

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

Andrew C. Thomas*

September 28, 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, and to speculate about the nature of “streaky” outcomes in general.

The simplest mechanism to generate streaks is a series of independent Bernoulli trials with identical probability of success for each trial. This requires only two terms to specify, the number of trials and the probability of success – in this case, that at least one hit (or walk) will be achieved

*Visiting Assistant Professor, Department of Statistics, Carnegie Mellon University. Correspondence email: act@acthomas.ca. I thank Jim Albert, David Friedenber, Jay Kadane, Carl Morris, Joseph Richards, and Hal Stern for their helpful comments, and Samuel Arbesman for bringing the problem to my attention. An anonymous referee also deserves my thanks for pointing out (among other things) the permutation approach to the problem in Section 2.3.

by the batter in any particular game. In this simple case, the success probability is assumed to be constant throughout a player’s career. This was the approach of [Warrack \[1995\]](#) (among others) which discounts additional explanations for streaky behaviour and bolsters the notion that streaks simply happen as a consequence of repeated independent trials, by using a formula first provided in [Feller \[1968, Chapter 13\]](#) to estimate the distribution of a player’s longest streak in their career.

The next logical extension from this is to assume that a player’s batting average varies from year to year, and to calculate the distribution of longest streaks for the history of an entire league, an approach taken by [Arbesman and Strogatz \[2008a\]](#). Their method, uses simulation rather than asymptotic formulae, takes these 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 large number of times, the model generates a distribution for the longest historical streak (and the second-longest, and so forth.) 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, health or the number of plate appearances. [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, and in the context of replicating the history of baseball, there would be little to be seen as “supernatural” about the streak that we would not have seen one like it at some point in the past 140 years.

But there are still questions about the external validity of this model that must be addressed. One question is the ability of the model to reproduce other properties of the system that aren’t of direct interest, such as long hitting streaks that do not hold the record but are nonetheless impressive. If the burning question is “was the DiMaggio streak exceptional?” then there must be streak-lengths that are *expected* to have occurred, and the generative model should adequately replicate that expectation. In particular, the models of [Arbesman and Strogatz \[2008b\]](#) simulate many more streaks lasting at least 30 games than actually occurred. (A further demonstration of this is given in Section 3.4.) If we accept that some of these “lesser” streaks are to be expected in simulated histories, then this tendency to oversimulate must be corrected in order to validate any probability calculations on those maximal streaks of interest.

If the “expected” streaks cannot be adequately explained through an independent-trials model, then a mechanism may be introduced to allow dependence between two games. The phenomenon of the “hot hand” [Gilovich et al., 1985], or the idea that success in neighbouring games is positively correlated, has long been a strongly-held folk belief, with two apparent explanations for this type of phenomenon. The first is that there is a direct correlation between the outcomes for these two at-bats, so that the outcome is directly 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, though only the former could explain why historical streaks would be shorter than predicted by the independent draw model: a negative correlation on success between consecutive games.

Goals

This modelling approach is designed to simulate only one aspect of the history of baseball – the occurrence and length of the most extreme streaks – by fitting to quantities that would have been expected, namely those extreme streaks that are lower down the record list. Each position on the record list represents an order statistic, and each of these are highly correlated – for example, knowing that positions 9 and 11 are both 35-game hitting streaks determines that position 10 will also be a 35-gamer. Rather than try to fit each individual position on a highly correlated multivariate distribution, I choose to fit to a single statistic in each case, namely the sum of streaks 6 through 30, and proceed with the assumption that streaks of roughly this length would have occurred in a do-over of history. The reason for choosing this particular range is two-fold: on the top end, I do not know how many streaks should be considered extraordinary, and five seems a fair estimate; on the bottom end, there are few records about the longest streaks below those that are 30 games in length in the available record from 1871 until 1950, and the 30th streak is of length 30 games. Note that this analysis can easily be repeated with different assumptions and more thorough information on streak distributions should it become available.

Following a deeper review of other analyses of the streak problem in sports, I demonstrate how to build a simulation study using 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 (or game-by-game statistics), which also means that specific pitcher matchups are unknown in each case.

Using the yearly batting average as the true batting average, and using a fixed number of at-bats per game, overpredict the length of lesser-order streaks; both of these choices are known to inflate streak lengths¹.

There are several conclusions I have reached about this modelling approach for hitting streaks:

- A simulation that adjusts batting averages and simulates a varying number of plate appearances per game are insufficient to eliminate the systematic excess in lengths of simulated

¹For the batting average: regression to the mean suggests that repeating those seasons where a player’s batting average was exceptionally high, like Ted Williams’s .406 average in 1941, would yield a season with a high batting average, but most likely less than .406; shrinkage models attempt to correct for this.

hitting streaks (see Section 3.4.)

- Adding a source of game-to-game heterogeneity in batting ability is enough to bring the distribution of the simulated streak sum to center on the observed outcome, but there is an extreme time skew: the vast majority of the longest simulated streaks are in the pre-modern era, before 1900. As seen in Table 1, the longest streaks from the 1950s onwards are roughly the same collective length as those before World War II.
- By dividing baseball history into several epochs – the pre-modern era (1871-1900), the early modern era (1901-1940), and the post-war era (1950-present), temporarily withholding the 1940s as a “test” epoch – and applying a different source of heterogeneity to each (lots for the pre-modern era, less for the early modern, and very little for the postwar) – brings the sum streaks for each era into alignment. By first applying the heterogeneities of the early modern era, then of the post-war era, to the 1940s, I obtain a range of estimates for the exceptionality of the DiMaggio streak in this context.
- This reduction in day-to-day heterogeneity of ability over time is highly suggestive of a substantial reduction in the variability of opposing pitchers, as far as a batter’s ability to get a hit is concerned, which extends the notions of Gould [1986], who suggests a reduction in variability of hitters, and validates the work of McCracken [2001] in showing that the differences in modern pitcher ability do not revolve around preventing hits (on balls in play).
- Because each of these corrections forces the sum-total of the longest streaks downwards in order to line up with reality, no extra assumptions about “hot hand” streaky behaviour are necessary to explain the longest hitting streaks under this model. (This does not mean that there is no such effect in the game.)
- None of these methods are sufficient to explain the lengths of the top on-base streaks, which are much longer under simulations than their real-world counterparts. Proposed modifications to this scheme to dampen

2 Review: Streak Analysis and Player Performance

There are many strong conventional beliefs that streaky behaviour is common in many aspects of sports. An illustration of this in baseball forms an important plot point in the 1988 film [Bull Durham](#), 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. Despite this, all attempts at measuring a grand-scale dependence between random events in sports (including [Gilovich et al. \[1985\]](#); [Tversky and Gilovich \[1989a,b\]](#)) have 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 \[2008b\]](#) 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.

2.1 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 an inappropriate choice. 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 significant 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,
- 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

2.2 The Game over Decades

In previous simulation attempts (e.g. [Arbesman and Strogatz \[2008b\]](#)) there is a large concentration of extreme streaks that take place in the earlier eras of the game, but in reality, the top

²It 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.

streaks appear to be balanced over time (see Table 1). This is likely due to the input itself, as there was considerably more variability between league batters. This led to more .400-or-better hitters during these times [Gould, 1986], which lead in simulations to a much higher probability of continuing a streak, even though the batters of this era played fewer games in a season. Any model that attempts to reproduce lower-order historical streaks should differentiate by time, in one way or another. In fact, there is considerable interest in asking why the record-holder for the streak did *not* come from before 1900, given the high batting averages of the day, begging for its inclusion as a separate object of study.

One way to do this is simply to divide the game into a series of “epochs”, and to fit parameters for each epoch corresponding to its lesser streaks. While a feasible minimum would be one continuous period of time from 1871 until the present, I show in Section 4.1 that this is insufficient. While a plausible maximum could be a series of epochs lasting one year at a time, the data on streaks preceding the 1940s does not allow for this level of precision [Rockoff and Yates, 2009], typically limited to those few instances when hitting streaks of 30 games or more have been recorded.

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, I choose to model the history of the game in four epochs: the pre-modern era (1871-1900), the early modern era (1901-1939), the 1940s (1940-1949) and the present era (1950-2009). The first epoch represents a time when the rules of the game were still in flux, and results (shown in Section 4.1) demonstrate its distinguishing differences from the second era. The goal is then to fit any parameters to the lesser record streaks in each era; I withhold the 1940s from this fitting for the time being, since it contains the streaks of interest, then apply a range of the parameters fitted from the two adjacent epochs to this time period. This allows us to obtain a range of estimates for the time period of interest.

The exercise of fitting this model to a larger number of epochs, say decade-by-decade, or using “change point” methods to determine from the data where steady epochs with constant parameters might exist, is left for future analysis, largely dependent on the availability of detailed hitting streak data between 1871 and 1950.

2.3 Simulation versus Permutation

Rather than using a method of direct simulation, McCotter [2008] pursued the question of hitting streaks by a method of permutation: start with a player’s game-by-game statistics in a given year, permute the outcomes to create a hypothetical season, and record the streaks that result. By repeating this process for every player in every year from 1957 until 2006, each iteration of this permutation process creates a hypothetical history, and ten thousand iterations produces confidence intervals under this model. The author found that this permutation method creates *shorter* streaks in this period than would be observed, demonstrating that the observed streak counts in this period from 5 games to 35+ games were all higher in magnitude than their simulated means. However, the author also notes that when removing games in which the player does not start – making a pinch-hitting appearance – there is a considerable decrease in the discrepancy between truth and simulation, though the discrepancy remains. (I will return to this point in the conclusions.)

In response to this article, Albert [2008a] points out that the permutation test makes it difficult to remove the confounding factors of why streaks appear to be longer than chance would indicate – namely, the difference between the hot hand and nonstationarity. Additionally, the streak effect does not seem to hold up when examined on a yearly basis, which one might expect if the effect were evident, and that with 30,000 player seasons to consider, the significance tests are effectively

tests of a large sample size.

In this case, since the data are not yet available for the bulk of baseball history (including the 1940s) the method is not currently implementable for examining the game prior to 1950, but would likely prove illuminating to determining an approximate measure of the net departure from an independent Bernoulli model.

3 Initial Design

The available data from 1871 onwards are yearly statistics for player performance from the Lahman baseball database [Lahman, 2009]. These statistics form the basis for a game-by-game simulation of each player-season. Initially, there are two elements that must be accounted for, to generate the probability of a game beginning or extending a streak: number of plate appearances per game that gives the player the opportunity to extend the streak, and the probability of a hit in a single plate appearance. I first review the available data set and establish notation, then propose a method for simulating the number of plate appearances per game, and finally establish a method for choosing a hitting average other than simply taking the observed result.

3.1 Elements of the Data Set

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 :

	Hits	H_{ij}
Standard	Bases-on-balls	BB_{ij}
Quantities	Hits-by-pitches	HBP_{ij}
	At-bats	AB_{ij}
	Times Reaching Base	$RB_{ij} = H_{ij} + BB_{ij} + HBP_{ij}$
Derived	Plate Appearances	$PA_{ij} = AB_{ij} + BB_{ij} + HBP_{ij}$
Quantities	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 year³, $\widehat{HA}_j = \frac{\sum_i H_{ij}}{\sum_i PA_{ij}}$ and $\widehat{OBA}_j = \frac{\sum_i RB_{ij}}{\sum_i PA_{ij}}$.

3.2 Simulating the Number of Plate Appearances Per Game

This data set does not contain the distribution for the number of plate appearances per game, only the season totals. (While other databases, as previously mentioned, contain some data on the distribution of plate appearances per game, these do not extend across the entire history of the game, just the last 63 of a total of 139 seasons.) For this reason, I construct an approximation

³For 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.

based on the idea that the total number of at-bats in a game is, roughly, a **Negative Binomial distribution**: the total number of at-bats required to reach 27 outs (failures) in a nine-inning game. The probability of an out is just one minus the probability of reaching base safely.

The 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 aforementioned probability of an out. Over an entire simulated season of G_{ij} games, this would correspond to a total number of plate appearances $A_{ij} = \sum_k A_{ijk} \sim NBin(27G_{ij}, p_{ij})$.

A starting 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 consequences of this approximation that must 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. Related to this is the occurrence of a double play, in which two batters are retired on the same plate appearance, or outs that are made by baserunners during a player's plate appearance, which can only occur in cases of other success. Either occurrence will slightly lower the number of potential opportunities to extend a streak.

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, due in part to the Bull Durham criterion.) However, this circumstance may prove to be a concern in events that have been less celebrated, such as on-base streaks.

3.3 Adjusting Hitting and On-Base Averages

There is by no means a single method that is considered "best" when it comes to determining a player's actual level of ability at a particular time. The simplest method would simply be to use the raw batting/hitting average statistic as an estimate of true ability (this is, after all, the unbiased statistical estimator for the true ability) but among the many reasons for going beyond this simple measure is the idea that information about a player's other performances can improve the estimation of their true ability. For example, if Ted Williams had hit .400 in two consecutive years, with a large number of at-bats in each case, it would be far more plausible that he was a "true" .400 hitter in at least one of them, than in the case where he performed the feat the one time; as it stands, the accomplishment in 1941 was likely due to a combination of a high natural batting average (one that was less than .400, but higher than his career average of .344) and a lucky season. This underlying natural average will occasionally yield a .400 performance, but will

reach or exceed it far less than half the time. By failing to correct for this, overly optimistic batting averages will yield inflated hitting streaks.

Several possibilities present themselves for estimating the “true ability” of a player in a particular year. In each case I consider a method of partial pooling: all the data points are evidence of an underlying structure, and the estimate of their true ability is a weighted average of the observed outcome and this structure; the weights are determined by the uncertainty in both factors. For example, a .400 average from a 2-for-5 batting line is far more uncertain than one from a 200-for-500 line (indeed, the latter estimate has a tenfold-smaller standard error.) This method of “shrinking” an observed outcome toward a common source [Brown, 2008] has the advantage that the collective error over all our estimates is reduced. One possibility for a choice of structure would be to shrink toward the grand average value for all players in a given year, but this ignores the longitudinal information on each player.

For this study I consider each player to have their ability follow a quadratic career curve, but before fitting these curves, a more concrete definition of ability is necessary. Assuming that the mean level of hitting talent changes very slowly from year to year (an assumption supported by Berry et al. [1999], which uses a decade-level time scale to bridge various eras), and that any annual deviations in the mean talent are caused by systematic factors like changes in the rules, I choose the measure of ability to be the hitting average relative to the league average \widehat{HA}_j . In particular, I transform this by taking the logistic transformation,

$$Y_{ij} = \log \frac{HA_{ij}}{1 - HA_{ij}} - \log \frac{\widehat{HA}_j}{1 - \widehat{HA}_j}.$$

This quantity is centered at zero (corresponding to average ability), allows for Y_{ij} to be unbounded even though the averages themselves are bounded between zero and one, and the transformation is linear in the region around zero so that linear shifts in the transformed average are also linear in the realized average.

Since this quantity represents the ability that will rise and fall during a career, we model this with a quadratic curve with the year on the x-axis:

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

where $(\beta_{0i}, \beta_{1i}, \beta_{2i})$ are the quadratic coefficients for the career curve of player i .⁴

The number of plate appearances is also the weight of each term; that is, $\varepsilon_{ij} \sim N(0, \sigma^2/PA_{ij})$ so that we have 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 $X'_j = [1 \quad (year)_{ij} \quad (year)_{ij}^2]$, and X as the matrix of these terms across all years, standard weighted linear regression gives estimates for the quadratic coefficients as

$$\beta_i = \begin{bmatrix} \beta_{0i} \\ \beta_{1i} \\ \beta_{2i} \end{bmatrix} \sim N_3 \left((X'WX)^{-1} X'WY, \sigma^2 (X'WX)^{-1} \right).$$

⁴The next level of analysis would treat these quantities as fully exchangeable, rather than simply each having a quadratic form; that is, there is a common statistical process that generates all these curves for each player, and sharing information across players would improve the estimation of these curves. If there is interest in each player’s career curves with respect to the rest of the league, there would be good reason to include this explicitly; see Jensen et al. [2009]; Carruth and Jensen [2007] for examples. Since this process would yield more corrections to shorter careers, which are far less likely to have produced record streaks (due to fewer opportunities and presumed lesser ability), it is less essential to carry this out, and I leave it for a future analysis of the phenomenon.

The fitted curve has an inherent uncertainty, as do each of the data points. I use the Empirical Bayes method to combine these estimates, so that the predicted value of each point is the weighted average of the fitted curve and the data, with the weights corresponding to the uncertainties.⁵

The estimated career curve for player i corresponds to $Y_i = X_i\beta_i$, or

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

the corresponding uncertainty at each data point is a function of the number of plate appearances,

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

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

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

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

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

or a draw can be taken from the distribution for every simulated season, $\hat{Y}_{ij} = \hat{\mu}_{ij}$. For this analysis, the distributional approach is used, as it assumes that our estimate of player ability is uncertain.⁶

Given the corrected value of a player's relative ability, I apply the reverse transform to the data to get a point estimate for the hitting average,

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

An example of this procedure is given in Figure 1 for the career of Don Mattingly. The quadratic curve on his adjusted hitting average suggests that his best years were at the beginning of his career with a slow decline following (even though during this decline, his performance was always above the league average.) His performance in 1983 and 1990 is far below his career expectation, but each of these were also due to a small number of at-bats, due to his rookie season and injury respectively. The corrected values put his likely achievement at being slightly below his expected performance by the curve, but still far above these realized values.

⁵This corresponds to the idea that the initial estimate, $p(\hat{Y}|Y)$ is the "prior" distribution for the data, and that the regression expression $p(\beta|Y)$ is the likelihood, so that $p(\hat{Y}|Y, \beta)$, the replication value for the transformed hitting average, is the posterior distribution of the data itself.

⁶The 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.

Don Mattingly: Career Hitting Average Adjustment/Shrinkage

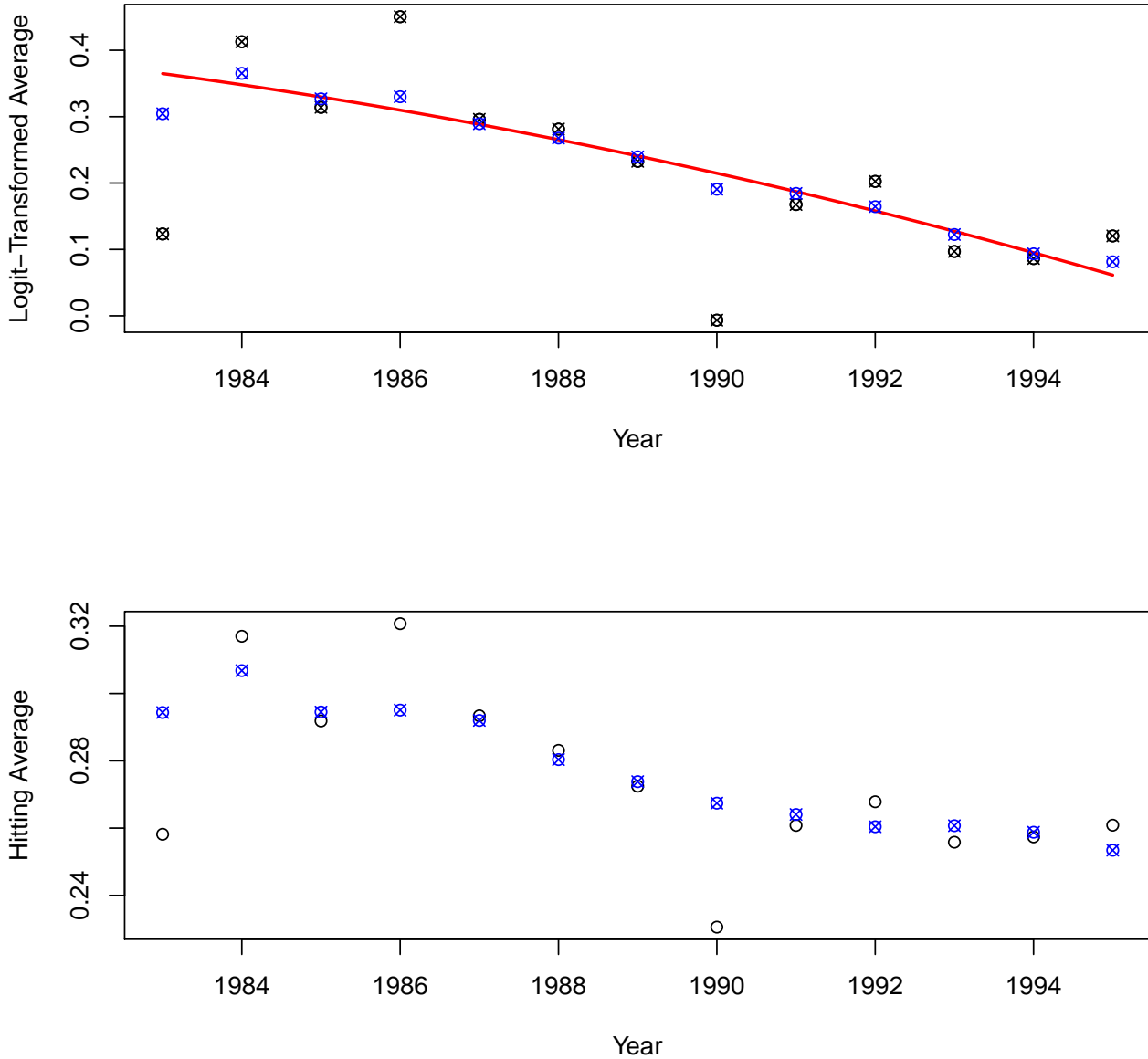


Figure 1: Top, the adjusted hitting measures for Don Mattingly, who played for the New York Yankees from 1983 until 1995 (7 games in 1982 are omitted.) Black points are his actual achievements; the red line is the quadratic regression line for these points with respect to the year; and the blue points are the estimates for the most likely “true” ability. Bottom, the observed statistics (black circles) and adjusted estimates (blue points) for hitting average. Note that below-average performances in 1983 and 1990, and overachieving performances in 1984 and 1986, are smoothed toward his career performance.

3.4 Adjustments to Seasonal Average and Varying Plate Appearances Are Insufficient to Explain Streaky Behaviour

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.

Having estimated a newly corrected hitting average for each player that smoothes out extraordinary performances, and designed an approximation for the distribution of at-bats per game, I now use these to run 100 simulated histories of Major League Baseball and record the sum of streaks six through thirty as described in Section 1. In reality, this equalled 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 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.

4 Introducing Game-to-Game Heterogeneity

The analysis so far has assumed that whatever differences that exist in ability from day to day are negligible to their impact on streaks, but the results of the previous section have essentially shown that this is insufficient. It is known that differences in the quality of the opposing team (pitchers and otherwise), weather conditions and minor injuries, among others, can have a substantial impact on the performance of a player. Because the data do not have the richness required to separate the impacts of these factors, I approximate these day-to-day factors as an additional source of heterogeneity in ability, essentially an added noise term for each game,

$$\widehat{HA}_{ijk} \sim N(\widehat{HA}_{ij}, \sigma_H^2),$$

so that σ_H is a characteristic degree of heterogeneity on hitting average. 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, though there is no reason why we cannot consider adjustments on other scales.

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 across a range of pitchers; this would likely be an indication of pure pitcher heterogeneity beyond what would normally be seen in a major league sport.

Lower Order Streaks with No Heterogeneity

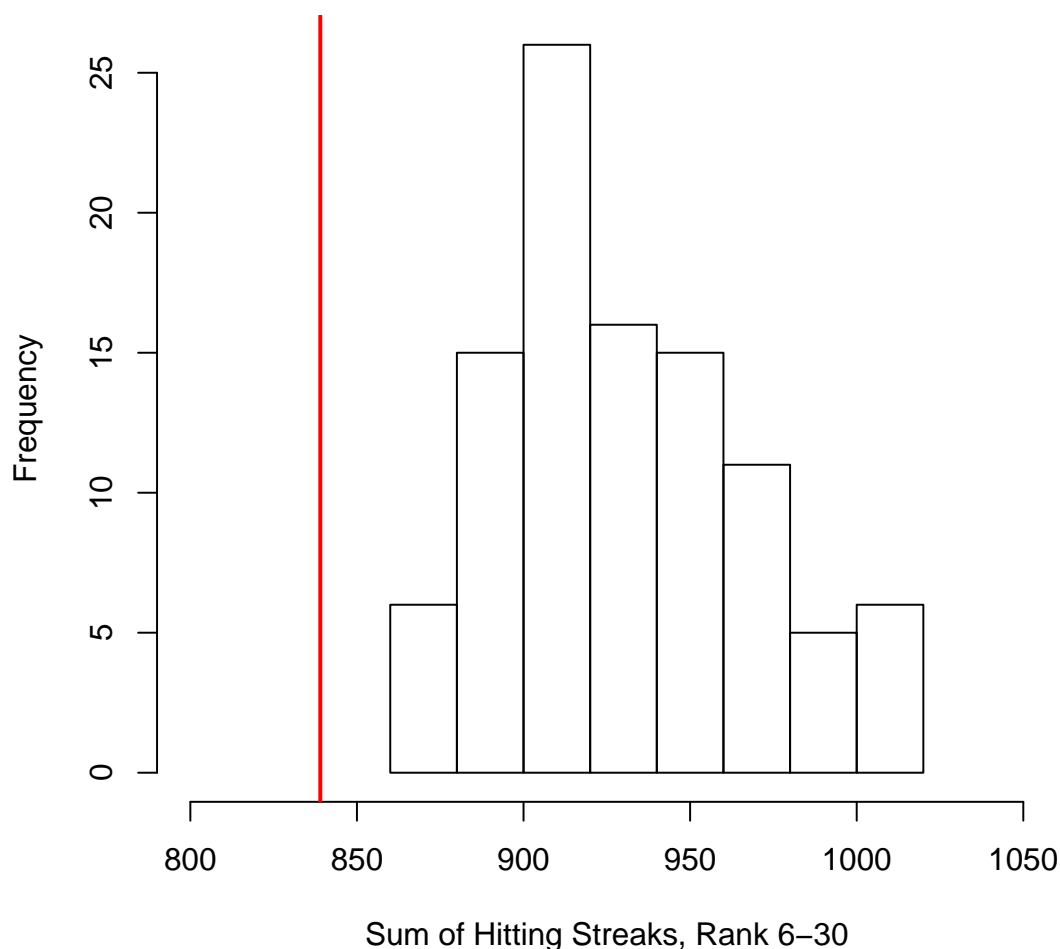


Figure 2: After simulating 139 seasons of baseball 100 times over using adjusted hitting averages and simulated numbers of plate appearances, a test statistic is calculated: the sum of the sixth-through thirtieth-longest hitting streaks, equal to 839. If the model is valid, and streaks of these lengths are considered likely events, then the model should produce a statistic that is comparable to reality. Even after smoothing for seasons that are outliers for each player, simulations (in the histogram) produce streaks that in aggregate are always longer than in reality (the solid vertical line) for hitting streaks.

4.1 Simulations 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), 500 simulations of each scenario are performed.

The results of these simulations are displayed in Figure 3. For each trial I note the sixth through thirtieth top streaks overall; 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 bulk of the longest simulated streaks is located in the pre-1940 period.

This suggests that the results would be improved by fitting several epochs independently. First I consider the use of only two epochs, 1971-1939 and 1871-1939 and 1950-2009 respectively, and in this case fit the top fifteen streaks in each era respectively, withholding the 1940s and the DiMaggio streak from fitting. The value of heterogeneity of $\sigma_H = 0.06$ 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, from 1871-1900 (co-inciding with the introduction of the American League, the World Series and the beginning of baseball's modern era) and 1901-1939, shows strong differences between these two halves. 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](#)].

4.2 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 for pre-1900 baseball, and the fairly-heterogeneous model for the beginning of the modern era, to the relatively uniform model 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

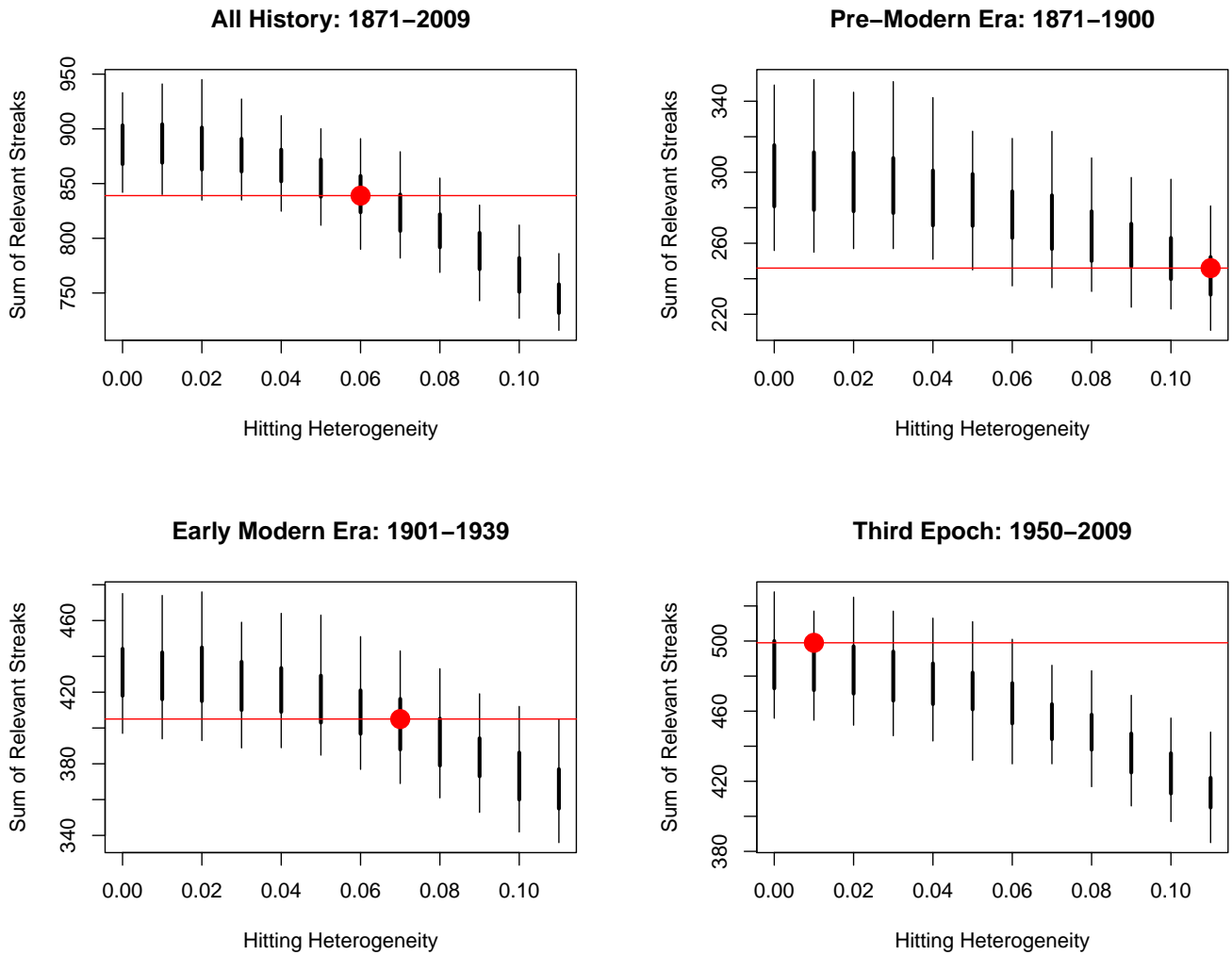


Figure 3: The distribution of the sums of the relevant lower-order streaks over 100 simulations, as a function of the additional heterogeneity σ_H . Thick bars represent the interquartile range, thin bars represent the central 95% confidence interval. The horizontal line represents the observed historical sum of streaks for each era under consideration. A red 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.

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 the 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% (76 times in 500); considering only the modern era, that probability drops to 4.8% (24 times in 500). 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 in 12.6% of cases. This suggests that the DiMaggio streak is certainly exceptional under the assumption that streak lengths in lower positions on the all-time list are to be expected, though it does not indicate a level of exceptionality that would be explained by DiMaggio having an exceptionally streaky hitting nature.

The sharp discrepancy in heterogeneity 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 for this early time period. 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. The next step would be to incorporate a larger database of streak information from this time, when and if it should become available.

5 An Analysis of On-Base Streaks is Less Conclusive

Being satisfied that the model produces valid results for much of the hitting streak behaviour in history, I now move to on-base streaks, which have received considerably less attention in both the public and academic presses, including the record-holding 84-game on-base streak of Ted Williams in 1949. 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 (1901-1939), 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 on walks 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

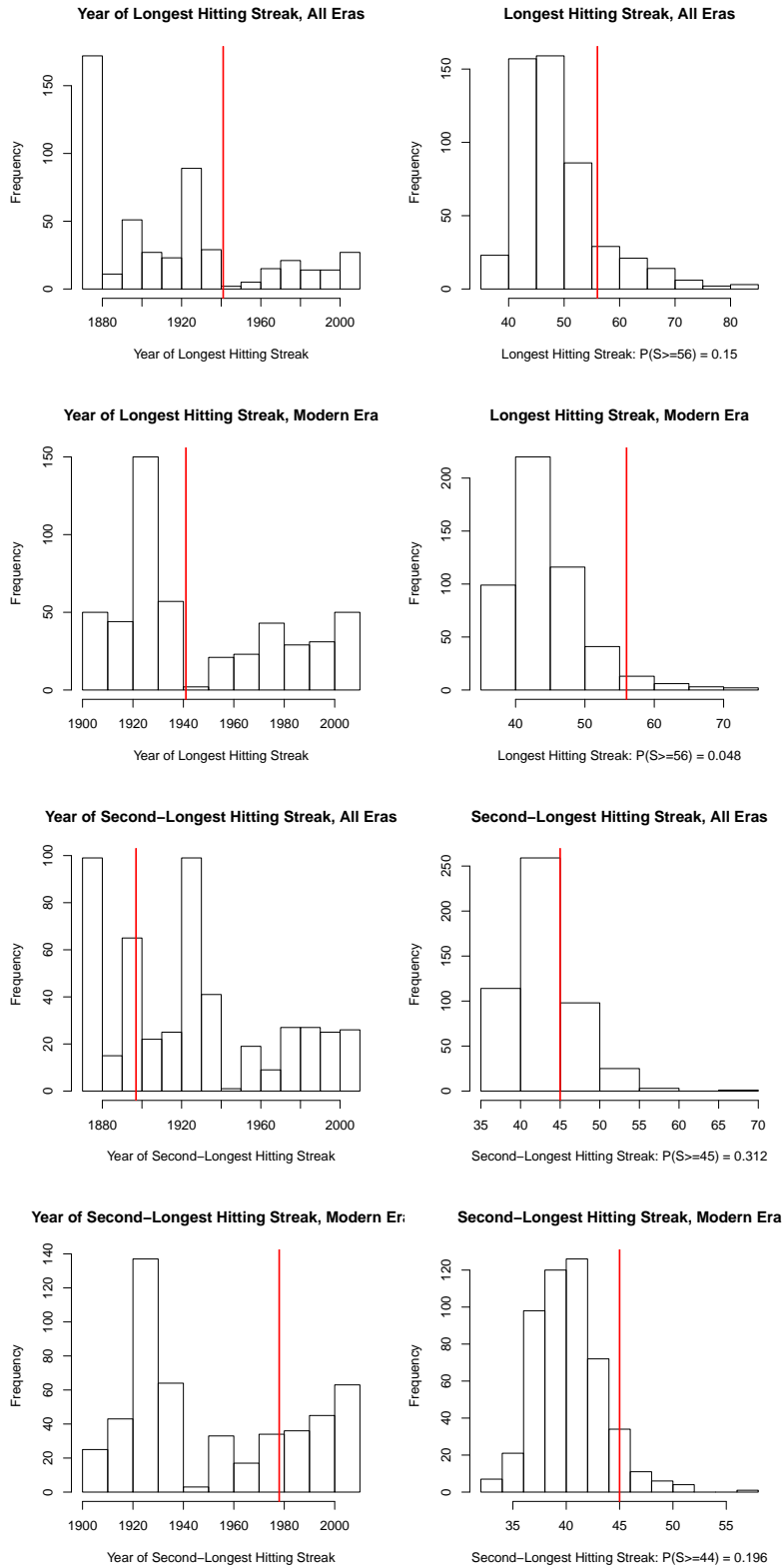


Figure 4: The behaviour of the longest and second-longest hitting streaks in simulations (histograms) and reality (red vertical lines). Streaks across all of baseball history (1871-2009) and the modern era (1901-2009) are each considered. The spikes in the 1870s, 1890s and 1920s are due largely to a small number of players.

Lower Order Streaks with No Heterogeneity

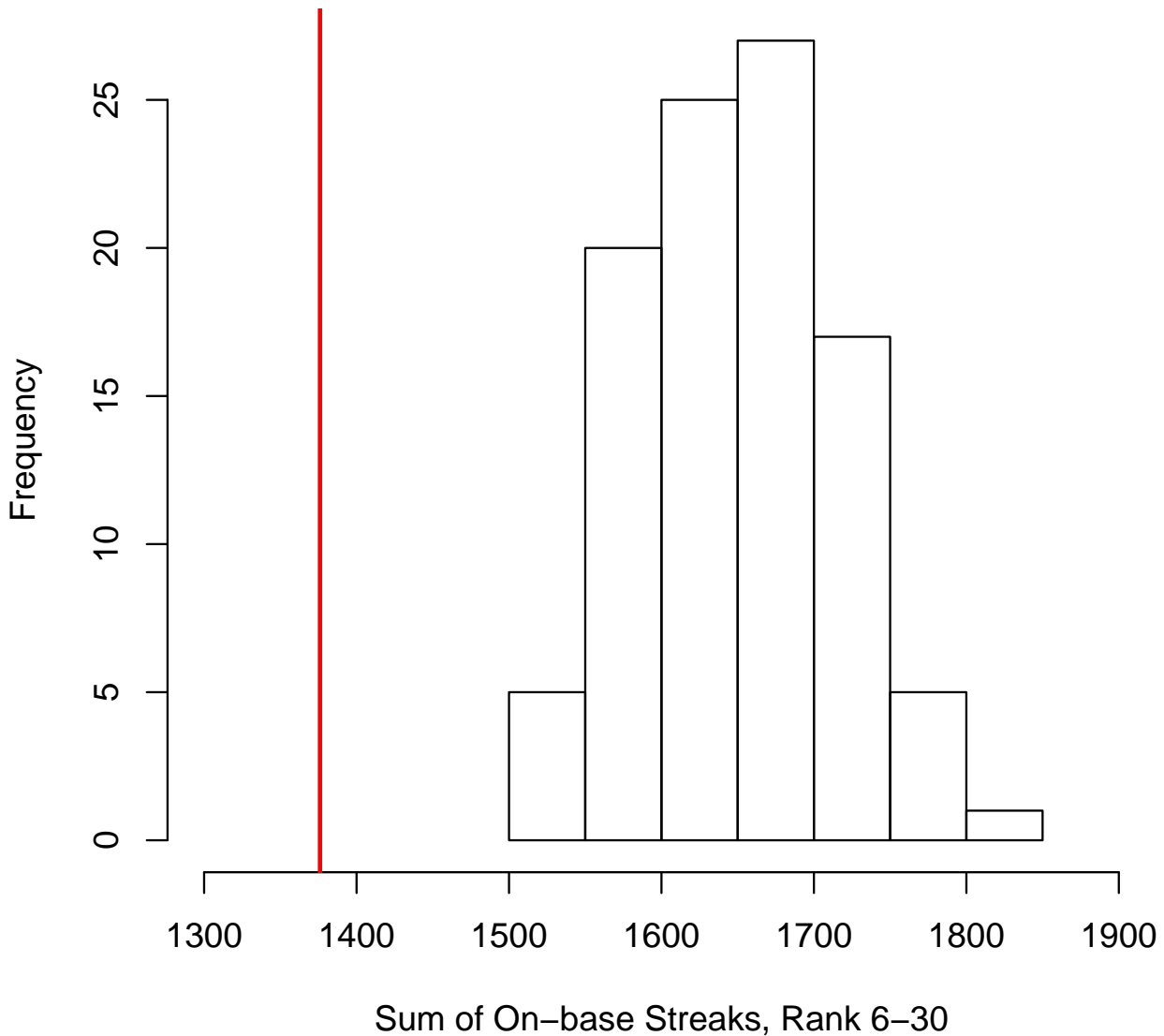


Figure 5: After simulating 139 seasons of baseball 100 times over without adding additional game-to-game heterogeneity in the ability to achieve a base on balls (while using the added heterogeneity in hitting ability for each era, as in the previous section), a test statistic is calculated: the sum of the sixth- through thirtieth-longest on-base streaks between 1871 and 2009, a total of 1387 streak-games. 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.

lesser streaks are to be expected, the model needs additional heterogeneity.

Figure 6 shows the behaviour of on-base streaks for a variety of heterogeneities, 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 test 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.

5.1 Adding Streak-Inducing or Streak-Damping to Simulations

To include a game-to-game dependence in the likelihood of continuing a streak, a Markov-type component can be included in the model, so that the probability of reaching base within one game is affected by whether a similar event was observed in the previous game. In particular, let μ_B represent the change in the walking average following a game in which the player reached base safely, and $-\mu_B$ be the change in walking average otherwise. For positive values, this will increase the expected length of both on-base streaks and slumps [Stern and Morris, 1993].⁷

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* 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 during 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. I must therefore conclude that the factors that produce on-base streaks are beyond the scope of this modelling approach.

⁷This is not the exact form of the streak adjustment originally proposed by Stern and Morris [1993], in which the bonus was implemented on an at-bat basis rather than by game-by-game.

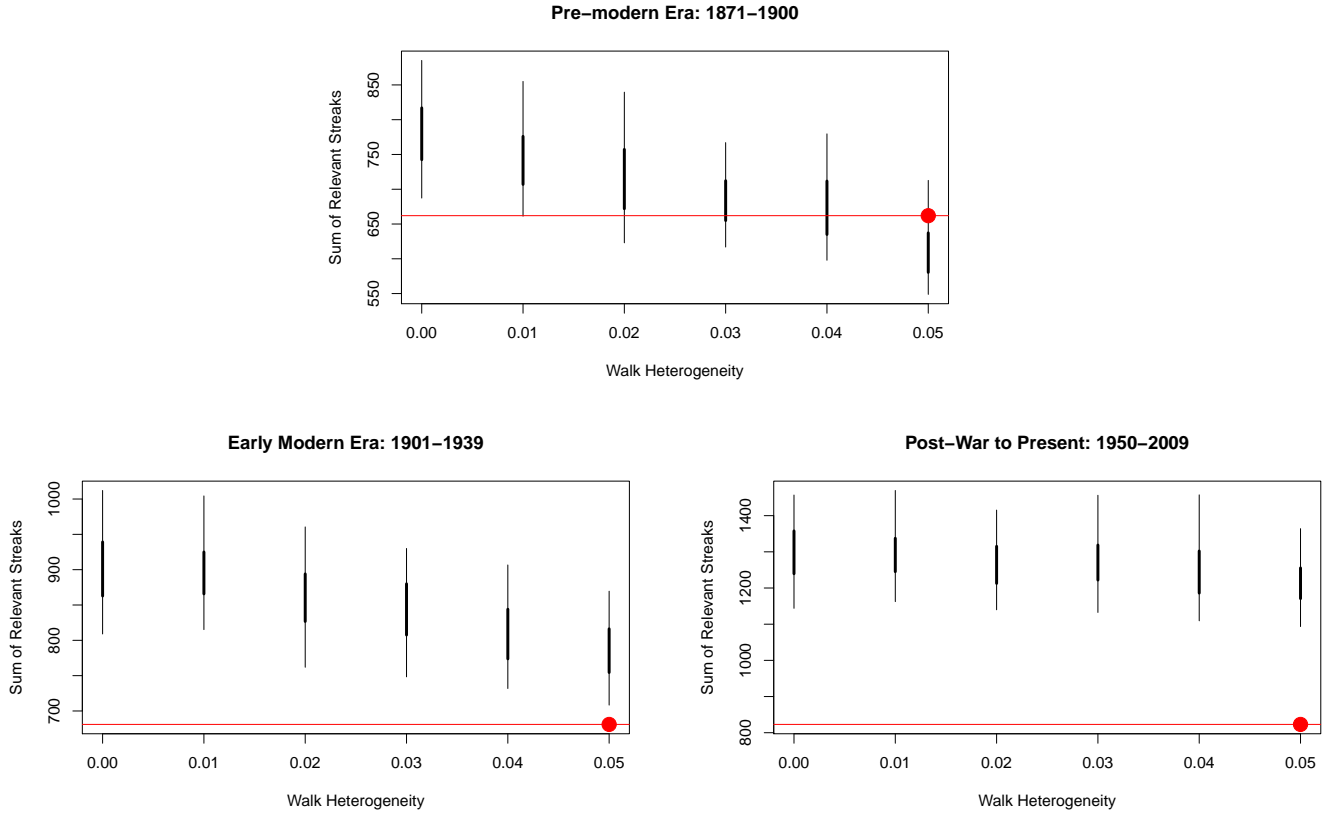


Figure 6: The distribution of the sums of the relevant lower-order streaks, as a function of the supplemental heterogeneity of walk σ_B ; thick bars represent the interquartile range, thin bars represent the central 95% confidence interval. The horizontal line represents the observed historical sum of lower-order streaks for each era. 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 σ_H in each case corresponds to the optimal values found in Section 4.1. While a reasonable value of on-base heterogeneity in the pre-modern era produces a model whose streak statistics are comparable to reality, no reasonable value has the same effect for the modern era.

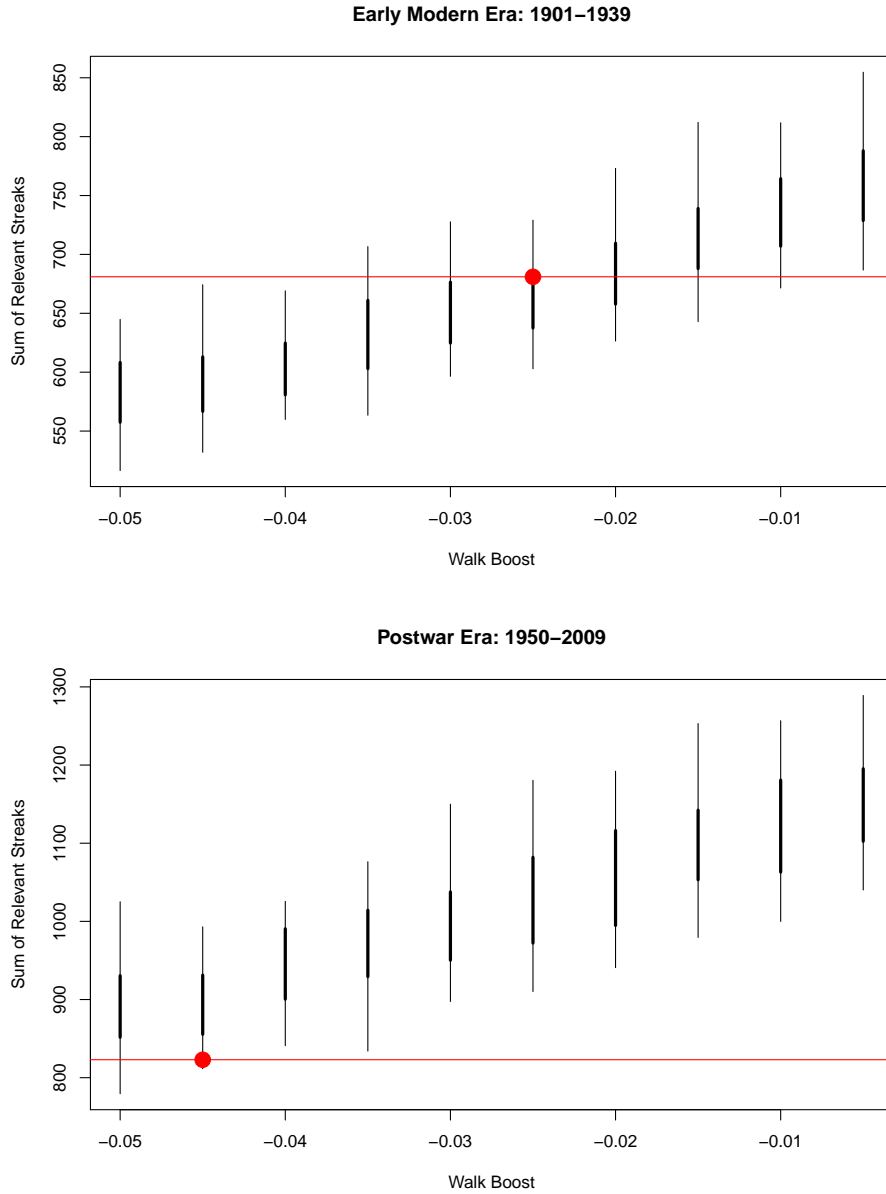


Figure 7: The mean sums of squared errors for on-base streaks as a function of the added streak bonus term for walks, in the early modern (1901-1939) and present-day (1950-2009) eras. Hitting heterogeneity in each case corresponds to the optimal values found in Section 4.1, and walk heterogeneity equals 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.

5.2 Possible Factors to Improve the Modelling of 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.⁸ 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 time⁹. 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.

6 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.

6.1 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

⁸Source: <http://www.baseball-almanac.com/hitting/hiobp3.shtml>.

⁹This streak was ended with a game in which Bonds failed to reach base in four plate appearance.

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.

6.2 The Role of Permutations

The permutation approach considers all factors except for game-to-game dependence that would contribute to streakiness (both in the sense of “hot hand” and nonstationarity), whereas the Bernoulli-type simulation approach begins with each factor to be separately included. As the simulation methods have suggested that a small degree of heterogeneity is necessary with no dependence between games, and permutations have suggested that there is a positive streak bonus, it is likely that the degree of heterogeneity is understated by the simulation method, and that the combination of a small positive bonus and a larger heterogeneity is most likely to be the correct balance. However, this combination is not identifiable with currently available data, especially for streaks before 1950. A permutation approach on game data from this era, when available, may serve to improve the estimates for total heterogeneity as well as isolate the magnitude of any streak effects.

6.3 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 (as in the aforementioned example with Ted Williams. 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.

6.4 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.

6.5 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.

B Software

The R code and data for the procedures I have given is included as a supplement to this manuscript. For most analyses, 100 replications are sufficient to check whether the model is producing an accurate representation of historical streak behaviour.

References

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

Hitting Streak			On-Base Streak		
Player	Year	G	Player	Year	G
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
Gene DeMontreville	1897	36	Ed Delahanty	1896	56
Fred Clarke	1895	35	Bill Joyce	1896	56
Ty Cobb	1917	35	Arky Vaughan	1936	56
George Sisler	1925	34	Billy Hamilton	1896	55
George McQuinn	1938	34	Ty Cobb	1915	55
George Davis	1893	33	Bill Joyce	1894	54
Hal Chase	1907	33	Ray Blades	1925	54
Rogers Hornsby	1922	33	Luke Appling	1936	53
Heinie Manush	1933	33	Danny Lyons	1887	52
Ed Delahanty	1899	31	Ty Cobb	1914	52
Nap Lajoie	1906	31	Tris Speaker	1920	52
Pete Rose	1978	44	Orlando Cabrera	2006	63
Paul Molitor	1987	39	Duke Snider	1954	58
Jimmy Rollins	2006	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 1: The top 15 hitting and on-base streaks for play before 1940 (top) and after 1949 (bottom). These were obtained in [Spatz \[2007\]](#).

- Albert, J. (1999): “Bridging Different Eras in Sports: Comment,” *Journal of the American Statistical Association*, 94, 677–680.
- Albert, J. (2008a): “Great Streaks,” *By the Numbers*, 18.
- Albert, J. (2008b): “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.
- 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, 1131–1152.
- Carruth, M. and S. Jensen (2007): “Evaluating Throwing Ability in Baseball,” *Journal of Quantitative Analysis in Sports*, 3, 2.
- 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.
- Jensen, S., K. Shirley, and A. Wyner (2009): “Bayesball: A Bayesian hierarchical model for evaluating fielding in major league baseball,” *Annals of Applied Statistics*, 3, 491–520.
- 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.
- McCotter, T. (2008): “Hitting Streaks Dont Obey Your Rules: Evidence That Hitting Streaks Arent Just By-Products of Random Variation,” *The Baseball Research Journal*.
- 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.