That's the Second-Biggest Hitting Streak I've Ever Seen! Verifying Simulated Historical Extremes in Baseball

Andrew C. Thomas*

April 2, 2010

Abstract

There is considerable interest in two consecutive game streak records in baseball, namely the celebrated 56-game hitting streak of Joe DiMaggio and the less famous 84-games-reaching-base streak of Ted Williams. How likely would these records be predicted to occur if the history of Major League baseball were repeated? I strive to answer this question through simulated replication using a series of Bernoulli-type models. I assume that the number of games played by each player in each season is held constant, while the batting and on-base averages are estimated from and shrunk towards the career trends of each player to smooth over outlying seasons. These simulation models are then verified against streaks that might be expected to occur, such as all-time streaks ranked 6 through 30, and are allowed to vary over time to reflect the changing distribution of opposing pitching. I find that a validated model for predicting hitting streaks contains no "hot hand" effect, suggests that the variability of opposing pitching has decreased markedly in the past 140 years, and that under this model the DiMaggio streak can be considered exceptional, while validated models for on-base streaks require considerably more complexity, including but not limited to a term that *dampens* on-base streaks.

1 Introduction

To fans and observers of baseball, there are fewer accomplishments held in as high regard as Joe DiMaggio's 56-game hitting streak, set during the 1941 season. This streak is especially noteworthy considering that the next three on the list lasted 45, 44 and 41 games respectively, putting forth the notion that there is something particularly magical about DiMaggio's accomplishment. It's that extra boost that has caused commentators like Stephen Jay Gould (Gould [1989]; PBS [2000]) to claim that the streak goes well beyond what the history of the game would suggest for the likeliest outcome.

This implies that any "likely" outcomes would be produced not only by hitters with high batting averages, but also those who are more likely to get a hit following another hit [Gilovich et al., 1985]. There are two typical explanations for this observation. The first is that there is a direct correlation between the outcomes for these two at-bats, so that the outcome is directly

*Visiting Assistant Professor, Department of Statistics, Carnegie Mellon University. Correspondence email: act@acthomas.ca. I thank Jim Albert, David Friedenberg, Jay Kadane, Carl Morris, Joseph Richards and Hal Stern for their helpful comments, and Samuel Arbesman for bringing the problem to my attention.

dependent on the previous result; a player's streaky behaviour is a consequence of their success (or failure) and is reinforced by their actions. The second is that there is an underlying nonstationarity that does not depend directly on the observed outcomes, but still indicates periods of high or low success; in essence, a player's outcomes could be from a "hot" period one week and a "cold" one the next, but these periods are independent of the actual plate appearance outcomes [Larkey et al., 1989]. Either or both of these explanations can explain why streaks would be longer than would merely be expected by chance.

Whatever the true underlying reasons, these sorts of epic accomplishments spin many folk tales, and Gould is but one of many to investigate the issue from several different points of view. One focus, seen in Warrack [1995] among others, is to investigate the behaviour of a single player's realized and potential streaks using a fixed-parameter Bernoulli model, which discounts both explanations for streaky behaviour and bolsters the notion that streaks simply happen as a consequence of repeated independent trials. This is used to estimate the probability of at least one hit during a game as a mechanism for simulating a player's season, or for estimating the distribution of the longest observed streak using a formula first provided in Feller [1968, Chapter 13].

Bringing this to a grander scale than one player at a time, Arbesman and Strogatz [2008a] take the approach of creating multiple simulations of the entire history of the game. This is accomplished one player at a time using the following steps:

1. Take the historical statistics of one player in the history of Major League Baseball, noting the division of each player's career into seasons. The performance statistics in this case are broadly grouped into hits (singles, doubles, triples and home runs), walks (including intentional walks and hits-by-pitches) and outs (everything else).

2. For each player-season, estimate the probability of a hit in any at-bat by their observed ratio of hits to plate appearances. (Call this term the "hitting average" to distinguish it from "batting average", the ratio of hits to hits plus outs, which removes walks and hits-by-pitches from consideration.)

3. Estimate the probability of a player getting at least one hit in a game given their hit-to-plate-appearance ratio p and the number of plate appearances per game n. This would correspond conceptually to a binomial with a non-integer number of trials and estimated as $p_{streak+} = 1 - (1-p)^n$.

4. Simulate the season by taking a Bernoulli trial for each game with the probability of starting or continuing a hitting streak as $p_{streak+}$.

Each player-season is simulated once to recreate the history of baseball; by repeating the simulation process, a distribution for the longest historical streak is obtained according to this model. An initial criticism of this method, as mentioned by Rockoff and Yates [2009], is that the hit probabilities for any given pair days should not be identical, owing to the differences in opposing pitching let alone other sources of heterogeneity such as weather or health. Arbesman and Strogatz [2008b] incorporate this into their updated analysis, though their conclusion does not change appreciably: that a hitting streak of at least 56 games would be observed in the history of the game roughly 49% of the time.

While this approach is a commendable start, there are a number of improvements that can be made to this method. One check that should be considered is the ability of the model to reproduce other properties of the system that aren't of direct interest, such as lesser hitting streaks, and other similar extreme events like consecutive games reaching base (when walks and hits-by-pitches are included). In particular, the models of Arbesman and Strogatz [2008b] simulate many more extreme streaks (30 games or more) than are present in history. (A demonstration of this is given in Section 3.) If we accept that some of these "lesser" streaks are reasonable to be observed in most theoretical repetitions of the game, then this tendency must be corrected in order to validate the probability calculations on the maximal streak.

1.1 Investigations of Streaky Behaviour

The previous simulation method assumes that there is no dependence between the likelihood of hitting or reaching base between any two games, or that such an effect is too small to measure. Despite conventional beliefs that streaky behaviour is common in many aspects of sports[1], all attempts at measuring a grand-scale dependence between random events in sports (including Gilovich et al. [1985]; Tversky and Gilovich [1989a,b]) has suggested that if there were a general non-zero dependence between two adjacent events, be they free-throw shots in basketball or plate appearances in baseball, the effect would be too small to measure with such limited data sets. This is essentially confirmed in a follow-up study [Larkey et al., 1989], though in rebuttal the authors demonstrate that streaky behaviour can be detected for a small fraction of players in the NBA.

Albright [1993] suggests that in a small but representative sample of baseball data, no streaky behaviour can be detected in hitting patterns. In their subsequent comments, Albert [1993] and Stern and Morris [1993] each suggest that this is a function of sample size as much as it is the proposed model, so that if there were any true streaky effects in the data, they would be too small to detect. This was followed by Albert [2008] in demonstrating the presence of some streaky behaviour in hits, strikeouts and home runs on an at-bat by at-bat basis; with an approximate mean of four plate appearances per game, it remains to be seen whether this effect would persist on a game-by-game basis.

1.2 Player Aging and Pooling Information

There is ample evidence to suggest that a player's realized batting average in any one year, as a direct estimate of their true ability, is a significantly biased estimate of their true ability. At a minimum, Brown [2008] demonstrates that for predicting the performance of a player for the remainder of a season, simply using the *league* average will yield more accurate predictions than a player's batting average to that point in a season. Several improved estimators for player performance were tested, including Empirical Bayes calculations and the James-Stein estimator originally used on batting average data in Efron and Morris [1975, 1977]. Each of these methods endorse the notion of "borrowing strength from the ensemble" – that is, given that there is an underlying common performance, an estimator that combines the uncertainty both within and between individual measures will often show improvement in predictive power.

As the data includes the history of each player, there is a signficant benefit to pooling information across a player's career in order to better estimate a player's "true" underlying average in any particular year. No matter what approach is taken, the goals are often the same:

- For the future, to predict the behaviour of a player given past performance; or,

[1]See for example the film Bull Durham (1988), in which a pitcher's winning streak is described as highly psychological in nature; namely that one should "never [fool] with a winning streak" when they may happen, by interfering with whatever factor the player believes is the prime factor for success.

Hitting Streak	Year		On-Base Streak	Year	
Joe DiMaggio	1941	56	Ted Williams	1949	84
Willie Keeler	1896-97	45	Joe DiMaggio	1941	74
Pete Rose	1978	44	Ted Williams	1941	69
Bill Dahlen	1894	42	Ted Williams	1948	65
George Sisler	1922	41	Orlando Cabrera	2006	63
Ty Cobb	1911	40	Johnny Tobin	1922	58
Paul Molitor	1987	39	Duke Snider	1954	58
Jimmy Rollins	2005-06	38	Barry Bonds	2003	58
Tommy Holmes	1945	37	Wade Boggs	1985	57
Gene DeMontreville	1896-97	36	George Kell	1950	57

Table 1: The longest hitting and on-base streaks in the history of Major League Baseball, including those that wrap between multiple seasons.

- For the past, to infer the most likely value for a player's true ability.

These models most often assume a that player's evolution of ability is concave in nature: that ability increases until reaching its peak, at which point ability tends to decline once again. This type of evolution has been noted for several different sports [Berry et al., 1999] and for various elements of productivity as it relates to age in many disciplines [Marchetti, 2002]; it is considered validated by the idea that the absence of these patterns in the case of Roger Clemens' [Bradlow et al., 2008] and Mark McGwire's [Albert, 1999] respective career performances are highly suggestive of unnatural physical enhancement for each player. A quadratic form is a simple yet flexible function for fitting this to player ability as a function of age [Morris, 1983; Albert, 1999] and is a natural starting point for any corrections to yearly averages for each player.[2]

1.3 The Game Over Decades

One of the perplexing results of previous historical simulations is the sheer imbalance of streaks over time in the simulated world with respect to the real world. While the top streaks appear to be balanced over time (see Table 1), many approaches that take raw batting average as an input produce historical streaks that are concentrated in the earlier days of the game, when there was considerably more variation between league batters, producing hitters who had seasons hitting .400 or better [Gould, 1986]. Any model that attempts to reproduce lower-order historical streaks should account for the broad tendencies of each era, notably that the top streak length are roughly the same in each era (see Table 2).

This notion can also be extended so that the model should include as many "epochs" of baseball as are necessary; the minimum would be one continuous period of time from 1871 until the present, and a plausible maximum could be epochs lasting one year at a time. I seek a model for the game

[2] It is worth mentioning that other systems do not necessarily require concavity in their career trajectories. The PECOTA estimation method, for example, compares a player's age-dependent career trajectory to all others in the history of the game and selects a series of nearly compatible players using nearest-neighbour matching; the careers of those compared players then serve as an estimate for the future performance of a player and can be conceived as an inferred distribution of past performance as well. Despite its success in the popular press for its predictive benefits, I cannot use this method for inference and simulation, mainly as the implementation of this algorithm is proprietary. This is not a major setback, however, as its success has mainly been used for prediction of individual trajectories in future observations, not as a corrective mechanism for earlier inferences on ability.

that is as minimal as necessary to explain the prevalance of hitting and on-base streaks to avoid any overfitting of observed streak lengths, and there is a rarity of streaks lasting more than 30 games overall; as a result, I strive to use as few epochs as necessary. Since the stated goal is to examine the likelihood of the hitting and on-base streaks of DiMaggio and Williams, which both occurred in the 1940s, a three-epoch would be advised using pre-1940, the 1940s, and post-1949 period as the epochs of interest, withholding the middle period from any model fitting.

1.4 Methods and Goals

Having considered these approaches, I turn to the driving questions behind this investigation: how unlikely to occur are the record-setting 56-game hitting streak of Joe DiMaggio and the 84-game on-base streak of Ted Williams, as shown in Table 1? In order to answer these questions I need several elements:

1. A data set of player performance through the history of baseball. For this I use the Lahman baseball database [Lahman, 2009], which contains the yearly statistics of all hitters in Major League Baseball from 1871 until 2009. This does not contain the game-by-game performances of each player or the performance by each plate appearance, which also means that specific pitcher matchups are unknown in each case.

2. A method to better infer the most likely batting average and on-base average for a player in each season, taking into account the player's entire career. This is covered in Section 2.1.

3. A method to simulate the distribution of the number of plate appearances each player makes within a (nine-inning) game, shown in Section 2.2.

4. A quantity that models the additional variability in these quantities on a daily basis. In general, this can account for opposing pitching and fielding, weather conditions, or any other phenomena that can be seen as an uncorrelated daily process. This is demonstrated in Section 2.3.

5. A method to include game-by-game correlation of outcomes, either positive (to encourage streaky behaviour) or negative (to dampen it). By default this parameter is set to zero for both hitting and on-base. This is shown in Section 2.4.

6. A test of fit for less relevant quantities that should still be expected in the model, namely the lengths of lesser streaks. Note that in this approach, I assume that the occurrence of said lesser streaks is unremarkable; that is, that streaks of these lengths would be expected in most instances of the history of baseball.

A full exposition of this treatment and the results it yields is done progressively, first by producing the null model with steps 1-3, and then verifying the fit of this with the test in step 6. Each element is only added as necessary to produce a series of simulated streaks that are validated against those historical outcomes we expect to observe.

2 Design

Here I lay out the elements involved in simulating the history of Major League Baseball as previously listed, beginning with the records for each player and then the required smoothing of averages over a player's career. I then show a method for estimating the number of plate appearances per game before adding additional daily heterogeneity and/or direct game-to-game dependence.

I first begin with the data on each player in each year. All together, this is equivalent to the following data table for player i in year j:

Standard Quantities	Hits	H_{ij}
	Bases-on-balls	BB_{ij}
	Hits-by-pitches	HBP_{ij}
	At-bats	AB_{ij}
Derived Quantities	Times Reaching Base	$RB_{ij} = H_{ij} + BB_{ij} + HBP_{ij}$
	Plate Apperances	$PA_{ij} = AB_{ij} + BB_{ij} + HBP_{ij}$
	Hitting Average	$HA_{ij} = \frac{H_{ij}}{PA_{ij}}$
	On-Base Average	$OBA_{ij} = \frac{RB_{ij}}{PA_{ij}}$

Compare also the grand averages across the league in any particular year[3], $\widehat{HA}_j = \frac{\sum_i H_{ij}}{\sum_i PA_{ij}}$ and $\widehat{OBA}_j = \frac{\sum_i RB_{ij}}{\sum_i PA_{ij}}$.

2.1 Adjusting Hitting and On-Base Averages

Each of the quantities presented in the previous section are necessary to estimate the most likely value of the hitting average in each year under the assumption of a quadratically-based career curve for any player; this procedure is also conducted for the most likely on-base average. The following method is used to estimate the most likely batting average:

1. Set $Y_{ij} = logit(HA_{ij}) - logit(\widehat{HA}_j)$. This allows for the accounting of year-to-year fluctuations in league average.

2. Fit the transformed yearly data to the curve

$$Y_{ij} = \beta_{0i} + \beta_{1i}(year)_{ij} + \beta_{2i}(year)^2_{ij} + \varepsilon_{ij}$$

   with weights corresponding to the number of plate appearances in each year; that is, $\varepsilon_{ij} \sim N(0, \sigma^2/PA_{ij})$ giving a weight matrix with diagonal terms $W_{jj} = PA_{ij}$ and zeroes on the off-diagonal, so that each year's performance is considered conditionally independent of the others. Representing $X'_j = [1\,(year)_{ij}\,(year)^2_{ij}]$, this gives estimates for the predicted transformed hitting average curve as

$$\beta = \begin{bmatrix}\beta_0\\ \beta_1\\ \beta_2\end{bmatrix} \sim N_3\left((X'WX)^{-1}X'WY, \sigma^2(X'WX)^{-1}\right).$$

[3]For most of the history of baseball, the American League and National League (and American Association prior to 1901) have remained essentially separate in their operations; this suggests that it may be prudent to consider the averages within each league for adjustment rather than across all of baseball. This is unnecessary for two reasons: first, league hitting averages are roughly equal; second, the use of league averages is only to make transformations that are subsequently undone, and their minuscule difference is unimportant to this operation.

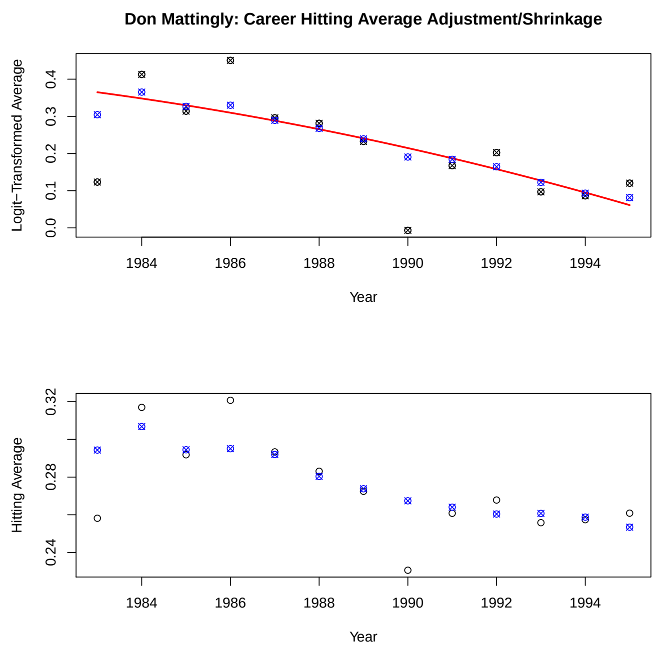


Figure 1: Steps showing how the estimated performance for a player in a particular year can be adjusted to take into account both the number of plate appearances and the player's overall career curve. This diagram is for the career of Don Mattingly, who played for the New York Yankees from 1982 until 1995; his first season contains only 7 games and 13 plate appearances and is not considered in the adjustment or simulation.

3. Given the uncertainties in the data and in the fitted curve, use Empirical Bayes estimates for the most likely average in each season. This corresponds to a weighted average of the likelihood for each point and the corresponding estimate for the hitting curve. If the estimated curve corresponds to

$$Y_i^{(curve)} \sim N(\mu^{(curve)} = X(X'WX)^{-1}X'WY, \Sigma_i^{(curve)} = \sigma^2 X(X'WX)^{-1}X'),$$

   and the uncertainty at each data point is

$$y_{ij}^{(point)} \sim N(\mu_{ij}^{(point)} = y_{ij}, \sigma_{ij}^{2\,(point)} = \frac{\sigma^2}{PA_{ij}}),$$

   then the shrinkage factor is calculated as the ratio of the curve's inverse variances (precisions) to the sum of each,

$$B_{ij} = \frac{\frac{1}{\sigma^{2\,(curve)}}}{\frac{1}{\sigma^{2\,(point)}} + \frac{1}{\sigma^{2\,(curve)}}}$$

   to obtain the Empirical Bayes predictive distribution

$$\widehat{\mu}_{ij} \sim N(B_{ij}\mu^{(curve)} + (1 - B_{ij})\mu_{ij}^{(point)}, \frac{1}{\frac{1}{\sigma^{2\,(point)}} + \frac{1}{\sigma^{2\,(curve)}}}.$$

   The mean value will be closer to the fitted curve than to the observed data point in those years with low PA. Either the mean of this distribution can be used,

$$\widehat{Y}_{ij} = B_{ij}\mu^{(curve)} + (1 - B_{ij})\mu_{ij}^{(point)},$$

   or a draw can be taken from the distribution for every simulated season, $\widehat{Y_{ij}} = \widehat{\mu_{ij}}$. For this analysis, the distributional approach is used, as it assumes that our estimate of player ability is uncertain.[4]

4. Apply the reverse transform to the data to get a point estimate for the hitting average,

$$\widehat{HA}_{ij} = logit^{-1}(\widehat{Y}_{ij} + logit(\widehat{HA}_j)).$$

An example of this procedure is given in Figure 1 for the career of Don Mattingly.

There are many extensions that can be made to the basic model. For example, there is considerably more pooling that can be accomplished through the methods of Berry et al. [1999] by compiling information across time. For the purposes of this investigation, I find it sufficient to smooth out those seasons that are outliers with respect to the rest of a player's career.

[4]The analyses were conducted with both approaches and gave virtually identical results, suggesting that the uncertainty in the average is dwarfed by the uncertainty in the process itself.

2.2 Simulating the Number of Plate Appearances Per Game

As this data set does not contain the distribution for the number of plate appearances per game, I approximate this with a distribution based on the Negative Binomial nature of the batting order. A full nine innings represents a total of 27 failures, so that the total number of plate appearances, by all players, in a single game (labelled k) can be approximated as $A_{ijk} \sim NBin(27, p_{ij})$, where p_{ij} represents the probability of an out. Over an entire simulated season of G_{ij} games, this would correspond to a total number of plate appearances $A_{ij} = \sum_k A_{ijk} \sim NBin(27G_{ij}, p_{ij})$.

One player's share of this will then be approximately one-ninth of the total. If the same player were to bat in every position in the lineup, the total number of plate appearances would be $A_{ij} = \sum_k A_{ijk} = 9 * PA_{ij}$; the effective failure probability of such a team during one at-bat is then estimated as $p_{ij} = \frac{r}{EA_{ij}} = 27 * \frac{G_{ij}}{9 * PA_{ij}}$.

An estimate for the distribution of at-bats for any player is to draw for the game-length outcome A_{ijk}, divide this by nine, and round to the nearest integer. When simulated, this approach produced a slight upward bias in the total number of at-bats for a season; a slight correction to the failure probability allows the expected value of each A_{ijk} times the number of games to match the observed number of plate appearances.

There are several minor flaws with this approach that should be noted. First, this of course discounts those games that are ended before nine innings due to weather. This is an event rare enough to ignore in explicit cases, and if necessary can be approximated with day-to-day heterogeneity.

Second, in games where the home team leads after eight and one half innings, the bottom half of the ninth inning is not played; this occurs when the home team has proven more successful, i.e. when the player has had more opportunities to bat, mitigating the consequences of this circumstance.

Third, there are cases where players are given the day off and brought in to make a pinch-hit appearance. Given the importance that hitting streaks have in the lore of baseball, it is unlikely that a Major League manager would allow a player on a hitting streak to have a day off (nor would any such player be likely to request one.) However, this circumstance can prove to be a concern in events that are less celebrated such as on-base streaks.

2.3 Game-to-Game Heterogeneity

The simulation mechanism proposed for a single game is now straightforward: draw from the number of plate appearances PA_{ijk} in game k, and draw from a Binomial distribution with this number of trials, and success probability $\widehat{HA}_{ij}$ in the case of hitting streaks; if this draw is greater than or equal to one, the streak continues.

Variability in the daily success probability can be induced by adding a noise term; essentially, we have $\widehat{HA}_{ijk} \sim N(\widehat{HA}_{ij}, \sigma_H^2)$, so that σ_H is a characteristic degree of heterogeneity. The addition of this noise can be made linearly or as a part of a linear transformation such as a logit or probit; in the linear case, hitting averages can be bounded to fall within the $[0, 1]$ interval. I choose a linear addition in this step purely to make the difference in probability equal for all players, rather than using a transformation that may generate a probability adjustment that is conditional on true performance at this stage.

Prior beliefs on the degree of heterogeneity have to be considered before adding parameters to a model that may not make any direct physical sense. Stern and Sugano [2008] conduct an examination of hitter and pitcher heterogeneity using Empirical Bayes methodologies and show that for a selection of opposing pitchers, New York Yankee shortstop Derek Jeter can be estimated to have extremely low variability of performance, with $\sigma_H \approx 0.005$. At the grand scale of this analysis, such a small degree of heterogeneity may not be necessary to include. At the same time, it would be highly implausible to see a batter's match-up heterogeneity to be more than, say, $\sigma_H = 0.1$, so that a player's hitting average could vary by as much as 0.4 between pitchers; this would likely be an indication of pure pitcher heterogeneity beyond what would normally been seen in a major league sport.

2.4 Adding Streak-Inducing or Streak-Damping Behaviour

To include a game-to-game dependence in the likelihood of continuing a streak, a Markov component can be included in the model so that the probability of a hit or of reaching base within one game is affected by whether a similar event was observed in the previous game. Let μ_H represent the change in the hitting average following a game in which a hit was recorded, and $-\mu_H$ be the change in average if no hit was registered. (Let μ_B be the corresponding value for the rate of bases-on-balls.) For positive μ_H, this will increase the expected length of both hitting streaks and hitting slumps given a player's baseline hitting average.

As stated, the addition of this dependence suggests that random draws need to be taken sequentially, which can lead to much slower computational time. In practice, the computation and simulation can be implemented by first generating a series of uniform variates between -1 and 0 at the beginning of the operation to represent each plate appearance in a game, adding the game's hitting average to each value, and recognizing that a hit has occurred in a plate appearance if the corresponding variate is greater than zero. Each draw can then be sequentially adjusted based on the previous outcome to reflect an increased or decreased probability of a hit, without requiring the regeneration of any Bernoulli random variables; this approach leads to a considerable speed-up in systems where random variate generation is computationally expensive.

2.5 Software

I have now described all the pieces necessary for this investigation of historical streak lengths in Major League Baseball. The R code and data for this procedure is included supplementally on the author's website. For most analyses, 100 replications are sufficient to check whether the model is producing an accurate representation of historical streak behaviour.

3 Analysis of Hitting Streaks

Since the underlying assumption of this model is that some streaky behaviour is expected, and since streaks are the only output that I seek to reproduce, I note that the model should be able to produce historical streaks that are not considered "magical" in the same sense; namely, streaks of lower order on the record list should be predictable. At the same time, streaks are highly dependent with respect to their neighbours, as it is difficult to examine each position independently. To that end, I consider the sum of lower-order streaks to be the test statistic in each case, namely the sixth- through thirtieth-longest streaks of each type. For hitting streaks, this is the sum of Ty Cobb's 40-game streak to two 30-game streaks in positions 29 and 30, for a total of 839 streak-games.

As each replication of the model produces a streak table and a streak sum of each type, the p-value approach is immediately applicable, as the fraction of simulated values that are greater than the observed value: if the observed value differs greatly from the model predictions, the model

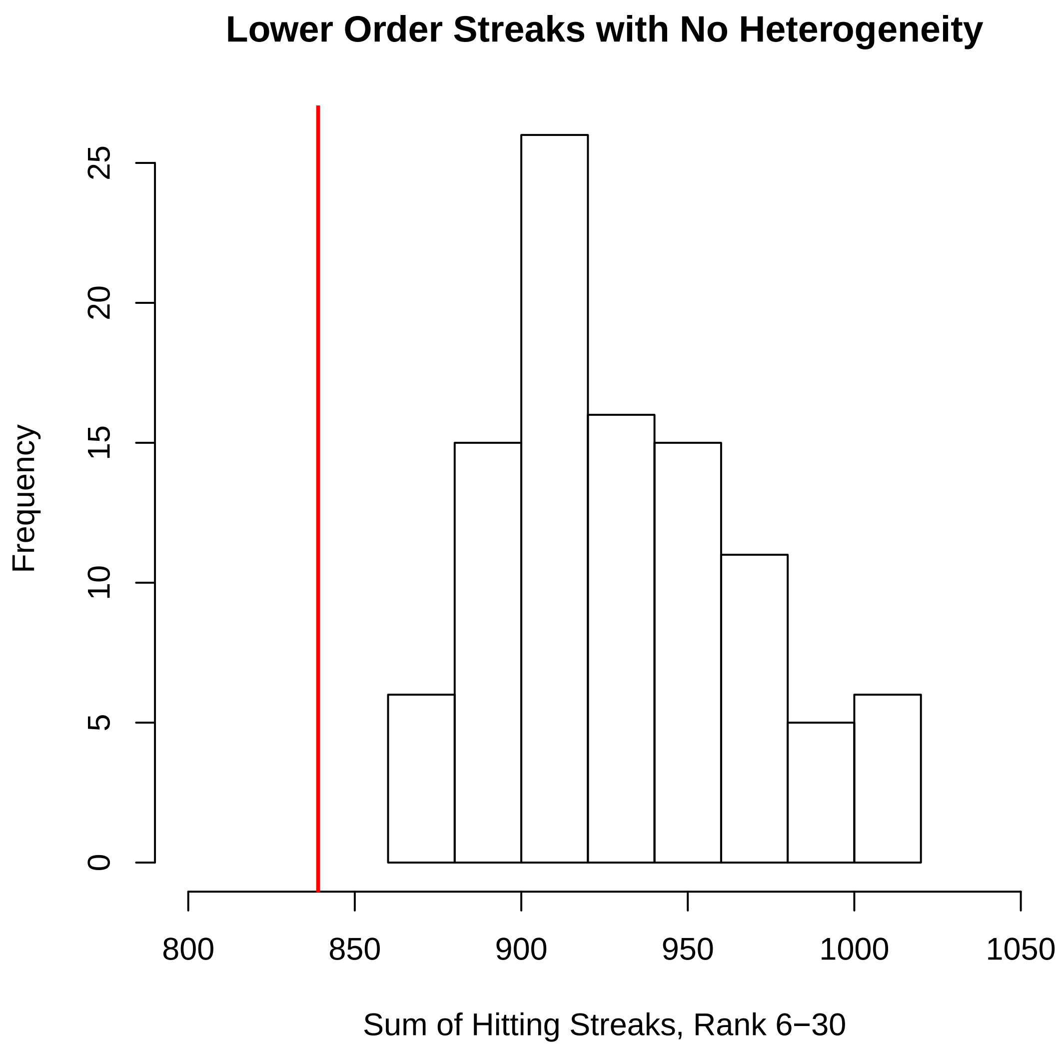


Figure 2: After simulating 139 seasons of baseball 100 times over without adding additional game-to-game heterogeneity in ability, a test statistic is calculated: the sum of the sixth- through thirtieth-longest hitting streaks, equal to 839. If the model is valid, and streaks of these lengths are considered likely events, then the model should produce a statistic that is comparable to reality. Even after smoothing for seasons that are outliers, simulations (in the histogram) produce streaks that in aggregate are always longer than in reality (the solid vertical line) for hitting streaks.

can be rejected. As seen in Figure 2, this is exactly the case for the model without heterogeneity for hitting streaks. The cumulative lengths of hitting streaks are greater for all simulations than they are for the real data, even after smoothing the extreme values against the career curves of players, yielding simulated p-values less than 0.01 in each case. This model cannot satisfactorily explain the cumulative lengths of lesser streaks; since I assume that these lesser streaks are to be expected, the model needs additional heterogeneity.

3.1 Models with Game-to-Game Heterogeneity in Hitting Ability

To gauge which models will be adequate to the task of simulating streaks, so that lower-order streaks will be of comparable length, I now introduce heterogeneity into the simulation process. With hit standard deviation σ_H selected to be a value between 0 and 0.11 (in intervals of 0.01), 100 simulations of each scenario are performed.

The results of these simulations are displayed in Figure 3. For each trial and each streak type I note the sixth through thirtieth top streaks overall, as well as the top 15 streaks from 1871-1939 and 1950-2009 respectively. The validity of the trial parameters is again established by checking that the sum of the comparison streaks is within the range of the sums in the simulations for each quantity; the goodness of fit is found with the sum of squared differences for each position.

The simulations best fit the real data overall for $\sigma_H = 0.06$, suggesting that there is considerable variation between pitchers' ability to stop a streak in progress. However, at this value there is a clear imbalance between different epochs the pre-1940 and post-1949 epochs. This value of variability is appropriate for the earlier era – in fact, better performance is achieved for this era with $\sigma_H = 0.08$ or $\sigma_H = 0.1$ – but smaller additional variability for the later era, as small as zero but taking the minimum mean SSE at $\sigma_H = 0.01$, is sufficient to produce these expected hitting streaks.

Splitting the earlier epoch in two shows another strong difference between the two halves. Dividing the game into the "pre-modern" era, before the creation of the American League in 1901 (and the annual World Series, beginning in 1903), and the first decades of the modern era, from 1901 until 1939, suggests that a great deal of the "required" heterogeneity is needed in the pre-modern era, with an optimal value of $\sigma_H = 0.1$ under the simulations; the quantity needed for the beginning of the modern era is $\sigma_H = 0.06$.

Whatever the optimal estimated values are in each epoch, there is a distinct decrease in day-to-day variance required to produce simulated streaks that are validated by the observed data. It is interesting to note that this validates an unrelated point about the variability of player performance over time. Gould [1986] has suggested that the variability of hitters has decreased over time as the players approach the physical limits of human ability, combined with a much larger pool of talent from which to draw. These findings suggest that the same rule applies to pitchers as time advances: as the sport evolves and the talent pool increases, the relative abilities of pitchers have contracted as well. This would also explain the small estimated difference of certain players against a field of pitchers [Stern and Sugano, 2008].

3.2 The DiMaggio Streak Was Special, For the Modern Era

I have produced a reliable and robust model for historical hitting streak production that is consistent with perceptions of the game, and run a large number of simulations under these assumptions. It now remains to collect these simulations and check the 56-game DiMaggio streak, as well as the Keeler, Rose and Sisler streaks of 45, 44 and 41 games, and see where these lie in the historical record. This was accomplished by "grafting" the simulations under the very-heterogeneous model

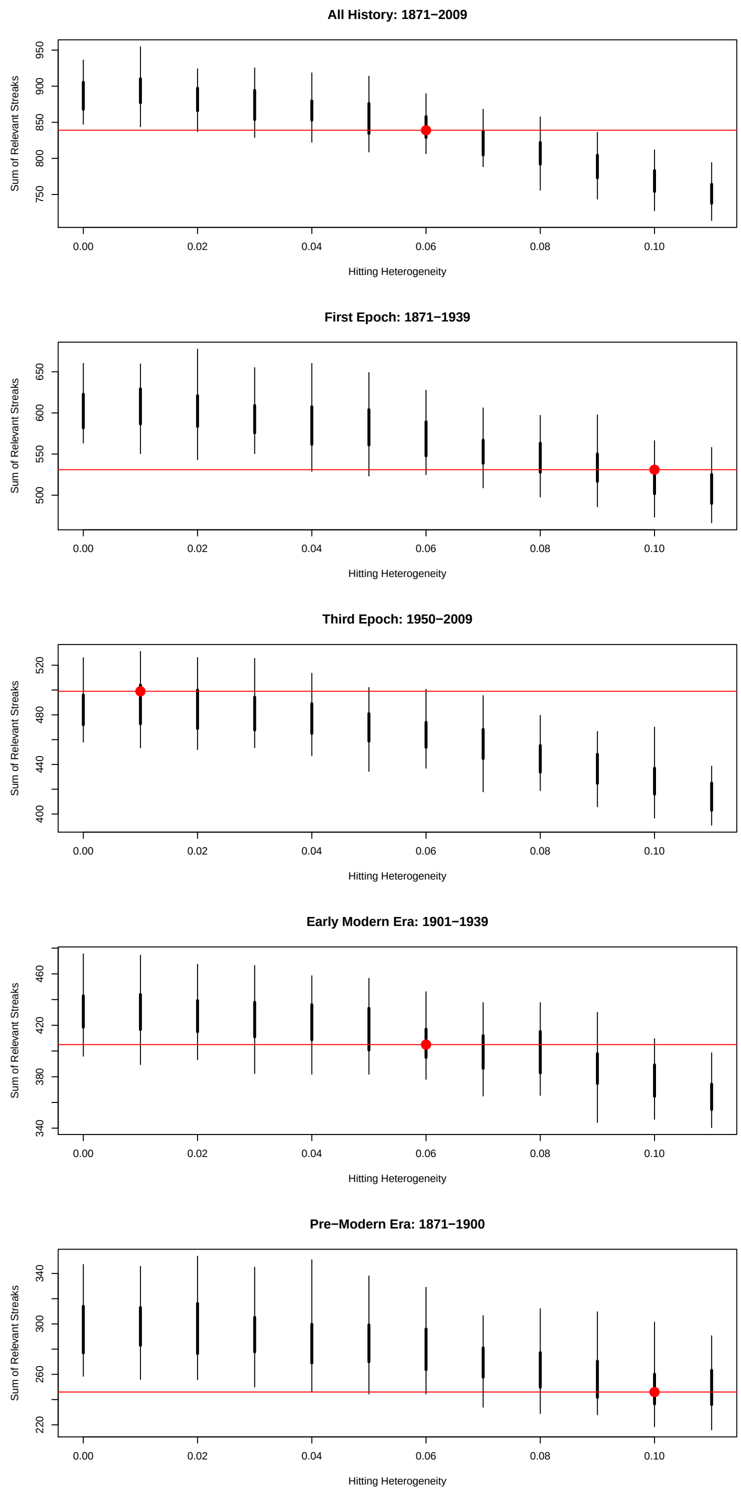


Figure 3: Each column represents the distribution of the sums of the relevant lower-order streaks, as a function of additional heterogeneity; thick bars represent the interquartile range, thin bars represent the central 95% confidence interval. The horizontal line represents the observed historical value in each case. A point on the horizontal line represents that the indicated column has the minimum sum of squared discrepancies between each simulated streak and its historical counterpart. For the history of organized baseball, the fit is best for a heterogeneity of roughly 0.06. However, when dividing the history into multiple epochs, the ideal level of heterogeneity appears to decrease over time, from 0.1 in the pre-1900 era, 0.06 from 1901-1939, decreasing to 0.01 from 1950 until the present day.

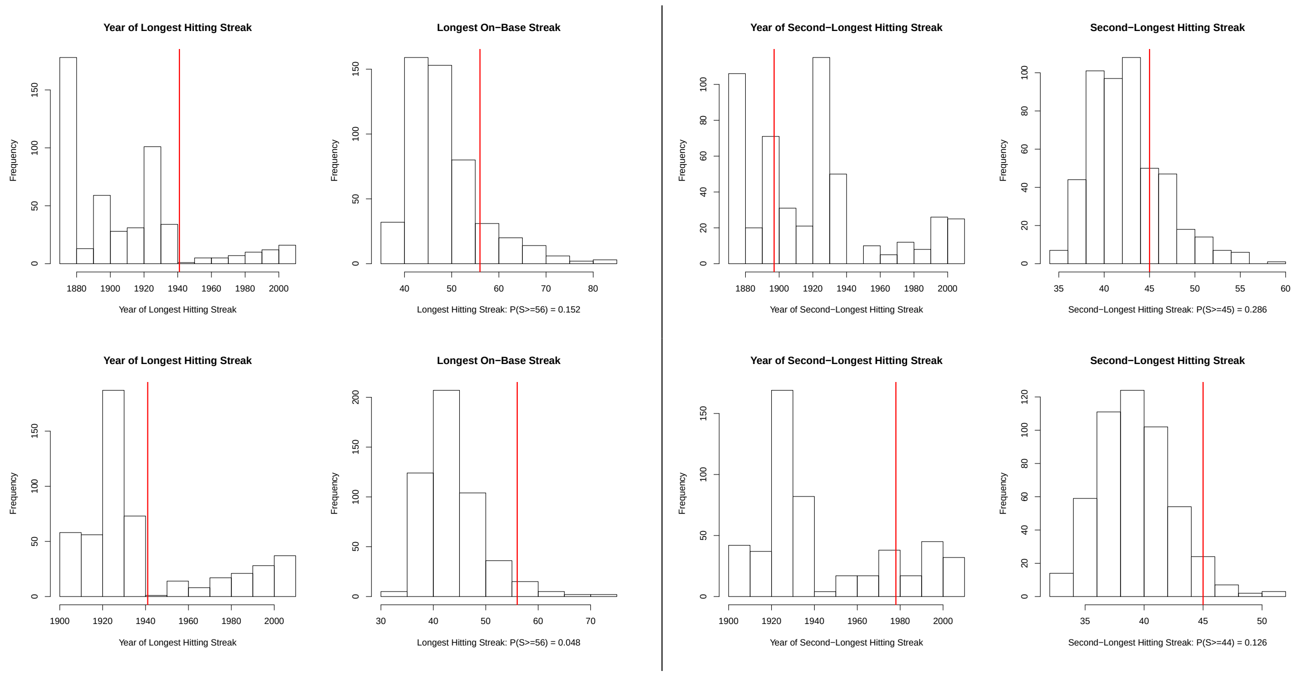


Figure 4: Left, the behaviour of the longest hitting streak in simulations and reality; on the right, the second-longest hitting streak is considered. The top represents the full period of time, 1871-2009; the bottom represents the modern era 1901-2009. The spikes in the 1870s, 1890s and 1920s are due largely to a small number of players.

for pre-1900 baseball and the fairly-heterogeneous model for the beginning of the modern era to the relatively uniform pitching of 1950 onwards. The only remaining parameter is to choose what model should be used for comparison in the 1940s, or whether a "compromise" heterogeneity value would be more appropriate.

This method explicitly excludes the period in the 1940s from the model fitting process as this is the area where the prediction is to be made. Given that lower heterogeneity between pitchers favors longer streaks, and that the goal is to find a lower bound on the extremeness of the DiMaggio streak, choosing the period to have the heterogeneity of the later era gives the batters of the 1940s a period that would favour streaky behaviour, and is the most conservative choice for estimating the probability of a more extreme streak.

Figure 4 shows the results of 500 simulations under this scheme. With the pre-modern history of the game included, the probability under this model of observing a more extreme hitting streak is estimated as 15.2%; considering only the modern era, that probability drops to 4.8%. These values are similar for the second-longest streak: with all history, Willie Keeler's 45-game streak is equalled or exceeded in 28.6% of simulations; with the modern era, Pete Rose's 44-game streak is exceeded 12.6% of the time. This suggests that the DiMaggio streak is certainly exceptional, though not necessarily "magical" under the assumption that streak lengths in lower positions on the all-time list are to be expected; where this does begin to approach the stuff of legend is in the modern era, from 1901 until the present.

The sharp discrepancy between these situations, and the spikes in the 1870s, 1890s and 1920s, suggest that even with a four-epoch model there is still more to investigate. In particular, the sheer number of simulated hitting streaks that lead the pack from the 1870s, in an era with far fewer games per season, is enough to cast doubt on the model's suitability for the pre-modern era, especially since only 7 hitting streaks of 30 games or more were available to validate the model in this era.

4 An Analysis of On-Base Streaks Is Less Successful

Being satisfied that the model produces valid results for hitting streak behaviour, I now move to on-base streaks, which have received considerably less attention. It would seem practical to begin with the four-epoch hitting streak model and to assume that the game-to-game heterogeneity in the ability to draw a walk is added to that for getting a hit, so that the act of reaching base can be treated with the same simulation method as for a hit, except with on-base average substituted for hitting average.

The effective rate of walks is considerably smaller than that for hits on average. Given this, a reasonable upper bound for walk heterogeneity σ_B is on the order of 0.05; this would suggest that the walk rate allowed by pitchers would vary from roughly no walks to upwards of 20% of the time in 95% of cases. As this walk rate is considerably higher than that for most pitchers who remain at the major league level, it suggests a reasonable upper bound for the effect without adding any game-to-game dependence.

The test statistic for model fit is once again a sum of "expected" streaks in lower positions. For the pre-modern era, the top twelve streaks total 662 streak-games (streaks less than 50 games long are not in the current record). For the early modern era, the top thirteen streaks total 681 streak-games. For baseball post-1949, the top fifteen streaks total 823 streak-games.

As seen in Figure 5, the model without heterogeneity of any kind generates test statistics well in excess of the observed value, yielding simulated p-values less than 0.01 in each case. This model cannot satisfactorily explain the cumulative lengths of lesser streaks; since I assume that these lesser streaks are to be expected, the model needs additional heterogeneity.

Figure 6 shows the behaviour of on-base streaks for a variety of heterogeneity parameters, from $\sigma_B = 0$ to $\sigma_B = 0.05$, a reasonable maximum value for the degree of external heterogeneity, during each of the three epochs under consideration. Each column contains 100 simulated histories, with vertical position representing the sum total of the top streaks in each epoch. For the pre-modern era, a walk heterogeneity of $\sigma_B = 0.05$ is sufficient to produce simulated streaks that are consistent with the observed data. This does not continue for the modern era; in particular, while simulations of the early modern era appear to be approaching consistency with observed streaks (without reaching it), the lack of agreement is striking up until the present, with the average top streak length predicted by the model roughly 50% bigger, on the order of 90 games rather than 60.

On its face, the current modelling approach is inconsistent with producing streaks of the appropriate length. Extra measures need to be taken to propose a plausible grand model for on-base streaks.

4.1 Streak-Dampening Behaviour in On-Base Records Is Insufficient

While historical hitting streak patterns can be explained without needing to add a term for explicit dependence, there seems to be little way to explain the grand pattern of historical on-base streaks without one, as they are considerably longer in simulations than they are in reality. The fact that the simulated streaks are longer than their observed counterparts suggests that a term to *dampen*

Figure 5: After simulating 139 seasons of baseball 100 times over without adding additional game-to-game heterogeneity in ability, a test statistic is calculated: the sum of the sixth- through thirtieth-longest on-base streaks, equal to 1387. If the model is valid, and streaks of these lengths are considered likely events, then the model should produce a statistic that is comparable to reality. Even after smoothing for seasons that are outliers, simulations (in the histogram) produce streaks that in aggregate are always longer than in reality (the solid vertical line) for on-base streaks.

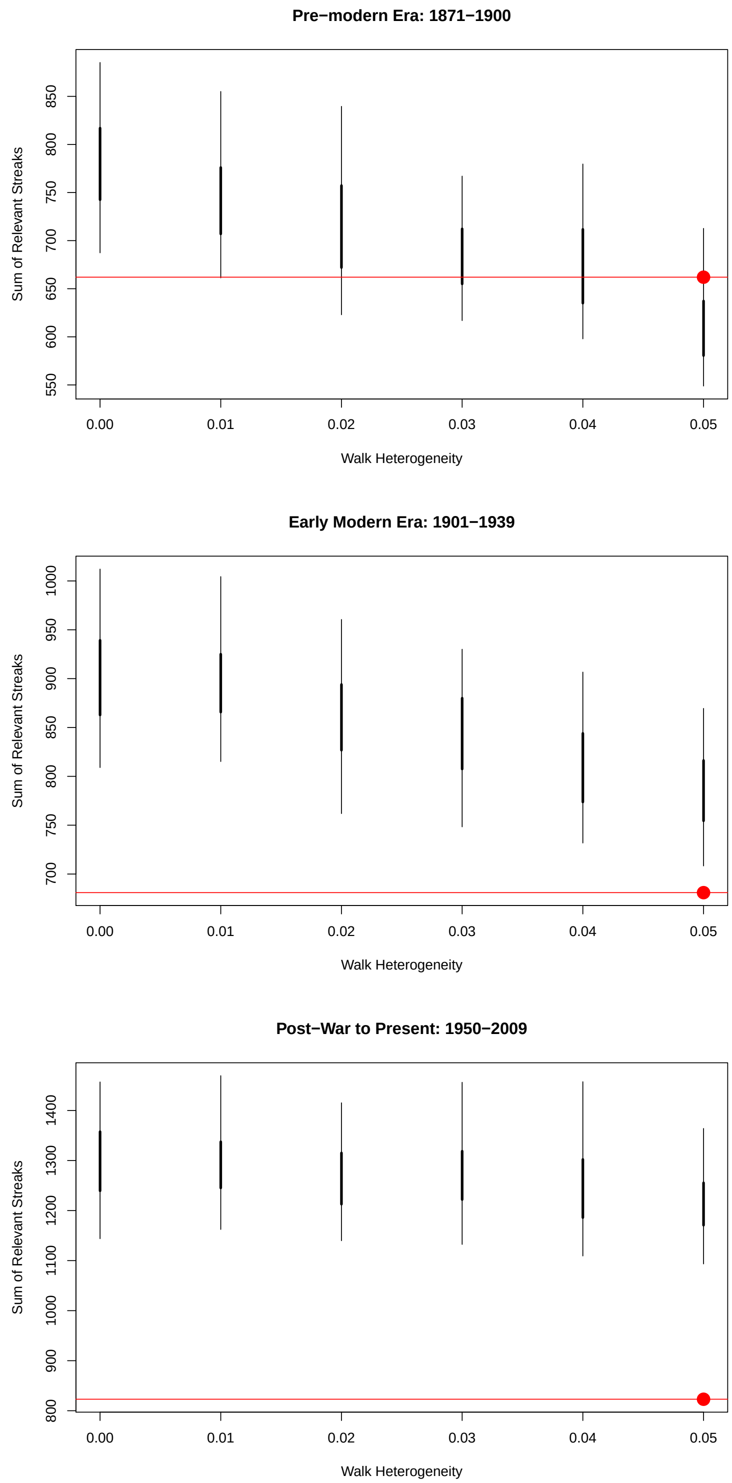


Figure 6: Each column represents the distribution of the sums of the relevant lower-order streaks, as a function of additional heterogeneity; thick bars represent the interquartile range, thin bars represent the central 95% confidence interval. The horizontal line represents the observed historical value in each case. A point on the horizontal line represents that the indicated column has the minimum sum of squared discrepancies between each simulated streak and its historical counterpart. These plots correspond to the pre-modern (1871-1900), early modern (1901-1939) and present-day (1950-2009) eras. Hitting heterogeneity in each case corresponds to the optimal values found in Section 3.1. While a reasonable value of on-base heterogeneity in the pre-modern era produces a model whose streak lengths are as expected, no reasonable value has the same effect for the modern era.

a streak is appropriate for the model, a notion that goes against both lay and expert knowledge on the nature of streaks in sports.

For the sake of constructing a plausible model, I adapt the current set-up to include a streak bonus term μ_B, which is the increase in the walk rate in games where the batter reached base the previous game, and the decrease of the walk the rate following a game without reaching base. A negative bonus therefore indicates a streak-damping term.

Figure 7 shows the results of simulations in the modern era for various streak bonus values; streak bonuses of $\mu_B = -0.025$ and $\mu_B = -0.045$ for the early and later modern eras are adequate to the task. But there is a major issue with this method, in that it is extremely unbalanced: players who are more likely to reach base in a game than not are effectively having their on-base averages lowered overall. Even in the extreme case – suggesting that the chances of reaching base on days following a zero-base game is as close to certainty is possible – the model still suggests that the only way to achieve reasonable historical streak values with a grand-scale method is to artificially lower the on-base averages of the very players most likely to set the streaks.

4.2 Possible Factors to Improve Modelling On-Base Streaks

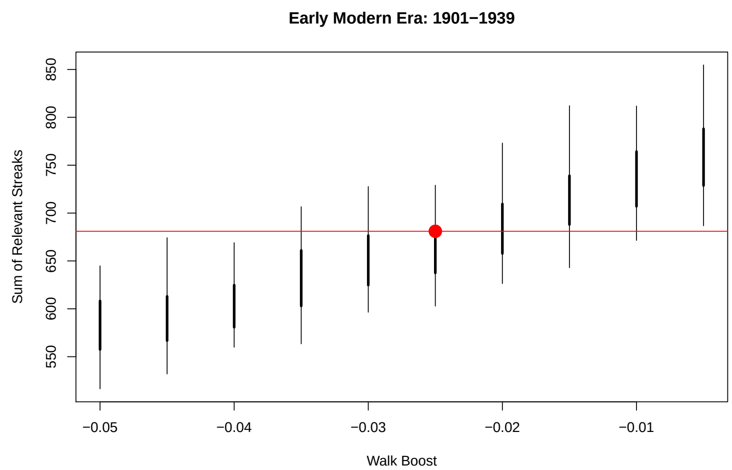
It would seem that of the models that can act on the grand scale to estimate historical on-base streaks, each makes too many unreasonable modelling choices to achieve the goal. Rather than assume that the problem is analogous to that for hitting streaks, it would be more worthwhile to examine where the assumptions between the two models differ.

As the recognition of a hitting streak may stop a manager from sitting a "hot" player, no such recognition has existed for on-base streaks or walks. Indeed, the appreciation of the base-on-balls as an offensive tool has fluctuated over time, reaching the height of popularity in the 1920s before its resurgence associated with the rise of the Oakland Athletics as documented by Lewis [2004] in the past 15 years. Even Ted Williams's record for consecutive games reaching base has rarely received attention as a record worth approaching, as it has been perceived as a consolation prize when a hitting streak comes to an end; indeed, DiMaggio's second-place 74-game on-base streak begins with his 56-game hitting streak, and held the record for eight years. This element then calls into question whether a manager would rest a player on an on-base streak, then bring them in for a pinch-hit appearance and end said streak without respect for the record.

There is also considerably more variability in on-base average among elite players. Barry Bonds is currently the record-holder for highest on-base average in a season with .609 in 2004; his 2002 and 2003 values of .582 and .529 are numbers 2 and 9 on the top 10 list as well.[5] From these numbers alone, one would expect him to have lengthy on-base streaks as well, and his 2003 streak of 58 games ranks number 8 all time[6]. However, those simulations that fit to lower-order on-base streaks consistently show Bonds at the top of the leaderboard, often with streaks exceeding 100 games in length. Why Bonds' streaks are not higher on the list may be explained by chance, but they may be better explained by the sheer number of intentional walks taken by Bonds during each of these seasons – 68, 61 and 120 in 2002, 2003 and 2004 respectively, the top three totals in the history of the recording of the statistic – which are tactical decisions by managers, an element of the game I have not attempted to model. If further analyses of these records are carried out with a better understanding of the intentional walk, perhaps there will be better insight into why Bonds did not perform as well in 2004, especially as he is far from the only player to collect a large number of intentional walks.

[5] Source: http://www.baseball-almanac.com/hitting/hiobp3.shtml.

[6] This streak was ended with a game in which Bonds failed to reach base in four plate appearance.



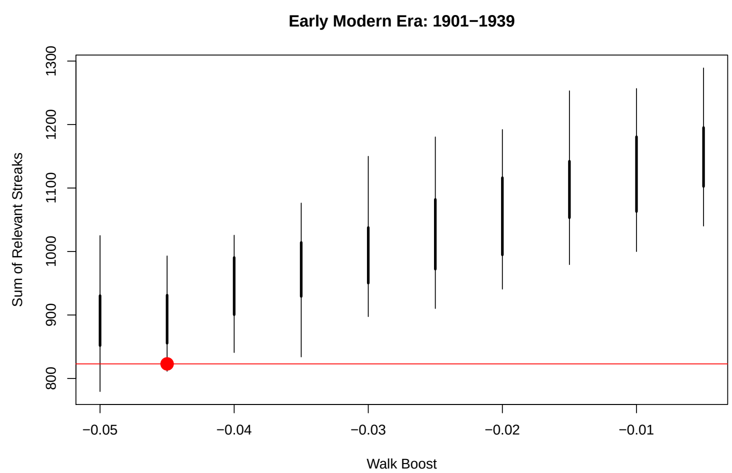


Figure 7: The mean sums of squared errors for on-base streaks as a function of the added streak bonus term for walks, in the early modern (1901-1939) and present-day (1950-2009) eras. Hitting heterogeneity in each case corresponds to the optimal values found in Section 3.1, and walk heterogeneity equals the maximum allowed value of $\sigma_B = 0.05$. In the early modern era, a streak bonus of $\mu_B = -0.025$ is optimal for producing on-base streaks that are consistent with observed data. In the present era, a streak bonus of $\mu_b = -0.045$ is the optimal value. Whether these models are to be believed together as a plausible explanation for the lengths of on-base streaks is left as an exercise to the reader.

5 Conclusions: Grand-Scale Analyses and Future Investigations

While there is too much uncertainty, even with 140 years of data, to produce definitive answers to the hitting streak or on-base streak problems, this modelling approach has revealed several important details of the history of the game.

5.1 Heterogeneity Is Important, And Useful

Day-to-day heterogeneity in ability, either in hitting or on-base average, must be incorporated in these simulation models to produce lower-order streaks that match the observed data. Interestingly, this heterogeneity appears to be decreasing over time for hitting – from roughly $\sigma_H = 0.1$ in the pre-modern era to $\sigma_H = 0.01$ in the present day. Rough (and presently unverified) estimates for heterogeneity in walks have remained high throughout history.

This story is consistent with the observations of McCracken [2001], which suggested that a pitcher's ability to prevent hits on balls in play is, in general, wildly overstated, and that the most consistent ability level of a pitcher is in actions that are independent of fielders, namely walks and strikeouts. In the later epoch (the subject of most of McCracken's analyses) the parameters that can best produce streaks of the correct magnitude correspond to these same concepts: there is considerable heterogeneity necessary for on-base streaks, corresponding to a wide distribution in pitcher ability for walks, and virtually no true underlying heterogeneity for hitting streaks. This analysis suggests that there was a strong difference among pitchers, or in the condition of the game, in the earlier epochs, periods not explicitly studied by McCracken [2001].

This suggests that in terms of differences of ability for the action of balls in play, pitchers have likely come as close to the upper limits proposed by Gould [1986] as can be detected. However, if there is any such upper limit for defense-independent statistics, it would appear that the limit of this human ability is still a long way off, and that it is this difference in pitcher ability that can still be adequately exploited both by hitters at the plate, and by organizations in their scouting departments.

5.2 Additional Bayesian Considerations

The proposed modelling approach goes as far as assigning a fixed starting batting average to a player in each year, estimated by Empirical Bayes shrinkage toward a career curve. The motivation for choosing this method was to smooth out averages that are unusually high compared to the rest of the career for the player; for Ted Williams, it is extremely unlikely that during 1941 he was a true .406 hitter, as that would imply that in roughly half the time, he would have performed even better than that; considering the rest of his career, his accomplishment is much more likely to be a combination of a high batting average (but not as high as .400) and a degree of luck. And at the crudest scale, this smoothing approach is sufficient to reproduce those lower-order streaks.

The next logical step would be to construct a model where information is pooled across players, not merely over the career of a single player. This is the approach taken by Berry et al. [1999] though not in the context of this particular problem. The only remaining question would be whether it would produce a substantially different result for hitting streaks; such a pooling scheme would not likely cause the on-base streak problem to disappear.

5.3 Individual Contributions Are Still Vital

This analysis has focused on a few specific quantities as considered across almost 140 years of history, and as such is focused only on grand trends with less attention to individual properties. For painting a broad picture, this is sufficient for hitting streaks but not for on-base streaks. It would seem likely that future investigations of on-base streaks will need to incorporate the idiosyncratic behaviour of individuals in order to estimate grand trends.

However, just because these modelling choices may be applicable at the grand level does not mean that they much apply to individuals or subgroups. While the hitting streak model suggests that there need not be a game-to-game hot-hand (or cold hand) effect, this in no way discounts the notion that *some* players may truly have this kind of trend in their data (see Larkey et al. [1989] for more on this notion). It may prove incredibly difficult to detect, in the same way that the ability to get a hit in "clutch" situations may be a factor with only a small fraction of the thousands of players throughout the history of the game; in fact, if there is a true effect for some players, it may be nearly impossible to detect thanks to multiple comparison problems.

5.4 Will Either Record Fall?

With pitcher variability for allowing hits at an all-time low, the circumstances are certainly present for someone to make a run at the DiMaggio streak. Assuming that the talent distribution is much like it was in the past century, these results suggest that the likelihood of the hitting streak record being broken in the next century are less than one in fifty.

The model for on-base streaks is on far less solid ground, since these are much more variable to pragmatic decisions by managers. The emergence of a single player as unique as Barry Bonds in his late career is an event that defies prediction in a standard statistical model, though this does give an indication of the type of player who would be likely to set the new record: a patient, high-average, high-power hitter whose presence would encourage opposing pitchers to allow walks more often.

A Streak Records

For the eras of interest, 1871-1939 and 1950-2009, the top 15 hitting and on-base streaks are given. Note that within each category, the distributions in each era are quite similar.

References

Albert, J. (1993): "A Statistical Analysis of Hitting Streaks in Baseball: Comment," *Journal of the American Statistical Association*, 88, 1184–1188.

Albert, J. (1999): "Bridging Different Eras in Sports: Comment," *Journal of the American Statistical Association*, 94, 677–680.

Albert, J. (2008): "Streaky Hitting in Baseball," *Journal for Quantitative Analysis in Sports*, 4, URL http://www.bepress.com/jqas/vol4/iss1/3.

Albright, S. C. (1993): "A Statistical Analysis of Hitting Streaks in Baseball," *Journal of the American Statistical Association*, 88, 1175–1183.

Player	Hitting Streak Year (End)	Games	Player	On-Base Streak Year (End)	Games
Willie Keeler	1897	45	Bill Joyce	1891	64
Bill Dahlen	1894	42	George Van Haltren	1893	60
George Sisler	1922	41	Cupid Childs	1892	57
Ty Cobb	1911	40	Jake Stenzel	1895	57
Fred Clarke	1895	35	Ed Delahanty	1896	56
Ty Cobb	1917	35	Bill Joyce	1896	56
George Sisler	1925	34	Arky Vaughan	1936	56
George McQuinn	1938	34	Billy Hamilton	1896	55
George Davis	1893	33	Ty Cobb	1915	55
Hal Chase	1907	33	Bill Joyce	1894	54
Rogers Hornsby	1922	33	Ray Blades	1925	54
Heinie Manush	1933	33	Luke Appling	1936	53
Ed Delahanty	1899	31	Danny Lyons	1887	52
Nap Lajoie	1906	31	Ty Cobb	1914	52
Sam Rice	1924	31	Tris Speaker	1920	52
Pete Rose	1978	44	Orlando Cabrera	2006	63
Paul Molitor	1987	39	Duke Snider	1954	58
Jimmy Rollins	2005.5	38	Barry Bonds	2003	58
Luis Castillo	2002	35	George Kell	1950	57
Chase Utley	2006	35	Wade Boggs	1985	57
Benito Santiago	1987	34	Ryan Klesko	2002	56
Willie Davis	1969	31	Jim Thome	2002	55
Rico Carty	1970	31	Derek Jeter	1999	53
Ken Landreaux	1980	31	Shawn Green	2000	53
Vladimir Guerrero	1999	31	Alex Rodriguez	2004	53
Stan Musial	1950	30	Jim Wynn	1969	52
Ron LeFlore	1976	30	Greg Gross	1975	52
George Brett	1980	30	Tony Phillips	1993	52
Jerome Walton	1989	30	Frank Thomas	1996	52
Sandy Alomar Jr.	1997	30	Gary Sheffield	2002	52

Table 2: The top 15 hitting and on-base streaks for play before 1940 (top) and after 1949 (bottom). These were obtained in Spatz [2007].

Arbesman, S. and S. Strogatz (2008a): "A Journey to Baseball's Alternate Universe," *New York Times*, URL http://www.nytimes.com/2008/03/30/opinion/30strogatz.html.

Arbesman, S. and S. H. Strogatz (2008b): "A Monte Carlo Approach to Joe DiMaggio and Streaks in Baseball," Unpublished manuscript.

Berry, S. M., C. S. Reese, and P. D. Larkey (1999): "Bridging Different Eras in Sports," *Journal of the American Statistical Association*, 94, 661–676.

Bradlow, E., S. Jensen, J. Wolfers, and A. Wyner (2008): "Report Backing Clemens Chooses Its Facts Carefully," *New York Times*, URL http://www.nytimes.com/2008/02/10/sports/baseball/10score.html.

Brown, L. D. (2008): "In-season Prediction of Batting Averages – A Field Test of Empirical Bayes and Bayes Methodologies," *The Annals of Applied Statistics*, 2, 113152.

Efron, B. and C. Morris (1975): "Data Analysis Using Stein's Estimator and its Generalizations," *J. Amer. Stat. Assoc*, 70, 311–319.

Efron, B. and C. Morris (1977): "Stein's Paradox in Statistics," *Scientific American*, 236, 119–127.

Feller, W. (1968): *An Introduction to Probability Theory and Its Applications*, volume 1, John Wiley and Sons.

Gilovich, T., R. Vallone, and A. Tversky (1985): "The hot hand in basketball: On the misperception of random sequences," *Cognitive Psychology*, 17, 295–314.

Gould, S. (1986): "Entropic Homogeneity Isn't Why No One Hits .400 Any More," *Discover*, 7, 60–66.

Gould, S. J. (1989): "The Streak of Streaks," *Chance*, 2.

Lahman, S. (2009): "Sean Lahman's Baseball Archive: Data from 1871-2009," URL http://www.baseball1.com/, online resource.

Larkey, P., R. Smith, and J. Kadane (1989): "It's Okay to Believe in the Hot Hand," *Chance*, 2, 22–30.

Lewis, M. (2004): *Moneyball : The Art of Winning an Unfair Game*, WW Norton and Company.

Marchetti, C. (2002): "Productivity Versus Age," Technical report, Richard Lounsbery Foundation.

McCracken, V. (2001): "Pitching and Defense: How Much Control Do Hurlers Have?" *Baseball Prospectus*, URL http://www.baseballprospectus.com/article.php?articleid=878.

Morris, C. N. (1983): "Parametric Empirical Bayes Inference: Theory and Applications," *Journal of the American Statistical Association*, 78, 47–55.

PBS (2000): "Joe DiMaggio: A Hero's Life," URL http://www.pbs.org/wgbh/amex/dimaggio/maps/maptxt.html, online resource.

Rockoff, D. M. and P. A. Yates (2009): "Chasing DiMaggio: Streaks in Simulated Seasons Using Non-Constant At-Bats," *Journal of Quantitative Analysis in Sports*, 5.

Spatz, L., ed. (2007): *The SABR Baseball List and Record Book: Baseball's Most Fascinating Records and Unusual Statistics*, Scribner.

Stern, H. and A. Sugano (2008): "Baseball Decisions and Small Samples," *Chance*, 20, 40–47.

Stern, H. S. and C. N. Morris (1993): "A Statistical Analysis of Hitting Streaks in Baseball: Comment," *Journal of the American Statistical Association*, 88, 1189–1194.

Tversky, A. and T. Gilovich (1989a): "The Cold Facts About the 'Hot Hand' in Basketball," *Chance*, 2.

Tversky, A. and T. Gilovich (1989b): "The 'Hot Hand': Statistical Reality or Cognitive Illusion?" *Chance*, 2, 31–34.

Warrack, G. (1995): "The Great Streak," *Chance*, 8.