Reassessing the likelihood of Joe DiMaggio's 1941 hitting streak

A new model for estimating expected hitting streaks

For consideration for the 2014 Jack Kavanagh Memorial Youth Baseball Research Award

Cyrus Hettle
Senior, University of Kentucky
cyrus.h@uky.edu

Abstract

In 1941, Joe DiMaggio had a hit in fifty-six consecutive baseball games. Several researchers, including Nobel Prize-winning physicist Edward Purcell, have tried to calculate the probability of this remarkable run of luck, with conflicting results. However, all of these models have failed to account for many sources of variance that can significantly affect the chance of a streak. In fact, variance in the number of plate appearances a player has, in the ability of the pitchers he faces, or in several other factors, can drastically lower the probability of a streak occurring, particularly one as long as DiMaggio's.

I present a modified Monte Carlo model incorporating these additional sources of variance, and use this model and a 700,000-game simulation to give a more nuanced analysis of the likelihood of DiMaggio's accomplishment. Finally, I consider current research in hitting streak distribution, and suggest how the modified model can be used to investigate the phenomenon of games clumping together in hitting streaks beyond what chance would predict.

1 Introduction and historical background

1.1 DiMaggio and the Streak

Many historians of baseball have considered Joe DiMaggio's streak of 56 consecutive games with a hit to be the most remarkable and unbreakable record of all time. The streak has an aura of improbability stemming from DiMaggio's somewhat mythic status and the gap between the length of DiMaggio's run and the 44-game stretch of both of the second-longest streaks, accomplished by Willie Keeler and Pete Rose. Even the number 56 itself has an inextricable association with DiMaggio's feat which is unmatched by any other record, except perhaps the massive lifetime totals of 714 and 755.

Figure 1: DiMaggio kissing his bat, 1941

To give a brief and prosaic account of the streak, DiMaggio played 56 consecutive games with a hit from May 15 to July 16, 1941. He broke Keeler's record on July 2, and began another 16-game streak on July 18, the day after the streak was snapped.[1] DiMaggio also holds the second-longest hitting streak in minor-league history, a 61-game run in the 1933 Pacific Coast League.

Two natural questions arise from studying DiMaggio's streak: how improbable was it for DiMaggio to have this 56-game streak, and how improbable was it for anyone, not just DiMaggio, to rack up 56 consecutive games with a hit? In many ways, these questions are easier to study than the improbability of other records, because a hitting streak is a phenomenon that occurs within a single season, or possibly two. Calculating the probability that a baseball player would hit 714 home runs, as Babe Ruth did, requires a model of the probability that a power-hitting freak like Ruth would even exist in the history of major-league baseball, a question that is most likely unanswerable. However, with just a few pieces of information, it is straightforward to obtain a rough estimate of the chance of a streak of any length for any player using elementary probability theory; see the next section for an explanation.

[1] DiMaggio walked on July 17, though he did not have a hit that day, giving him an 83-game streak of reaching base. Curiously, this very impressive streak is *not* the record; Ted Williams reached base in 84 consecutive games in 1949.

1.2 The basic model

A fundamental piece of information that is needed to estimate the chance of a player having a lengthy hitting streak is an estimate of that player's chance of having a hit in any individual game. Let P_H denote this probability. Then the basic model estimates that in every game,

$$P_H = 1 - \left(1 - \frac{H}{PA}\right)^{PA/G}$$

where H, PA, and G are the number of hits, plate appearances, and games for the player during the entire season in question. If we plug DiMaggio's 1941 statistics into this formula, the basic model says that DiMaggio's probability of getting a hit in any one game should be

$$P_H = 1 - \left(1 - \frac{193}{622}\right)^{622/139} \approx .81.$$

Now, to compute the chance of a streak, let S_L be the probability of having a hit in every one of L games; in other words, of having a streak of length L in a season of exactly L games.[2] Then $S_L = (P_H)^L$. Calculating P_H and S_L in this way is straightforward (and entirely correct), if we assume that:

- the result of each plate appearance is independent;
- the player has the same chance of having a hit in every plate appearance;
- the player has the same number of plate appearances in every game.

Of course these are all incorrect assumptions, but it may appear that over the course of a long hitting streak they will even out and produce an accurate result. Yet as I will show in the following example, this variance is actually extremely significant, especially when we are considering streaks as long as DiMaggio's. As I will explain in Section 2, the modified model corrects each of the three assumptions that go into the basic model. This increased variability leads to significantly more accurate results.

1.3 An example: how the basic model fails when there is variance

To show how much variance affects the probability of a streak of any length, consider the case of two imaginary hitters, Charlie Consistent and Virgil Variable. Each player is of exactly equal ability: he has a .3 chance of getting a hit in any plate appearance. However, Charlie has exactly 4 plate appearances in each game, while Virgil has 3 plate appearances one game, 4 in the next, 5 after that, then goes back to 3, etc.

Common sense would suggest that both players have the same probability of having a three-game hitting streak, but because of the variance in his number of plate appearances, Virgil actually has a slightly lower chance.

[2] If we are interested in the probability of having a 56-game streak in more than 56 games, such as the 139 games DiMaggio played in 1941, the calculations become significantly more complicated. For an excellent summary, see [8].

To see this, recall that each player has a H/PA average of .3; thus each one's chance of having a hit in a game with 4 plate appearances is

$$P_H = 1 - (1 - .3)^4 = 1 - (.7)^4 \approx .76.$$

We calculate the chance of a hit in Virgil's 3 and 5 plate appearance games similarly. Charlie and Virgil's chances of having a hit in each of three games and of having a three-game streak are in Table 1.

Player	G1 PA	G1 P_H	G2 PA	G2 P_H	G3 PA	G3 P_H	Streak chance
Charlie Consistent	4	.76	4	.76	4	.76	.439
Virgil Variable	3	.66	4	.76	5	.83	.416

Table 1: Streak chances and variability: Charlie is slightly better than Virgil.

Charlie is about 5.5% more likely than Virgil to have a three-game hitting streak, which isn't much. But what if we are talking about longer streaks? (See Table 2.)

Player	S_3	S_{10}	S_{25}	S_{56}
Charlie Consistent	.439	.064	.0010	.00000021
Virgil Variable	.416	.054	.0006	.00000008
Charlie's advantage	5.5%	18.5%	66.6%	262.5%

Table 2: Charlie and Virgil's chances of having streaks of length 3, 10, 25, and 56.

Charlie and Virgil are equally skilled: either is just as likely to have a hit in a given plate appearance. But just by varying Virgil's plate appearances from game to game, without even changing the average, he is doomed to have 56-game hitting streaks less than half as often as Charlie. This small initial change in variance leads to drastic differences when we look at streaks as long as DiMaggio's. In fact, the basic model treats Charlie and Virgil like they are exactly the same because it assumes Virgil has the same number of plate appearances in every game. If I ask the basic model to predict Virgil's chance of a 56-game hitting streak, it will give me Charlie's probability, which is very far off from the Virgil's true probability.

The way variance is affecting Virgil in this example is quite subtle. His average P_H is

$$\frac{G1P_H + G2P_H + G3P_H}{3} = \frac{.66 + .76. + 83}{3} = .75,$$

just barely below Charlie's .76. But doesn't this contradict the fact that his H/PA is exactly the same as Charlie's?

Not at all! The key observation is that for the purposes of a hitting streak, having one hit is just as good as having two or three or even more, and therefore the extra PA gained in going from 4 to 5 is less significant than the PA lost from going from 4 to 3. For PA #5 to matter, Virgil would need to go hitless in his first four plate appearances and then get a hit in #5. This will happen with probability $(1 - .3)^4 \cdot .3 = .072$. But for PA #4 to matter, Virgil only needs to go hitless in his first

three plate appearances before hitting safely on his fourth, and this will happen with the greater probability $(1 - .3)^3 \cdot .3 = .102$. Therefore, since the 3 PA games hurt Virgil more than the 5 PA games help his chance of continuing a streak, his average P_H is going to be slightly less than Charlie's.

Let's introduce a new player, Sam Slightly Worse. Sam is consistent like Charlie in that his P_H is exactly the same in every game, but instead of his P_H being .76 like Charlie, it is .75, like Virgil. It should be clear that Sam's chance of having a 56-game streak will fall somewhere in between Charlie's and Virgil's. Before reading on, think for a moment and guess who Sam's chance will be closer to. Is it more significant that his P_H is a little bit less than Charlie's, or that he is far more consistent than Virgil?

OK, time to reveal the answer. Charlie's chance was about 1 in 4.6 million, and Virgil was trailing behind at 1 in 12.5 million. It turns out that Sam is right at 1 in 10 million, much closer to Virgil. What this means is that the reason Virgil fared so much worse than Charlie is (mostly) not because his P_H varied wildly from game to game, but because the variance caused his overall average P_H to be less. Indeed, 77.6% of his lower chance of a streak comes from the lower average, while only 22.4% results from game-to-game difference in P_H.

1.4 Previous models and estimates

Nobel Prize-winning physicist Henry Purcell was one of the first to estimate the improbability of the Streak. As the great biologist and author Stephen Jay Gould reports in [9], Purcell calculated that to have a probability greater than 50 percent that a run of even fifty games will occur once in the history of baseball, the history of baseball would have to include either four lifetime .400 batters or fifty-two lifetime .350 batters over careers of one thousand games.

Gould, a lifelong DiMaggio fan, summarized Purcell's results thus:

> Nothing ever happened in baseball above and beyond the frequency predicted by coin-tossing models. There is one major exception, and absolutely only one...sequence so many standard deviations above the expected distribution that it should not have occurred at all. DiMaggio's streak is the most extraordinary thing that ever happened in American sports [9].

The basic model appears to have been introduced by Charles Blahous in [4], although Blahous uses at-bats in the place of plate appearances, which gives a slightly less accurate result. Blahous also recognized the variability of plate appearances from game to game and the difficulty of assigning games a PA/G ratio which may not be an integer, and corrected for this in his model. However, he assumed that the variance of pitching ability from game to game would not impact the final result, a faulty assumption which is corrected in my model. Blahous concluded that the chances of DiMaggio having a 56-game streak in 1941 were about 1 in 746.

Unfortunately, Blahous made a small technical error in his probability which ends up significantly overstating the chance of DiMaggio having a streak. Michael Frieman pointed out and corrected this error in [8]. Frieman also uses the basic model to arrive at P_H, the probability of having a hit in any one individual game, but because he correctly calculates the probability of stringing 56 consecutive games with a hit together, his final figure is a chance of 1 in 9,545.

Next to tackle the problem were Joe D'Aniello, in [6], and coauthors Bob Brown and Peter Goodrich, whose paper [5] was published in the same issue of the Baseball Research Journal as

D'Aniello's. Both papers focus on a pure Monte Carlo model of the problem, and both use a sample size of one million seasons. D'Aniello chose to vary DiMaggio's plate appearances according to his actual 1941 opportunities, including just two plate appearances in one rain-shortened game, while Brown and Goodrich used the basic model without any variance. As would be expected, the two papers reached rather different conclusions. D'Aniello computed the probability of a streak to be about 1 in 66,667, while Brown and Goodrich place it significantly higher, at 1 in 4,504.[3]

The problem received a revival at the hands of an New York Times op-ed ([3]) written by Samuel Arbesman and Steve Strogatz. Arbesman and Strogatz used the basic model, but examined the chance of a 56-game streak over all of baseball history, not just DiMaggio in 1941. They concluded that while it is not surprising that baseball's record books should note a streak of 56 games or more, it is remarkable that DiMaggio should hold the record, and not one of the many earlier hitters whom they estimate should have had been much more likely. In a later and longer paper ([2]), they add variance to their calculations. Instead of altering plate appearances or the chance of a hit in a particular plate appearance, they vary P_H itself from game to game. Unfortunately, even when the model includes huge amounts of variance in P_H, the number of streaks of shorter lengths (30-49 games) predicted by their model exceeds the statistically significant number in real life. As we saw when comparing Sam to Virgil, varying P_H itself while keeping the same overall average does not sufficiently alter the overall chance of streak.

2 A new model

2.1 Factors considered

To model both plate appearance and hitting probability more accurately, the model incorporates four sources of randomness that affect every baseball game: park effects, pitcher ability, platoon effects, and the performance of other batters, which determines the number of plate appearances any individual batter has. The following four subsections explains these factors' importance and how they were accounted for in the model. While they may be improved further (see Section 4.1), I believe that this new model adjusts for the most prominent factors, and any additional adjustments would provide minimal additional accuracy and may require a significant amount of either research effort or additional run time to achieve the same sample size of simulation.

2.1.1 Park effects

Park factor statistics, which express the effect of a park in a given season on run scoring, are normalized to 100. Hence the ratio of the park factor to 100 is ideally the ratio of the number of runs that would be scored in that park by an average team to the number of runs scored by that team in a neutral park. For studying hitting streaks, however, we are more concerned with the hit factor, which is the equivalent statistic for the chance of having a hit in a given plate appearance. Park factors are easily available for all parks in the modern era, for instance from [1]. Hit factors

[3] Brown and Goodrich and Frieman both used the basic model and the same technique of calculating the chance of a streak, as opposed to Blahous's fallacious method. The difference in their final results has two sources. First, Brown and Goodrich used DiMaggio's H/PA average not only from 1941, but from the period from 1936-1940, when his average was slightly higher. Also, Brown and Goodrich's result is derived from a random simulation of 1,000,000 seasons, and so is prone to some level of error. Frieman's calculation, on the other hand, is calculated purely from P_H.

are more difficult to compute and find, but by studying available hit factors from the last ten years I estimate that in general

$$HF \approx \sqrt{PF/100}$$

where HF is the hit factor normalized to 1.[4] See Table 3 for the results.

This result makes sense intuitively. Because most runs occur from the interaction of multiple hits, small increases in hit factor will result in slightly larger increases in the chance of multiple hits occurring consecutively.

	BOS	CHI	CLE	DET	NYY	PHI	STL	WAS
PF	102	99	95	110	99	98	103	95
HF	1.010	.995	.975	1.045	.995	.990	1.015	.975

Table 3: Park factors and hit factors for the 1941 American League.

2.1.2 Platoon effects

It was well-known even before the rigorous study of baseball statistics that batters usually hit better against opposite-handed pitchers. An estimate in [7] gives that right-handed batters such as DiMaggio on average improve their batting average by .017. Research in [11] and others find that most batters have a similar platoon split, so given the lack of lefty-righty splits for 1941 and much of the rest of DiMaggio's career, I used the .017 estimate to conclude that

$$\left(\frac{H}{PA}\right)_{vs.LHP} - \left(\frac{H}{PA}\right)_{vs.RHP} \approx .014786.$$

Because DiMaggio faced more right-handed than left-handed pitchers, the adjusted averages I used were respectively

$$\left(\frac{H}{PA}\right)_{vs.LHP} \approx .32027$$
$$\left(\frac{H}{PA}\right)_{vs.RHP} \approx .30548.$$

2.1.3 Pitcher ability

By far the most important factor in determining DiMaggio's chance of a hit in any given plate appearance is the quality of the pitcher throwing to him. Of course, for this purpose we care only about the pitcher's tendency to give up hits, and not some measure of runs allowed, so we begin with the pitcher's H/BF ratio. I adjusted this ratio for two factors: the quality of the pitcher's opposition, and the hit factor of his home ballpark. For instance, Washington pitchers benefited in 1941 from playing in the least run-friendly ballpark in the league, as well as not having to play against their own offense, which ranked second of the eight teams in hits.

Conveniently, this adjusted ratio also picks up the defensive ability of the opposing team.

[4] When the park factor is near 100, a reasonable approximation is $HF \approx \sqrt{PF/100} \approx \frac{PF/100}{2} + \frac{1}{2}$, because the behavior of $x^{\frac{1}{2}}$ in a neighborhood of 1 is approximately linear with slope $(x^{\frac{1}{2}})'(1) = \frac{1}{2}$.

2.1.4 Plate appearance variance

The previous three adjustments have introduced randomness into the probability of having a hit in any one plate appearance, instead of a constant probability of H/PA. This final correction adds further variance by changing the number of plate appearances from game to game. Instead of having the plate appearances for each game directly reflect DiMaggio's 1941 statistics, I generated 30 random plate appearances against the two pitchers in proportion to those pitcher's innings pitched. Each plate appearance was randomly assigned to be an out or a hit, adjusted for the pitcher's ability and the park's hit factor. From these 30 plate appearances, I then computed how many plate appearances the Yankees would have before they would end the game by reaching 27 outs (or 24 in a proportion of the home games equal to New York's winning percentage that year). Since DiMaggio batted fourth in every game he played in 1941, it is easy to compute using the ceiling function that

$$\text{DiMaggio's PA} = \left\lceil \frac{\text{New York PA} - 3}{9} \right\rceil.$$

This entire procedure results in DiMaggio having about .024 fewer plate appearances per game that he should, but this is a small enough error that it is easily corrected by granting an extra plate appearance to .024 · 700,000=16,800 of the games chosen at random.

2.2 Implementation

I generated 700,000 games of data in a semi-Monte Carlo manner. First, I took 100,000 games against each of seven other teams in the 1941 American League. These 100,000 were divided into 50,000 games home and away for the Yankees. These 700,000 games were permuted and divided into 139-game seasons. Next, each game was randomly assigned two pitchers from the opposing team's roster. Each pitcher played a number of times in the simulation proportional to his innings pitched in the season. The starting pitcher was randomly assigned an inning total ranging uniformly from 5 to 9, and the remaining innings were assigned to the reliever.

Having set up the game, it now remains to compute DiMaggio's chance of a hit given the park and opposing pitchers. To compute his number of plate appearances, I carried out the procedure in Section 2.1.4, and broke this total number into an integer number of plate appearances against each of the two pitchers. Finally, I compute the probability of a hit in that game to be

$$P_H = 1 - (1 - ((.31029 \cdot A_S + P_S)HF)^{PA_S}) \cdot (1 - ((.31029 \cdot A_R + P_R)HF)^{PA_R})$$

where .31029 is DiMaggio's $\frac{H}{PA}$, A is pitcher ability, P the appropriate platoon split, HF the park's hit factor, PA plate appearances, and the $_S$ and $_R$ subscripts indicate the starting and relieving pitcher respectively.

The result is a semi-Monte Carlo model, in contrast to a "pure" Monte Carlo method, which would generate a single result for the game: either a hit or no hit. In this model, although a specific opponent, number of plate appearances, opposing starter and reliever, etc. are chosen for each game, the result is a probability between 0 and 1 that DiMaggio would have a hit in that game. Counting events as rare as a 56-game hitting streak in a relatively small sample demands the additional precision from examining these probabilities when a pure Monte Carlo model may only produce one or two streaks in the entire sample.

3 Results

3.1 Game-to-game results

In this new model, the average probability of having a hit in any one individual game was approximately .8003, slightly lower than the basic model's constant of .81. This may at first seem to be a fault with the model, as studies such as [10] have assumed that the basic model gives an accurate estimate of a player's average chance of having a hit, and indeed it is good enough for many purposes. But the example of Charlie and Virgil provides an explanation for why P_H should be slightly behind the basic model's prediction, and actual data on P_H confirms the the basic model's exaggeration. Table 4 shows the top fifty players by batting average in 2013. E[GH] is the expected number of games with a hit, using the basic model, and GH is the actual number of games with a hit that player had in 2013. The final column shows the difference. Note that the basic model

Player	E[GH]	GH	±	Player	E[GH]	GH	±
Jose Altuve	112.66	114	1.34	Adam Jones	118.73	121	2.27
Nori Aoki	111.58	105	-6.58	Howie Kendrick	90.77	89	-1.77
Brandon Belt	101.69	95	-6.69	Jason Kipnis	105.46	104	-1.46
Carlos Beltran	106.31	101	-5.31	Adam Lind	94.60	90	-4.60
Adrian Beltre	123.54	120	-3.54	James Loney	111.04	106	-5.04
Michael Brantley	106.00	99	-7.00	Jed Lowrie	112.85	106	-6.85
Billy Butler	112.94	105	-7.94	Victor Martinez	117.21	114	-3.21
Marlon Byrd	103.91	104	.09	Joe Mauer	87.75	84	-3.75
Miguel Cabrera	116.47	118	1.53	Andrew McCutchen	117.40	111	-6.40
Robinson Cano	120.24	115	-5.24	Yadier Molina	102.64	93	-9.64
Matt Carpenter	121.43	115	-6.43	Daniel Murphy	119.71	119	-.71
Shin-Soo Choo	107.32	99	-8.32	Daniel Nava	93.67	87	-6.67
Allen Craig	101.11	99	-2.11	David Ortiz	101.78	98	-3.78
Michael Cuddyer	100.45	105	4.55	Dustin Pedroia	120.66	119	-1.66
Chris Davis	111.79	108	-3.79	Salvador Perez	97.63	95	-2.63
Josh Donaldson	113.88	105	-8.88	Buster Posey	103.20	97	-6.20
Jacoby Ellsbury	104.00	105	1.00	Alexei Ramirez	116.38	114	-2.38
Freddie Freeman	110.91	107	-3.91	Marco Scutaro	93.30	88	-5.30
Paul Goldschmidt	117.01	112	-5.01	Jean Segura	109.57	111	1.43
Carlos Gomez	102.53	93	-9.53	Mike Trout	118.53	120	1.47
Adrian Gonzalez	112.77	110	-2.77	Troy Tulowitzki	91.22	86	-5.22
Matt Holliday	101.82	101	-.82	Chase Utley	91.11	91	-.11
Eric Hosmer	119.16	116	-3.16	Shane Victorino	89.79	84	-5.79
Torii Hunter	111.91	107	-4.91	Joey Votto	115.70	118	2.30
Chris Johnson	106.38	100	-6.38	Jayson Werth	95.01	89	-6.01
				Total	**5373.52**	**5192**	**-181.52**

Table 4: The fifty leaders in batting average in 2013. Only nine of the fifty outperformed the basic model's estimate.

overestimates GH by about 3.6 games per hitter, a difference comparable to the difference between the basic model and my revised model.[5]

Figure 2, the histogram of P_H over the 700,000 games in the simulation, is approximately normal but with two prominent humps comprising the games with 4 and 5 PA. Note that the great majority of games fall within a small neighborhood of the average, and only about 1.7% of the games had a P_H of .65 or less.

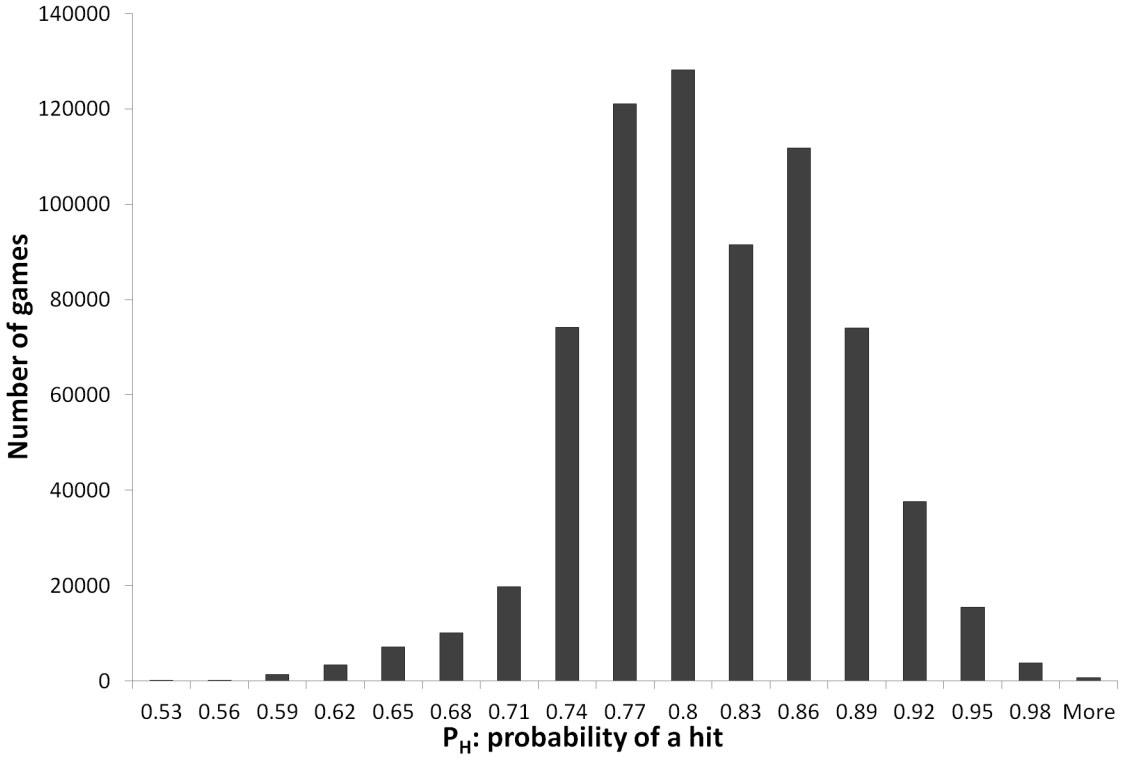

Figure 2: Histogram of the hit chances in individual games

3.2 Season-to-season results

In the 5,035 seasons computed in the new model, DiMaggio's average chance of having a 56-game hitting streak was approximately .00006716, or 1 in 14,890, with a histogram (Figure 3) which is skewed right. This chance is lower than that in any other cited study except [6], which used some more extreme variance in game-to-game PA. The maximum observed single-season chance was .000283, or 1 in 3,533, and the minimum was .0000149, or 1 in 67,114.

[5] A curious trend in the 2013 data is a weak negative correlation between H/PA and the degree of exaggeration by the basic model–essentially, the basic model appears to be slightly more accurate for better hitters.

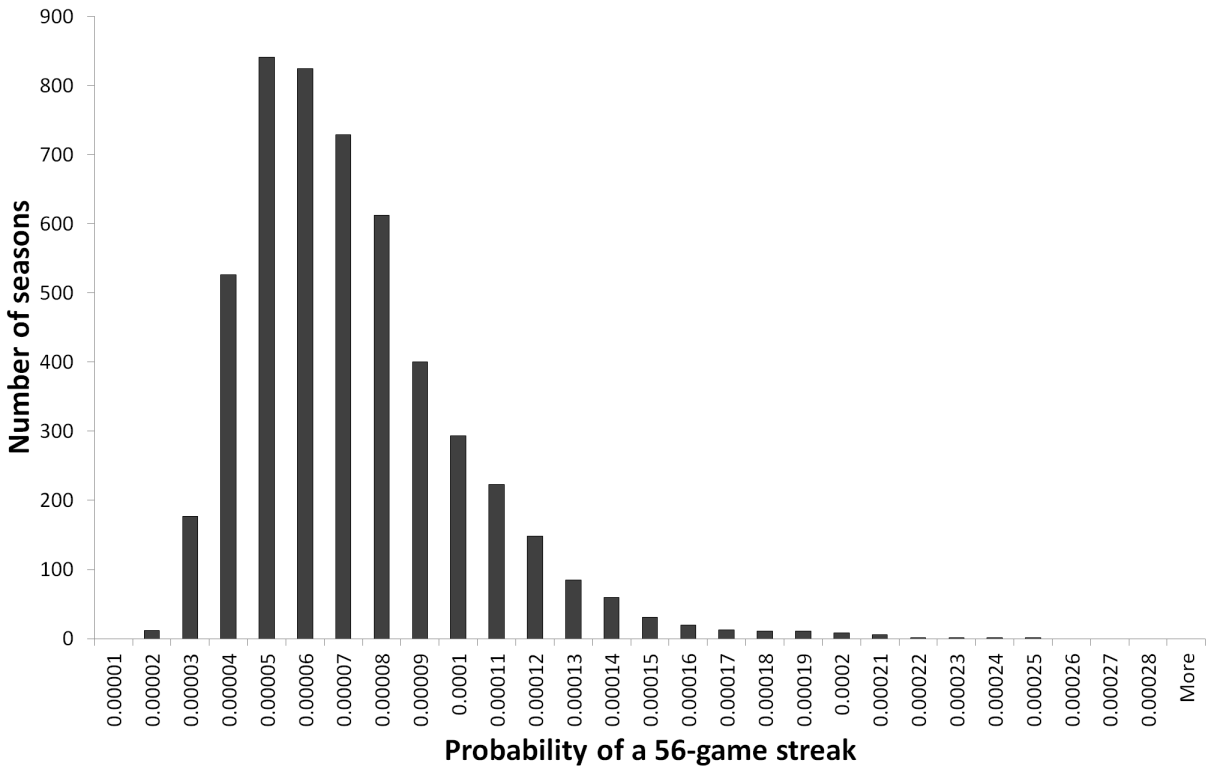

Figure 3: Histogram of the streak chances in individual seasons

4 Further research

4.1 Improving the model

While I believe that the model introduced in this paper is a significant improvement in accuracy over previous variations of the basic model, there are still many additional sources of both in-game and game-to-game variance that could be added. Here are a few ways the model could be modified, with suggestions for implementation for both DiMaggio and for any player in any season, as well as considerations for how adding each factor to the model might change the results.

The model currently uses a single hit factor, which for older seasons is based on a park factor (for more recent seasons, a separate hit factor is available.) Some older parks may be conducive to run-scoring by allowing more home runs, but not significantly improve the chance of having a hit. Asymmetric ballparks may also affect batters of different handedness to varying degrees, and even symmetric ballparks may have different effects on slap hitters than power hitters. Calculating all of these factors is a large project itself, but there may be a few aspects that could be reasonably incorporated into this model.

Many baseball games do not follow the model's primitive pattern of having exactly one relief pitcher, who may pitch up to four innings. While the current model is a good approximation of a batter facing multiple pitchers who may have different handedness and ability, it does not respect the tendency of pitchers to pitch in different roles (starter, middleman, closer, etc.) with highly specific numbers of innings. There is certainly a great deal of room for improvement in

the accuracy of this portion of the model, but I don't believe such improvement would change the results of the model in any appreciable way.

It is well-known that baseball has a home-field advantage, although this advantage is much smaller than in other sports, such as basketball or football. The model currently does involve designating games as home or away for purposes of determining park effects and the number of outs the Yankees are able to make before the game ends, but makes no adjustment for Yankee Stadium as opposed to, say, the White Sox' Comiskey Park, which has an identical hit factor. Such an adjustment would likely have a small but measurable negative effect on the chances of a streak, as it would add additional variance, thereby adding spread and lowering the average of P_H.

The sample size for the modified model was 700,000 randomly permuted games, 50,000 home and away against each other team in the American League. This results in many seasons with unbalanced schedules, where the Yankees play either many more strong teams or many more weak teams than they would with a balanced schedule. Improving the accuracy of the schedule by randomizing each season individually would improve the model and result in fewer outlying seasons that may skew the results too far in the positive direction.

Finally, every game played in the simulation went for exactly nine innings. During the 1941 streak, however, DiMaggio played four shortened games, in which he had only three, four, three, and two plate appearances, and four extra inning games. In the new model, two-plate appearance games are impossible, and games with three plate appearances happened only 3.1% of the time. Games with fewer or even more plate appearances are another real-world source of variability that should decrease the chance of a streak.

4.2 Why don't hitting streaks obey our rules?

One of the most fascinating sabermetric papers of the last decade is Trent McCotter's 2008 study "Hitting Streaks Don't Obey Your Rules" ([10]). Using very simple techniques, McCotter observed that hitting streaks occur more often when we look at games in their actual order than in some randomly permuted order. In other words, hitting streaks of any substantial length are slightly more likely than we would expect them to be. In the paper and subsequent notes, McCotter has dismissed several possible explanations for this phenomenon, and the sabermetrics community still lacks an explanation. One possible reason is that a streaking player is "on a roll" and there is a window of time where he is more likely to have a hit than his season or career averages would indicate. This hot-hand effect is often invoked by announcers and fans as a way of predicting future events: "Player X is on a hot streak and has 6 hits in his last 14 at-bats, so he should come in to pinch-hit for Player Y, who has a better batting average this season, but is only 3 for 13 in his last three games." Sabermetricians typically deny the existence of this effect, and cite studies of free throw shooting in basketball that find no correlation between the results of consecutive free throws.

By applying the modified model presented here, it would be possible to derive a much better understanding of the phenomenon. For instance, look at all hitting streaks of ten or more games in a season. If the phenomenon holds true, there will be more of these streaks than we might expect due to games with a hit coming in close succession. By applying the model with various modifications (the ones that are already present, the ones discussed in Section 4.1, and any others that seem plausible) to each of the streaks, we can isolate which modifications produce results that more accurately match reality. These will be the modifications that are likely to be responsible for

the phenomenon generally.

5 Conclusions

5.1 Thoughts on true player skill vs. statistics

All previous models, including our model and the basic model, have estimated the probability of a lengthy hitting streak for individual players in individual seasons, and are primarily based on a player's ratio of hits to plate appearances. It seems logical that to answer the question "If DiMaggio played 1941 again, what would his probability of having a 56-game hitting streak be?" we should examine his H/PA ratio, and that this would continue to be his ratio in subsequent retrials of the season. Yet DiMaggio had a lucky season in 1941, even without considering the streak. If DiMaggio were to play the 1941 season again, not only would he almost certainly not have another streak, but we would expect his H/PA ratio to decrease to something more resembling his true ability level. In a word, DiMaggio overachieved in 1941, and indeed, any baseball record over the period of a single season or less is very likely caused by overperformance with respect to the player's ability. Estimating streak chances based on a single season's averages is thus something of a fool's errand. When we are trying to estimate the probability of an event that has happened once in the entire history of major league baseball, using probabilities that are even slightly inaccurate will give significantly biased results. Being able to accurately estimate player skill beyond statistical manifestation is essentially impossible, and would have many more applications than predicting hitting streaks, but we do need to be aware that single-season statistics are not a perfect tool to predict streaks when it comes to individual players.

5.2 Summary

Previous studies of DiMaggio's streak have failed to include enough variance in both the number of plate appearances and the chance of having a hit in a given plate appearance. Some models, such as those found in [2], have varied P_H from game to game, but this fails to account for the drop in the average value of P_H brought on by in-game variance. By creating a model that is adjusted for hit factor, platoon effects, and pitcher ability, and using a random number of plate appearances, it is possible to estimate the chance of DiMaggio's feat much more accurately, at about 1 in 14,890. This model can be easily adapted to study multiple other problems, such as calculating the chance of other famous streaks, looking at the whole of baseball history and what the streak record "should" be, and examining the phenomenon of the hot hand and a preponderance of streaks that is more than chance would predict.

References

[1] *Baseball–Reference.com - Major League Statistics and Information*, `http://www.Baseball-Reference.com`.

[2] Samuel Arbesman and Steven H. Strogatz, *A Monte Carlo Approach to Joe DiMaggio and Streaks in Baseball*, http://arxiv.org/ftp/arxiv/papers/0807/0807.5082.pdf.

[3] ———, *A Journey to Baseball's Alternate Universe*, http://www.nytimes.com/2008/03/30/opinion/30strogatz.html?_r=0, March 2008, accessed on January 2, 2014.

[4] Charles Blahous, *The DiMaggio Streak: How Big a Deal was it?*, The Baseball Research Journal **23** (1994), 41–43.

[5] Bob Brown and Peter Goodrich, *Calculating the Odds*, The Baseball Research Journal **32** (2003), 31–40.

[6] Joe D'Aniello, *DiMaggio's Hitting Streak*, The Baseball Research Journal **32** (2003), 31–34.

[7] Dan Fox, *Schrodinger's Bat*, http://www.baseballprospectus.com/article.php?articleid=4970, April 2006, accessed on June 1, 2014.

[8] Michael Frieman, *56-Game Hitting Streaks Revisited*, The Baseball Research Journal **31** (2002), 11–15.

[9] Stephen Jay Gould, *The Streak of Streaks*, CHANCE **2** (1989), no. 2, 10–16.

[10] Trent McCotter, *Hitting Streaks Don't Obey Your Rules*, The Baseball Research Journal **37** (2008), 62–70.

[11] Tom Tango, Michael Lichtman, and Andrew Dolphin, *The Book*, Potomac Books, 2007.