# Where Have You Gone, Joe DiMaggio?

## Probability and Hitting Streaks

Phil Yates

Department of Mathematical Sciences
DePaul University

February 16, 2018

# 1941 Baseball Season

| Player | BA | HR | RBI | AB | H | OBP | SLG | WAR |
|---|---|---|---|---|---|---|---|---|
| Player A | .357 | 30 | 125 | 541 | 193 | .440 | .643 | 9.1 |
| Player B | .406 | 37 | 120 | 456 | 185 | .553 | .735 | 10.6 |

# 1941 Baseball Season

| Player | BA | HR | RBI | AB | H | OBP | SLG | WAR |
|---|---|---|---|---|---|---|---|---|
| Player A | .357 | 30 | 125 | 541 | 193 | .440 | .643 | 9.1 |
| Player B | .406 | 37 | 120 | 456 | 185 | .553 | .735 | 10.6 |

Who are these players?

# 1941 Baseball Season

| Player | BA | HR | RBI | AB | H | OBP | SLG | WAR |
|---|---|---|---|---|---|---|---|---|
| Player A | .357 | 30 | 125 | 541 | 193 | .440 | .643 | 9.1 |
| Player B | .406 | 37 | 120 | 456 | 185 | .553 | .735 | 10.6 |

Who are these players?

# 1941 Baseball Season

| Player | BA | HR | RBI | AB | H | OBP | SLG | WAR |
|---|---|---|---|---|---|---|---|---|
| Player A | .357 | 30 | 125 | 541 | 193 | .440 | .643 | 9.1 |
| Player B | .406 | 37 | 120 | 456 | 185 | .553 | .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: MLB.com

# Hitting Streaks

What is a hitting streak?

# 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

# 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

# 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

# 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

Basic concepts of probability needed to analyze hitting streaks:

**Sample Space**

All possible outcomes in an experiment. Denoted by $S$

# Probability and Hitting Streaks

Basic concepts of probability needed to analyze hitting streaks:

**Sample Space**

All possible outcomes in an experiment. Denoted by $S$

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

# Probability and Hitting Streaks

Basic concepts of probability needed to analyze hitting streaks:

**Sample Space**

All possible outcomes in an experiment. Denoted by $S$

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 $S$: 16 total outcomes

# Probability and Hitting Streaks

Basic concepts of probability needed to analyze hitting streaks:

**Sample Space**

All possible outcomes in an experiment. Denoted by $S$

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 $S$: 16 total outcomes

- "No hits in four at bats" – OOOO

# Probability and Hitting Streaks

Basic concepts of probability needed to analyze hitting streaks:

**Sample Space**

All possible outcomes in an experiment. Denoted by $S$

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 $S$: 16 total outcomes

- "No hits in four at bats" – OOOO
- "One hit in four at bats" – HOOO, OHOO, OOHO, OOOH

# Probability and Hitting Streaks

Basic concepts of probability needed to analyze hitting streaks:

**Sample Space**

All possible outcomes in an experiment. Denoted by $S$

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 $S$: 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

# Probability and Hitting Streaks

Basic concepts of probability needed to analyze hitting streaks:

**Sample Space**

All possible outcomes in an experiment. Denoted by $S$

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 $S$: 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" – HHHO, HHOH, HOHH, OHHH

# Probability and Hitting Streaks

Basic concepts of probability needed to analyze hitting streaks:

**Sample Space**

All possible outcomes in an experiment. Denoted by $S$

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 $S$: 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" – HHHO, HHOH, HOHH, OHHH
- "Four hits in four at bats" – HHHH

# Probability and Hitting Streaks

Basic concepts of probability needed to analyze hitting streaks:

**Complement of an Event $A$**

The set of all outcomes that are in the sample space $S$ but are not in the event $A$. Typically denoted as $\overline{A}$ or $A^c$.

# Probability and Hitting Streaks

Basic concepts of probability needed to analyze hitting streaks:

**Complement of an Event $A$**

The set of all outcomes that are in the sample space $S$ but are not in the event $A$. Typically denoted as $\overline{A}$ or $A^c$.

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

# Probability and Hitting Streaks

Basic concepts of probability needed to analyze hitting streaks:

**Complement of an Event $A$**

The set of all outcomes that are in the sample space $S$ but are not in the event $A$. Typically denoted as $\overline{A}$ or $A^c$.

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

# Probability and Hitting Streaks

Basic concepts of probability needed to analyze hitting streaks:

**Complement of an Event $A$**

The set of all outcomes that are in the sample space $S$ but are not in the event $A$. Typically denoted as $\overline{A}$ or $A^c$.

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

# Probability and Hitting Streaks

Basic concepts of probability needed to analyze hitting streaks:

**Complement of an Event $A$**

The set of all outcomes that are in the sample space $S$ but are not in the event $A$. Typically denoted as $\overline{A}$ or $A^c$.

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" – HHHO, HHOH, HOHH, OHHH

# Probability and Hitting Streaks

Basic concepts of probability needed to analyze hitting streaks:

**Complement of an Event $A$**

The set of all outcomes that are in the sample space $S$ but are not in the event $A$. Typically denoted as $\overline{A}$ or $A^c$.

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" – HHHO, HHOH, HOHH, OHHH
- "Four hits in four at bats" – HHHH

# Probability and Hitting Streaks

Basic concepts of probability needed to analyze hitting streaks:

**Complement of an Event $A$**

The set of all outcomes that are in the sample space $S$ but are not in the event $A$. Typically denoted as $\overline{A}$ or $A^c$.

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" – HHHO, HHOH, HOHH, OHHH
- "Four hits in four at bats" – HHHH

# Probability and Hitting Streaks

Basic concepts of probability needed to analyze hitting streaks:

**Complement of an Event $A$**

The set of all outcomes that are in the sample space $S$ but are not in the event $A$. Typically denoted as $\overline{A}$ or $A^c$.

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" – HHHO, 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

Basic concepts of probability needed to analyze 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
$P(A \text{ and } B) = P(A) \times P(B)$.

# Probability and Hitting Streaks

Basic concepts of probability needed to analyze 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?

Basic concepts of probability needed to analyze hitting streaks:

**Rules of Probability**

Complement Rule:

$$P\left(A^c\right) = 1 - P(A)$$

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?

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

# Probability and Hitting Streaks

Basic concepts of probability needed to analyze 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?

$$\begin{aligned} P(\text{at least one H in 4 at bats}) &= 1 - P(\text{zero H's in 4 at bats}) \\ &= 1 - P(\text{OOOO}) \end{aligned}$$

Basic concepts of probability needed to analyze 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?

$$\begin{aligned} P(\text{at least one H in 4 at bats}) &= 1 - P(\text{zero H's in 4 at bats}) \\ &= 1 - P(\text{OOOO}) \\ &= 1 - (0.700 \times 0.700 \times 0.700 \times 0.700) \end{aligned}$$

# Probability and Hitting Streaks

Basic concepts of probability needed to analyze 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?

$$\begin{aligned} P(\text{at least one H in 4 at bats}) &= 1 - P(\text{zero H's in 4 at bats}) \\ &= 1 - P(\text{OOOO}) \\ &= 1 - (0.700 \times 0.700 \times 0.700 \times 0.700) \\ &= 1 - 0.700^4 \end{aligned}$$

# Probability and Hitting Streaks

Basic concepts of probability needed to analyze hitting streaks:

## Rules of Probability

Complement Rule:

$$P\left(A^c\right) = 1 - P(A)$$

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?

$$\begin{aligned}
P(\text{at least one H in 4 at bats}) &= 1 - P(\text{zero H's in 4 at bats}) \\
&= 1 - P(\text{OOOO}) \\
&= 1 - (0.700 \times 0.700 \times 0.700 \times 0.700) \\
&= 1 - 0.700^4 \\
&= \mathbf{0.7599}
\end{aligned}$$

# 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?

# 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.

# 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.

$0.7599 \times 0.7599 = 0.7599^2 = \mathbf{0.5774}$

# 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.

$0.7599 \times 0.7599 = 0.7599^2 = \mathbf{0.5774}$

Five consecutive games?

# 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.

$0.7599 \times 0.7599 = 0.7599^2 = \mathbf{0.5774}$

Five consecutive games? $0.7599^5 = \mathbf{0.2534}$

# 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.

$0.7599 \times 0.7599 = 0.7599^2 = \mathbf{0.5774}$

Five consecutive games? $0.7599^5 = \mathbf{0.2534}$

Ten consecutive games?

# 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.

$0.7599 \times 0.7599 = 0.7599^2 = \mathbf{0.5774}$

Five consecutive games? $0.7599^5 = \mathbf{0.2534}$

Ten consecutive games? $0.7599^{10} = \mathbf{0.0642}$

# 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.

$0.7599 \times 0.7599 = 0.7599^2 = \mathbf{0.5774}$

Five consecutive games? $0.7599^5 = \mathbf{0.2534}$

Ten consecutive games? $0.7599^{10} = \mathbf{0.0642}$

Twenty consecutive games?

# 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.

$0.7599 \times 0.7599 = 0.7599^2 = \mathbf{0.5774}$

Five consecutive games? $0.7599^5 = \mathbf{0.2534}$

Ten consecutive games? $0.7599^{10} = \mathbf{0.0642}$

Twenty consecutive games? $0.7599^{10} = \mathbf{0.004122}$

# 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.

$0.7599 \times 0.7599 = 0.7599^2 = \mathbf{0.5774}$

Five consecutive games? $0.7599^5 = \mathbf{0.2534}$

Ten consecutive games? $0.7599^{10} = \mathbf{0.0642}$

Twenty consecutive games? $0.7599^{10} = \mathbf{0.004122}$

Fifty-six consecutive games?

# 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.

$0.7599 \times 0.7599 = 0.7599^2 = \mathbf{0.5774}$

Five consecutive games? $0.7599^5 = \mathbf{0.2534}$

Ten consecutive games? $0.7599^{10} = \mathbf{0.0642}$

Twenty consecutive games? $0.7599^{10} = \mathbf{0.004122}$

Fifty-six consecutive games? $0.7599^{56} = \mathbf{0.0000002101}$

# 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?

# 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?

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

# 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?

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

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

# 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?

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

Probability of at least one H in 4 at bats in one game?
$1 - 0.643^4 = \mathbf{0.8290599}$

Fifty-six consecutive games?

# 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?

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

Probability of at least one H in 4 at bats in one game?
$1 - 0.643^4 = \mathbf{0.8290599}$

Fifty-six consecutive games? $0.8290599^{56} = \mathbf{0.00002759}$

# Potential Problems

One big assumption that we made in the previous calculation?

# 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!**

# 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 | $(1 - 0.643^1) \times (1 - 0.643^7) = \mathbf{0.3408}$ |
| 2 | 6 | $(1 - 0.643^2) \times (1 - 0.643^6) = \mathbf{0.5451}$ |
| 3 | 5 | $(1 - 0.643^3) \times (1 - 0.643^5) = \mathbf{0.6535}$ |
| 4 | 4 | $(1 - 0.643^4) \times (1 - 0.643^4) = \mathbf{0.6873}$ |
| 5 | 3 | $(1 - 0.643^5) \times (1 - 0.643^3) = \mathbf{0.6535}$ |
| 6 | 2 | $(1 - 0.643^6) \times (1 - 0.643^2) = \mathbf{0.5451}$ |
| 1 | 7 | $(1 - 0.643^7) \times (1 - 0.643^1) = \mathbf{0.3408}$ |

# Potential Problems

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

# 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?

# 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 game.

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:

# 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:

$$\left(1 - 0.643^2\right)^3$$

# 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:

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

# 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:

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

# 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:

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

# 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?

# 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

# 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?

# 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 Probability

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?

# Simulation: Another Way To Estimate Probability

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!

# Simulation: Another Way To Estimate Probability

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?

# Simulation: Another Way To Estimate Probability

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? $S = \{1, 2, 3, 4, 5, 6\}$.

# Simulation: Another Way To Estimate Probability

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? $S = \{1, 2, 3, 4, 5, 6\}$.

The purple numbers are "hits" and the gold numbers are "outs."

# Simulation: Another Way To Estimate Probability

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? $S = \{1, 2, 3, 4, 5, 6\}$.

The purple numbers are "hits" and the gold numbers are "outs."

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

# Simulation: Another Way To Estimate Probability

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? $S = \{1, 2, 3, 4, 5, 6\}$.

The purple numbers are "hits" and the gold numbers are "outs."

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

Problem?

# Simulation: Another Way To Estimate Probability

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? $S = \{1, 2, 3, 4, 5, 6\}$.

The purple numbers are "hits" and the gold numbers are "outs."

Each person should roll the die four times. That would be a "game." If a 1 or 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 Probability

Assuming consecutive at 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 Probability Distribution** with:

# Simulation: Another Way To Estimate Probability

Assuming consecutive at 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 Probability Distribution** with:

- $n$ – number of trials – the number of at bats in a given game

# Simulation: Another Way To Estimate Probability

Assuming consecutive at 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 Probability 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

# Simulation: Another Way To Estimate Probability

Assuming consecutive at 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 Probability 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.

# Simulation: Another Way To Estimate Probability

Assuming consecutive at 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 Probability 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?

# Simulation: Another Way To Estimate Probability

Assuming consecutive at 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 Probability 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

First read in the hits and at bats for DiMaggio's 1941 season into `R`.

# Simulation of Joe DiMaggio's 1941 Season

First read in the hits and at bats for DiMaggio's 1941 season into `R`.Then run the `streaks` program. What's the program do?

# Simulation of Joe DiMaggio's 1941 Season

First read in the hits and at bats for DiMaggio's 1941 season into `R`.Then run the `streaks` program. What's the program do?

- Runs $k$ simulated "baseball seasons." Here the season is DiMaggio's 1941 season. We will set the number of simulates seasons to be 10,000.

# Simulation of Joe DiMaggio's 1941 Season

First read in the hits and at bats for DiMaggio's 1941 season into `R`.Then run the `streaks` program. What's the program do?

- Runs $k$ simulated "baseball seasons." Here the season is DiMaggio's 1941 season. We will set the number of simulates seasons to be 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.

# Simulation of Joe DiMaggio's 1941 Season

First read in the hits and at bats for DiMaggio's 1941 season into `R`.Then run the `streaks` program. What's the program do?

- Runs $k$ simulated "baseball seasons." Here the season is DiMaggio's 1941 season. We will set the number of simulates seasons to be 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 randomly generate the number of hits in each game by using a Binomial Distribution, where the probability of success is $p = 0.357$.

# Simulation of Joe DiMaggio's 1941 Season

First read in the hits and at bats for DiMaggio's 1941 season into `R`.Then run the `streaks` program. What's the program do?

- Runs $k$ simulated "baseball seasons." Here the season is DiMaggio's 1941 season. We will set the number of simulates seasons to be 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 randomly 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 10,000 of these longest streaks!

# Simulation of Joe DiMaggio's 1941 Season

First read in the hits and at bats for DiMaggio's 1941 season into `R`.Then run the `streaks` program. What's the program do?

- Runs $k$ simulated "baseball seasons." Here the season is DiMaggio's 1941 season. We will set the number of simulates seasons to be 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 randomly 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 10,000 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 – here 10,000.

# Simulation of Joe DiMaggio's 1941 Season

10,000 Simulations of DiMaggio's 1941 Season:

| *Method* | *Max* | *40+* | *50+* | *56+* |
|---|---|---|---|---|
| Constant AB's | 75 | 57 | 8 | 2 |
| Varying AB's | 57 | 41 | 2 | 1 |

Estimated probability of DiMaggio hitting safely in at least 56 straight games:

# Simulation of Joe DiMaggio's 1941 Season

10,000 Simulations of DiMaggio's 1941 Season:

| *Method* | *Max* | *40+* | *50+* | *56+* |
|---|---|---|---|---|
| Constant AB's | 75 | 57 | 8 | 2 |
| Varying AB's | 57 | 41 | 2 | 1 |

Estimated probability of DiMaggio hitting safely in at least 56 straight games: $\dfrac{1}{10,000} = 0.0001$

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 1930-1952 (both leagues), 1953 (American League), and 2008-2017 (both leagues). These seasons are not included in the analysis about to be discussed.

# Simulation & Retrosheet Data

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

- $\mathbf{AB}_{ij} = (AB_{ij1}, AB_{ij2}, \ldots, AB_{ijk})$
- 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):
$$H_{ijk} \sim \text{Binomial}\,(AB_{ijk}, p_{ij})$$
- 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 $j$:
$$\mathbf{AB}^1_{ij}, \mathbf{AB}^2_{ij}, \ldots, \mathbf{AB}^m_{ij}$$
- The number of hits a player gets in each game in the $m^{\text{th}}$ simulated season is
$$\mathbf{H}^{*m}_{ij} \sim \text{Binomial}\left(\mathbf{AB}^{*m}_{ij}, p_{ij}\right)$$
- Any run of hits in $\mathbf{H}^{*m}_{ij}$ 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.

# Results: 1000 Simulated Baseball Histories

Top 10 Maximum Hitting Streaks:

| | | Simulated Season | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Player* | *Year* | *40+* | *50+* | *56+* | *Min* | *Q1* | *Q2* | *Q3* | *Max* |
| Harry Heilmann | 1921 | 34 | 9 | 5 | 10 | 18 | 22 | 27 | 91 |
| Rogers Hornsby | 1922 | 54 | 10 | 3 | 11 | 19 | 23 | 29 | 89 |
| Felipe Alou | 1966 | 4 | 3 | 2 | 9 | 14 | 16 | 21 | 75 |
| Julio Franco | 1991 | 3 | 1 | 1 | 8 | 14 | 16 | 20 | 74 |
| Alex Rodriguez | 1996 | 10 | 3 | 2 | 9 | 15 | 18 | 22 | 72 |
| Rogers Hornsby | 1921 | 21 | 3 | 1 | 6 | 14 | 17 | 22 | 71 |
| Ichiro Suzuki | 2004 | 34 | 8 | 5 | 11 | 18 | 22 | 27 | 69 |
| Jimmy Rollins | 2007 | 2 | 1 | 1 | 7 | 13 | 15 | 18 | 64 |
| Rogers Hornsby | 1922 | 23 | 4 | 2 | 7 | 15 | 18 | 23 | 63 |
| Ralph Garr | 1974 | 13 | 3 | 2 | 9 | 15 | 18 | 22 | 63 |

Hitting Streaks in 18,607,000 Simulated Player-Seasons:

| *Max* | *40+* | *50+* | *56+* |
|---|---|---|---|
| 91 | 2237 | 284 | 85 |

# Results 1000 Simulated Baseball Histories

Hitting Streaks in 1000 Simulated Baseball Histories:

| *Max* | *40+* | *50+* | *56+* |
|---|---|---|---|
| 91 | 252 | 165 | 70 |

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

$$\frac{85}{18,607,000} = 0.000004568 = 0.0004568\%$$

The estimated probability of a hitting streak of at least 56 games occurring at **some point** in baseball history (technically 64 seasons of AL-NL and 4 seasons of just NL):

$$\frac{70}{1000} = 0.07 = 7\%$$

# Simulation of Hitting Streaks

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

# 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 the season

Solution?

# 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 the season

Solution?

- Vary the batting average over the course of a simulated season!

# 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 the season

Solution?

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

# 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 the 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?

# 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 the 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?