Hitting Streaks in Seasons Using Non-Constant Batting Averages

Philip Yates [1] David Rockoff [2]

California State Polytechnic University, Pomona [1] Iowa State University [2]

Background

In an article from the *New York Times* (March 30, 2008) called “A Journey to Baseball’s Alternative Universe,” Samuel Arbesman and Steven Strogatz ran simulations of baseball seasons to estimate the probability of long hitting streaks. They treated a player’s at bats per game as constant across all games in a season, which greatly overestimates the probability of long streaks, simulated 10,000 baseball histories and tabulated which player held the longest streak, who that player was, whe he did it, and how long his streak was. Rockoff and Yates (2009) ran simulations that vary at bats, using Retrosheet game data for all of major league baseball from 1920-1929 and 1954-2008, as well as for the National League in 1911 and 1953. We ran three additional simulations, each with a diferent treatment of a player’s batting average: Beta random variable, correlated based on performance over a 15 game “neighborhood”, and correlated based on performance over a 30 game “neighborhood.”

Constant vs. Variable At-Bats

- To illustrate how constant at-bats can overestimate the likelihood of long hitting streaks, Warrack (1995) makes a Jensen’s inequality argument by approximating the probability of getting at least one hit in B at-bats as

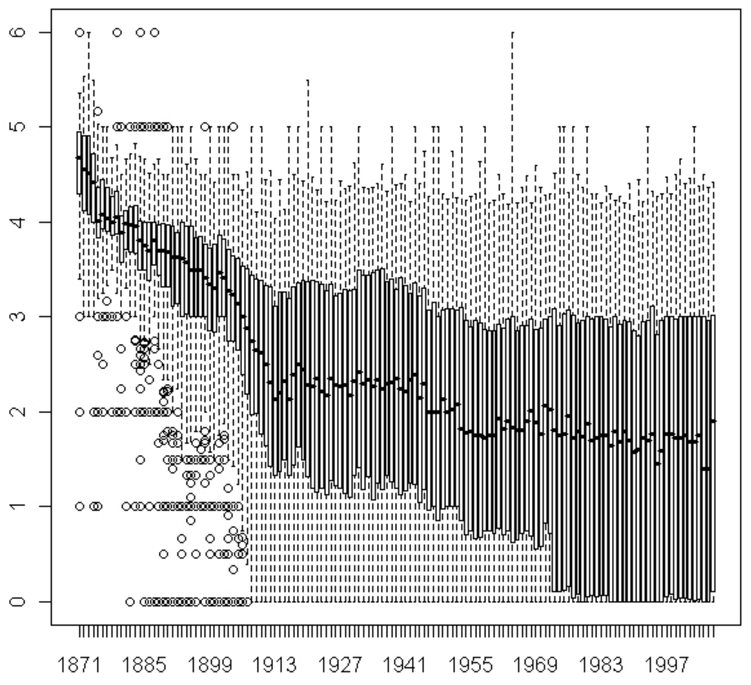
$$p = 1 - (1 - A)^B,$$

where A is the player’s batting average. This is concave in B.

- Rockoff and Yates (2009) use this argument to illustrate how low at-bat games hurt chances for getting a hit in a game more than high at-bat games help. Extended over long stretches, this effect is telescoped.

B_1	B_2	p	B_1	B_2	p
4	4	**0.577**	2	6	**0.450**
3	5	**0.547**	1	7	**0.275**

- **At-bats per game have spread out and decreased over time:**



Simulations and Results

- We obtained game data from Retrosheet. For each season in their database, this data includes multitudes of information on every single plate appearance in the major leagues, including unique game identifier, batter, and whether the appearance resulted in an at-bat. Thus we were able to determine the number of at-bats for each player in each game in every season.
- The number of at-bat for a given player, which is bootstrapped, from the player’s actual distribution of game at-bats during that season
- In our simulations, we treated a player’s chance of a hit in a given at-bat in three different ways:

Method #1 – Beta: A player’s chance of a hit in a given at bat is a Beta random variable with α is the number of hits in a season and β is the number of outs in a season;

$$f(x|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}, \quad \alpha > 0, \quad \beta > 0, \quad x \in [0,1]$$

Method #2 – Binom-15: Let the probability of a hit vary game to game using a neighborhood of 15 games. For example, in game 50, the probability should be reflected in his performance in games 35–65. Using the new hit probabilities, generate a new array of hits. Run a few iterations.

Method #3 – Binom-30: Similar to Method #2, but using 30 game neighborhoods. For example, in game 50, the probability should be reflected in his performance in games 20–80.

- 1000 baseball “histories” were simulated.

Results: The following table lists the top 20 performances in the simulations. Note that the method denoted as “Binom” is the method presented in Rockoff and Yates (2009).

Streak	Player	Year	Method	56+	Games	AB	Avg	HR	RBI	Max
95	George Sisler	1922	Binom-15	23	142	586	.420	8	105	41
95	George Sisler	1920	Beta	8	154	631	.407	19	122	25
93	Tony Gwynn	1997	Binom-15	3	149	592	.372	17	119	20
91	Harry Heilmann	1921	Binom	5	149	.394	19	139	23	
90	George Sisler	1921	Binom-15	1	138	582	.371	12	104	
89	Rogers Hornsby	1922	Binom	3	154	623	.401	42	152	33
88	Freddie Lindstrom	1928	Binom-30	3	153	646	.358	14	107	
88	Dante Bichette	1998	Binom-30	1	161	662	.331	22	122	12
87	Lefty O’Doul	1929	Binom-15	13	154	638	.398	32	122	
87	Harry Heilmann	1921	Binom-15	9	149	602	.394	19	139	23
85	Rogers Hornsby	1921	Binom-15	4	154	602	.394	19	139	23
85	Derek Jeter	2000	Binom-15	1	148	593	.339	15	73	13
84	Paul Waner	1927	Binom-30	3	155	623	.380	9	131	23
83	Ian Kinsler	2008	Binom-15	1	121	518	.319	18	71	25
82	Heinie Manush	1928	Binom-30	8	154	638	.378	13	108	
82	George Sisler	1922	Beta	5	142	586	.420	8	105	41
80	Pie Traynor	1929	Binom-30	3	130	540	.356	4	108	
80	Lloyd Waner	1927	Binom-15	2	150	629	.355	2	27	
79	Al Simmons	1925	Binom-30	12	153	654	.387	24	129	23
78	George Sisler	1920	Binom-15	19	154	631	.407	19	122	25

Acknowledgements

- The information used here was obtained free of charge from and is copyrighted by Retrosheet. Interested parties may contact Retrosheet at 20 Sunset Road, Newark, DE 19711 or at www.retrosheet.org.
- Thanks to Cliff Blau for DiMaggio game-by-game-data.

Comparisons

Streaks in 18,607,000 simulated player-seasons

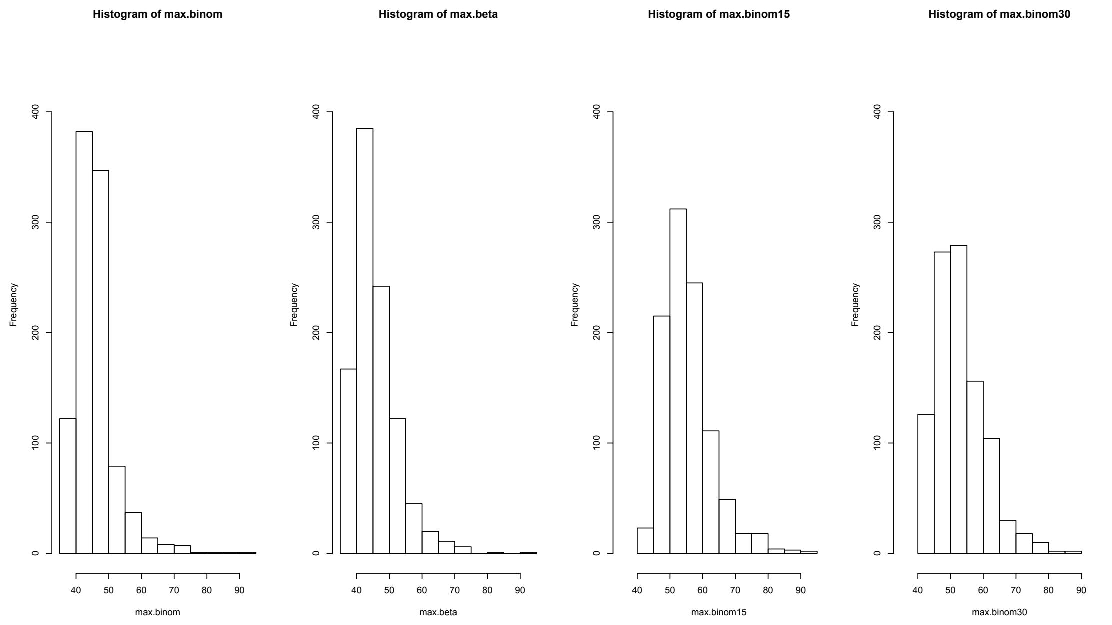
Method	Max	40+	50+	56+
Binom	91	2337	284	85
Beta	95	2248	270	88
Binom-15	95	11,328	1741	561
Binom-30	88	7190	1094	422

In 1000 histories

Method	Max	40+	50+	56+
Binom	91	252	165	70
Beta	95	899	244	84
Binom-15	95	1000	829	450
Binom-30	88	1000	662	343

10,000 simulations of DiMaggio ’41

Method	Max	40+	50+	56+
Constant AB	75	57	8	2
Binom	57	41	2	1
Beta	61	36	4	1
Binom-15	70	134	23	4
Binom-30	67	114	20	9



Limitations of Research

- Excludes majority of data prior to 1953 (1900’s, 1910’s, 1930’s, 1940’s)
- Doesn’t allow for the remote possibility of more than one long streak by a player in a simulated season
- Doesn’t account for day-to-day managerial and player choices, such as batting a guy leadoff if he has a long hitting streak going, or not taking a walk in the late innings
- Doesn’t account for multi-season streaks a la Jimmy Rollins (2005-2006)

References

- Arbesman, Samuel & Strogatz, Steven. “A Journey to Baseball’s Universe.” *The New York Times.* 30 March 2008.
- Rockoff, David M. & Yates, Philip A. (2009) “Chasing DiMaggio: Streaks in Simulated Seasons Using Non-Constant At-Bats.” *Journal of Quantitative Analysis in Sports*: Vol. 5: Iss. 2, Article 4.
- Warrack, Giles. (1995) “The Great Streak.” *Chance*, Vol. 8, No. 3, 41–43, 60.

http://www.csupomona.edu/~payates

payates@csupomona.edu[1] drockoff@iastate.edu[2]