

- "No hits in four at bats" - OOOO
- "One hit in four at bats" - HOOO,OHOO,OOHO,OOOH
- "Two hits in four at bats" - HHOO,HOHO,HOOH,OHHO,OHOH,OOHH
- "Three hits in four at bats" - HHOO,HHOH,HOHH,OHHH
- "Four hits in four at bats" - HHHH

Probability and Hitting Streaks

Complement of an Event A

The set of all outcomes that are in the sample space Ω but are not in the event A. Typically denoted as A^c.

Event A: A player gets at least one hit in four at bats in a single game.

- "One hit in four at bats" - HOOO,OHOO,OOHO,OOOH
- "Two hits in four at bats" - HHOO,HOHO,HOOH,OHHO,OHOH,OOHH
- "Three hits in four at bats" - HHOO,HHOH,HOHH,OHHH
- "Four hits in four at bats" - HHHH

Complement A^c: A player gets zero hits in four at bats in a single game.

- "No hits in four at bats" - OOOO

Probability and Hitting Streaks

Rules of Probability

Complement Rule:

$$P(A^C) = 1 - P(A)$$

Multiplication Rule for Independent Events: Let A and B be two independent events. Then

$$P(A \text{ and } B) = P(A) \times P(B)$$

What is the probability that a player with a 0.300 batting average gets at least one hit in four at bats in a single game?

$$P(\text{at least one H in 4 at bats}) = 1 - P(\text{zero H's in 4 at bats}) = 1 - 0.7^4 = 0.7599$$

Probability and Hitting Streaks

What is the probability that a player with a 0.300 batting average gets at least one hit in four at bats in two consecutive games?

The player needs to get at least one hit in four at bats in the first game **and** at least one hit in four at bats in the second game.

$$0.7599 \times 0.7599 = (0.7599)^2 = 0.5774$$

Five consecutive games?

$$0.7599^5 = 0.2534$$

Ten consecutive games?

$$0.7599^{10} = 0.0642$$

Fifty-six consecutive games?

$$0.7599^{56} = 0.0000002101$$

Joe DiMaggio: Probability of 56-Game Hitting Streak

In 1941, Joe DiMaggio had 541 at bats in 139 games.

3.892 at bats per game -- let's round to 4 at bats (can't have a fraction of an at bat)

What is the probability that Joe DiMaggio, a 0.357 hitter in 1941, gets at least one hit in four at bats in fifty-six consecutive games?

$$P(H) = 0.357 \implies P(O) = 0.643$$

Probability of at least one H in 4 at bats in one game?

$$1 - 0.643^4 = 0.8290599$$

Fifty-six consecutive games?

$$0.8290599^{56} = 0.00002759$$

Potential Problems

One big assumption that we made in the previous calculation?

- Joe DiMaggio had 4 at bats **in every single game during the streak!**

Why is this a problem? Let us assume he had 8 at bats in 2 games.

Game 1 AB's	Game 2 AB's	Prob. of H's in both games
1	7	$(1-0.643^{1})\times(1-0.643^{7})=$**0.3408**
2	6	$(1-0.643^{2})\times(1-0.643^{6})=$**0.5451**
3	5	$(1-0.643^{3})\times(1-0.643^{5})=$**0.6535**
4	4	$(1-0.643^{4})\times(1-0.643^{4})=$**0.6873**
5	3	$(1-0.643^{5})\times(1-0.643^{3})=$**0.6535**
6	2	$(1-0.643^{6})\times(1-0.643^{2})=$**0.5451**
7	1	$(1-0.643^{7})\times(1-0.643^{1})=$**0.3408**

Potential Problems

By assuming the same number of at bats in each game, it **inflates** or potentially **overestimates** the likelihood of a hitting streak

Solution?

Vary the at bats for each game1

During the 56-game hitting streak, DiMaggio had:

- 3 games with 2 at bats
- 11 games with 3 at bats
- 26 games with 4 at bats
- 16 games with 5 at bats

Source: Cliff Blau, a member of SABR (Society for American Baseball Research)

Joe DiMaggio: Probability of 56-Game Hitting Streak

When varying the at bats based on his **actual** at bats during the hitting streak, we can find the probability of a 56-game hitting streak:

$$\left(1-0.643^2\right)^3 \times \left(1-0.643^3\right)^{11} \times \left(1-0.643^4\right)^{26} \times \left(1-0.643^5\right)^{16}$$

Probability? **0.000007993**

Notice this is much smaller than **0.00002759** -- the probability of the 56-game hitting streak when assuming 4 at bats per game

Problem?

This probability is specific to these 56 games and not necessarily **any** 56 consecutive games over the course of an **entire** baseball season

Simulation: Another Way To Estimate!

Let's say we have a player that has a 0.333 batting average and has 4 at bats each game. The baseball season has 162 games in a season (154 when DiMaggio played). How can we estimate the probability of this player having a hitting streak as long as Joe DiMaggio's 56 game hitting streak?

We can use dice (or some other chance mechanism) to "simulate" the season!

Sample space for a single 6-sided die?

$$\Omega = \{\mathbf{1}, \mathbf{2}, 3, 4, 5, 6\}$$

The bold numbers are "hits" and the regular numbers are "outs."

Each person should roll the die four times. That would be a "game." If a **1** or a **2** occurs in **any** of the four rolls of a die, the player had a hit in that game.

Problem?

This is just one simulated "season." To estimate the probability we would need to repeat this process thousands of times.

Simulation: Another Way To Estimate!

Assuming consecutive bats are independent of one another and the batting average ("probability of getting a hit in an at bat") is the same for every at bat, the number of hits in a game follows a **binomial distribution** with:

- n - number of trials - the number of at bats in a given game
- p - probability of success - the probability of getting a hit in an at bat

For each player in a given season, n will vary from game to game while p will be treated as constant.

How do we do this?

Using a computer program (statisticians love R!)

Simulation of Joe DiMaggio's 1941 Season

- Run k simulated "baseball season." This could be set to a number like 1000 or 10,000
- Let $n_1, n_2, \ldots, n_{139}$ be the **actual** at bats during DiMaggio's 1941 season. A simulated baseball season will sample **with replacement** these at bats 139 times.
- For each "game" in the simulated baseball season, we will random generate the number of hits in each game by using a binomial distribution, where the probability of success is $p = 0.357$.
- We will look at the longest streak in each simulated baseball season. This streak is the number of consecutive "games" with a hit. We will have k of these longest streaks!
- Estimate the probability of having a hitting streak of 56 games by counting the number of simulated seasons with a streak of 56 consecutive games or longer and divide by the number of simulated seasons, k
- But really we are interested in the probability of there ever being a 56-game hitting streak by **any** player achieving it over a given 56-game stretch. How do we do this?

Simulation & Retrosheet Data

Rockoff and Yates (2009, 2011) analyzed play-by-play data from Retrosheet.org for:

- National League only: 1911, 1921, 1922, 1953
- American and National League: 1920-1929, 1954-2007

Since then they have added 1911 (American League), 1911-1919 (both leagues), 1921-1922 (American League), 1930-1952 (both leagues), 1953 (American League), 2008-2024 (both leagues)

These seasons have been added to update the results from Rockoff and Yates

Simulation & Retrosheet Data

The process for hitter i in season j who plays in k games that season:

- $\mathbf{AB}_{ij} = \left(AB_{ij1}, AB_{ij2}, \ldots, AB_{ijk}\right)$
- The number of hits a player i in season j gets in game k (assuming that at-bats over the course of a single game are independent of each other)

$$H_{ijk} \sim \text{Binomial}\left(AB_{ijk}, p_{ij}\right)$$

- A simulated season's worth of at bats are the at bats in each season sampled with replacement.

$$\mathbf{AB}^*_{ij} = \left(AB^*_{ij1}, AB^*_{ij2}, \ldots, AB^*_{ijk}\right)$$

- If m seasons are simulated, then for player i and season j:

$$\mathbf{AB}^1_{ij}, \mathbf{AB}^2_{ij}, \ldots, \mathbf{AB}^m_{ij}$$

Simulation & Retrosheet Data

- The number of hits a player gets in each game in the m^{th} simulation season is

$$\mathbf{H}_{ij}^{*m} \sim \text{Binomial}\left(\mathbf{AB}_{ij}^{*m}, p_{ij}\right)$$

- Any run of hits in $\mathbf{H}_{ij}^{*m}$ that are greater than zero denotes a hitting streak. The simulations will keep track of each player's maximum hitting streak in any given simulated season.

Simulation & Retrosheet Data

This is from 2024. 222,544 rows of data and 161 variables.

`2024plays`

Filter | Cols: « < 1 – 50 > »

	gid	event	inning	top_bot	vis_home	site	batteam	pittea
1	ANA202404050	6/P78S	1	0	0	ANA01	BOS	ANA
2	ANA202404050	NP	1	0	0	ANA01	BOS	ANA
3	ANA202404050	W	1	0	0	ANA01	BOS	ANA
4	ANA202404050	8/F89	1	0	0	ANA01	BOS	ANA
5	ANA202404050	NP	1	0	0	ANA01	BOS	ANA
6	ANA202404050	K	1	0	0	ANA01	BOS	ANA
7	ANA202404050	K	1	1	1	ANA01	ANA	BOS
8	ANA202404050	NP	1	1	1	ANA01	ANA	BOS
9	ANA202404050	K	1	1	1	ANA01	ANA	BOS
10	ANA202404050	2/P2F/FL	1	1	1	ANA01	ANA	BOS
11	ANA202404050	HR/F89XD	2	0	0	ANA01	BOS	ANA
12	ANA202404050	HR/F78XD	2	0	0	ANA01	BOS	ANA
13	ANA202404050	31/G34	2	0	0	ANA01	BOS	ANA

Retrosheet Data in R

```
library(tidyverse)

`2024plays` <- read.csv("2024plays.csv")

data2024 <- `2024plays` %>%
  filter(gametype=="regular") %>%
  group_by(batter) %>%
  mutate(h=single+double+triple+hr) %>%
  summarize(H=sum(h))

data2024.abs <- `2024plays` %>%
  filter(gametype=="regular") %>%
  group_by(batter,gid) %>%
  summarize(ab=sum(ab)) %>%
  mutate(game=1:n())

data2024.abs.wide <- data2024.abs %>%
  pivot_wider(id_cols=batter,values_from=ab,names_from=game,names_prefix="AB")

data2024 <- data2024 %>%
  inner_join(data2024.abs.wide,by="batter")

data2024 <- data.frame(Year=rep(2024,nrow(data2024)),data2024)

data2024 <- data2024 %>% filter(H>0)
```

Simulating Streaks in R

```
streaks <- function(sim){
data2024 <- read.csv("2024.csv",header=T,sep=",")

n <- nrow(data2024)
g <- ncol(data2024)
streaks.max.vector <- rep(0,n*sim)
streaks.max.matrix <-
  matrix(streaks.max.vector,ncol=sim)

# Simulations
for(k in 1:sim){

for(i in 1:n){
# Find player's At Bats
ab <- data2024[i,4:g]
tab <- t(ab)
tab <- na.omit(tab)
ab <- t(tab)

# Batting Average
avg <- data2024[i,3]/sum(ab)

# Simulated data
ab.sim <- sample(ab,length(ab),replace=TRUE)
hit.sim <- rbinom(n=ab.sim,size=ab.sim,p=avg)
```

```
# Figuring out streaks
streaks <- rep(0,80)
streak.number <- 1

for(j in 1:length(ab.sim)){
if(hit.sim[j]>0) streaks[streak.number] <-
    streaks[streak.number]+1
else streak.number <- streak.number+1
}

streaks.max.matrix[i,k] <- max(streaks)
}
}
return(streaks.max.matrix)
}
```

Results: 1000 Baseball "Histories"

		SIMULATED STREAKS								ACTUAL SEASON					
STREAK	PLAYER	YEAR	40+	50+	56+	MIN	Q1	Q2	Q3	MAX	AB	AVG	HR	RBI	OBP	STREAK
89	George Sisler	1922	47	12	7	11	19	24	29	89	586	0.420	8	105	0.467	41
84	George Sisler	1920	77	22	13	10	20	24	30	84	631	0.407	19	122	0.449	25
79	Al Simmons	1925	60	10	8	9	19	23	29	79	654	0.387	24	129	0.419	23
77	Ralph Garr	1974	6	2	1	8	15	18	22	77	606	0.353	11	54	0.383	14
74	Ichiro Suzuki	2004	31	4	2	10	18	22	27	74	704	0.372	8	60	0.414	21
70	Rogers Hornsby	1922	57	9	4	10	19	23	29	70	623	0.401	42	152	0.459	33
70	Ichiro Suzuki	2009	17	4	1	10	16	20	25	70	639	0.352	11	46	0.386	27
68	Pie Traynor	1923	4	1	1	7	14	16	20	68	616	0.338	12	101	0.377	24
68	Zack Wheat	1924	18	4	1	9	16	19	24	68	566	0.375	14	97	0.428	24
68	Freddie Lindstrom	1928	18	1	1	9	16	19	24	68	646	0.358	14	107	0.383	14
68	Jeff Heath	1941	2	1	1	7	13	15	18	68	585	0.340	24	123	0.396	19
52	Joe DiMaggio	1941	5	1	0	8	14	16	20	52	541	0.357	30	125	0.440	56

Results: 1000 Baseball "Histories"

Hitting streaks in 1000 Simulated Baseball Histories

Max	40+	50+	56+
89	979	357	131

The estimated probability of a single player having a hitting streak of at least 56 games:

$$\frac{140}{47,514,044} = 0.000002946497 = 0.0002946497\%$$

The estimated probability of a hitting streak of at least 56 games happening at **some point** in baseball history (1911–2024):

$$\frac{131}{1000} = 0.131 = 13.1\%$$

Simulation of Hitting Streaks

One of the big assumptions to the previous simulations and probability calculations?

- The batting average ("probability of getting a hit") is constant **for every single at bat** over the course of a season

Solution?

- Vary the batting average over the course of a simulated season!
- There are a variety of ways to do this!

Another **huge** assumption in these calculations?

- At bats are independent of one another. Why might this not be a good assumption to make? Is it a reasonable assumption to make?