Journal of Quantitative Analysis in Sports

Volume 4, *Issue* 1 2008 *Article* 3

Streaky Hitting in Baseball

Jim Albert*

*Bowling Green State University, albert@bgnet.bgsu.edu

Streaky Hitting in Baseball

Jim Albert

Abstract

The streaky hitting patterns of all regular baseball players during the 2005 season are explored. Patterns of hits/outs, home runs and strikeouts are considered using different measures of streakiness. An adjustment method is proposed that helps in understanding the size of a streakiness measure given the player's ability and number of hitting opportunities. An exchangeable model is used to estimate the hitting abilities of all players and this model is used to understand the pattern of streakiness of all players in the 2005 season. This exchangeable model that assumes that all players are consistent with constant probabilities of success appears to explain much of the observed streaky behavior. But there are some players that appear to exhibit more streakiness than one would predict from the model.

KEYWORDS: batting logs, consistent hitting, exchangeable model, Bayes factor

1 Introduction

Consider the batting logs for all players in the 2005 baseball season. We focus on the "regular" players who had at least 300 plate appearances during this season; this will exclude pitchers and part-time players who may have different batting tendencies from the regular players. By using different definitions of "batting success", we will focus on sequences of three types of hitting data.

- *Hitting data.* Here we focus only a player's official at-bats (excluding walks, hit by pitches, and sacrifice flies), and define a "success", coded by 1, if the player gets a hit, and 0 otherwise.
- *Strikeout data.* Again we consider only official at-bats and a player is "successful" (coded by 1), if he strikes out; if he doesn't strike out, the at-bat is coded as 0.
- *Home run data.* Consider only the at-bats where the batter puts the ball in-play; these will be the at-bats that are not strikeouts. Then a batter is successful (coded by 1) if he hits a home run, and 0 otherwise.

For a particular form of hitting data, one will obtain binary sequences of 0's and 1's for all regular players in a season. If one scans these data, one will find interesting patterns that suggest that particular players are streaky. One may find

- periods during the season where a player is very successful
- periods during the season where a player has limited success
- streaks or runs of batting success or no batting success for particular players

There is no doubt that these streaky patterns exist in baseball hitting data and there is much discussion in the media about these streaky patterns. The interesting question, from a statistical perspective, is: What do these streaky hitting patterns say about the streaky *abilities* of these players?

To discuss streaky ability, consider a simple probability model for the hitting for a single player. Suppose, for each at-bat, a hitter is successful with probability p, this probability remains the same value for all at-bats during the season, and outcomes for different at-bats are independent. We

will call this the *consistent p model* – this coin-tossing model represents the hitting for a player who is "truly consistent" during the season. A player with "streaky ability" deviates from the consistent p model. It is possible that the player's hitting probability p can vary across the season. Alternatively, the individual outcomes may be dependent and the probability of a hit in an at-bat may depend on the player's success or lack-of-success in previous at-bats.

2 Previous Work

There has been much interest in the detection of streaky ability since Gilovich et al (1985) and Tversky and Gilovich (1989) who claimed that any observed streakiness in sports data is simply people's misperception of the streaky patterns inherent in random data. For discussions on the statistical detection of streakiness, see Berry (1991), Larkey et al (1989), and Stern (1997). Alan Reifman has a web site (http://thehothand.blogspot.com/) devoted to hot-hand research. Bar-Eli et al (2006) give a survey of twenty years of hot-hand research.

Albright (1993) did an extensive analysis of streakiness of hitting data for many major league baseball players. By incorporating "streakiness" parameters in a regression model, he determined that these parameters were statistically significant for particular players. However, he failed to find convincing evidence for a general pattern of streakiness across players. Albert (1993), in his discussion of Albright's paper, introduced a streaky model for hitting performance. He assumed that a baseball hitter had two possible ability states, low and high, with corresponding hitting probabilities p_{LOW} and p_{HIGH}. A player would move between the two ability states according to a Markov Chain with a given transition matrix. One could then measure streaky ability for a given player by estimating the difference in hot and cold hitting probabilities $p_{HIGH} - p_{LOW}$. Once a streaky model has been defined, then one can investigate the power of different statistics, such as the longest streak of wins, in detecting values of parameters of the streaky model. Albert and Pepple (2001), using a Bayesian viewpoint, showed that a number of streaky statistics such as the longest streak and the total number of streaks were not very powerful in detecting true streakiness. Wardrop (1999) reaches similar conclusions from a frequentist perspective.

Albert (2004) looked at the streaky patterns of wins and losses for Major League baseball teams. He focused on Oakland's 20-game winning streak in

the 2002 season; by using a random effects model to model team competition, he concluded that one should expect to see streaks of similar lengths every 25 years of baseball. There appears to be more evidence for streaky ability at the individual sports level. Dorsey-Palmateer and Smith (2004) demonstrate that bowlers' patterns of strikes are not explainable by simple chance models, and Klaassen and Magnus (2001) demonstrate differences from the "independent and identically distributed" assumption for tennis point-by-point data.

3 Multiplicity

One general problem in the search for streaky ability is the issue of multiplicity. Since there are many peoples and batting opportunities for a single season, there are many opportunities for streaky behavior, and by focusing only on players who appear streaky, there is a selection bias.

To illustrate this multiplicity problem, consider the following simulation model for hit/out data (this model will be developed in a later section). Suppose we have 284 players where each player has a constant probability of success. (That is, each player is a consistent p hitter.) But the success probabilities among players vary – they come from a beta probability distribution with mean 0.274 and standard deviation 0.013. Suppose we simulate probabilities of success for the 284 players from this beta model, and then simulate independent sequences of successes and failures (much like flipping coins with varying probabilities of heads) for the 284 players using their actual 2005 number of at-bats. When we did this simulation once, we observed one player who had a hitless streak of 38 at-bats. We observed another player who got hits in nine consecutive at-bats. We also observe some players with some interesting short-term streaky behavior. To illustrate, Figure 1 plots the batting average of one player using a moving window of 30 at-bats over time. We see that during one 30 at-bat sequence during the season, the player had a batting average exceeding .500, and during another 30 at-bat period, his average was smaller than .100. Is this observed streaky behavior meaningful? No, this data were simulated assuming that each player was a consistent p hitter. We are observing this behavior since there we have data for many players and the extreme players in this group appear streaky due to multiplicity.

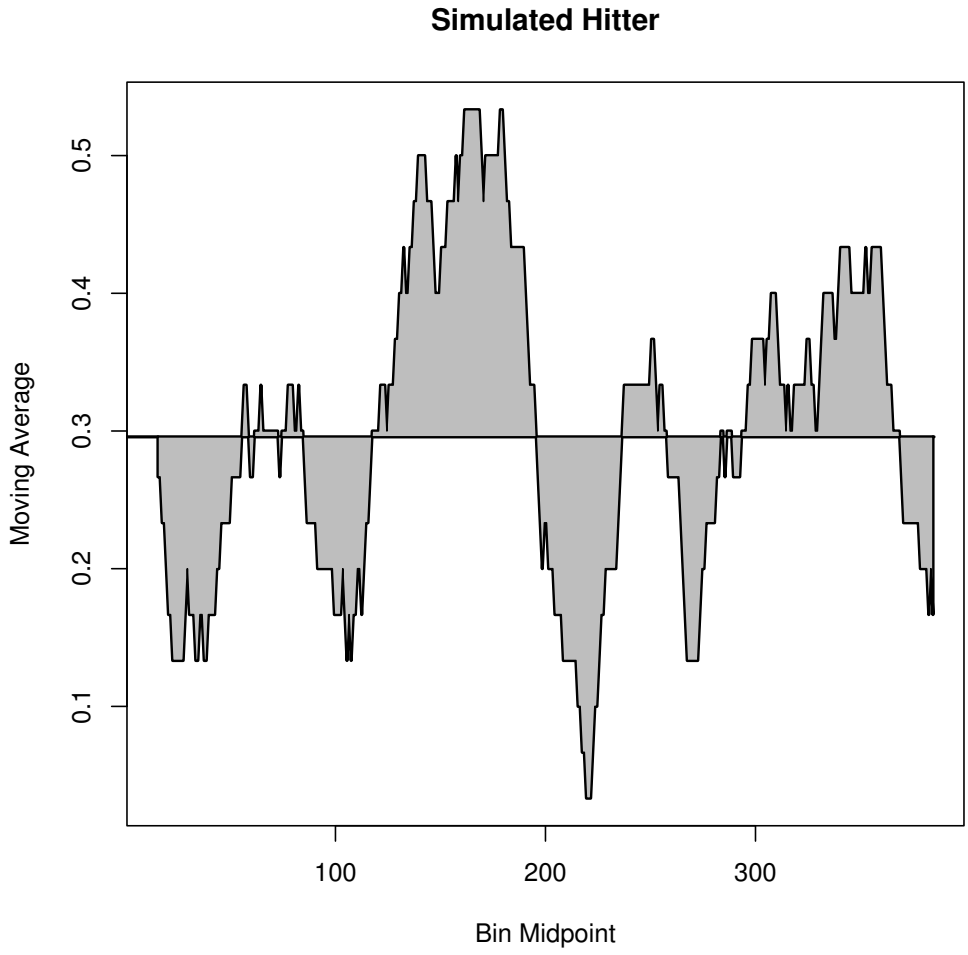

Figure 1: Moving average graph for hitting data for one "extreme" player simulated from a consistent p model.

4 Plan of the Paper

The intent of this paper is to investigate the streaky hitting patterns displayed by all regular players during a single baseball season. Specifically, we would like to address if a consistent-p model is suitable for describing the streaky behavior of this group of players. If it is not suitable, then are there general patterns of streakiness? Are there outliers, or players who display streakiness that is not consistent with the model?

We begin our study in Section 5 by considering the binary sequence of success/failure outcomes for a single player and describing a number of different statistics that can be used to measure patterns of streaky performance. Five of these streaky statistics are computed for all regular players in the 2005 season and Section 6 gives a description of these "streaky distributions." It can be difficult to understand the size of a particular streaky statistic, say the length of the longest run of successes, since it is confounded with the player's hitting ability and the number of hitting opportunities. Section 7 provides a simple adjustment procedure that looks at an individual's player's streaky statistic in the context of a hypothetical collection of players with the same ability and the same number of at-bats. We measure the extremeness of a player's streaky statistic by means of a p-value or the probability that the statistic from this hypothetical collection is at least as extreme as the player's statistic. By graphing these p-values for all players, we learn about the suitability of the consistent-p model across players. In Section 8, we improve this adjustment method by use of an exchangeable model to simultaneously estimate the hitting abilities of all players, and the posterior predictive distribution is used in Section 9 to assess the suitability of the model in predicting the observed streakiness among the players. In Section 10 we look at players that appear unusually streaky using our criteria, and Section 11 summarizes the main findings.

5 Streaky Statistics

We begin with a binary sequence of hitting outcomes for a player during a season. As an example, consider the batting performance of Carlos Guillen who appeared to be streaky in the 2005 season. Following is Guillen's sequence of batting outcomes for all at-bats for the 2005 season, where 1 and 0 correspond to a hit and out, respectively.

```
0 1 0 1 1 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 1 1 1 0 0 0 0 0 0
1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 1 0 1 0 0 1 1 0
1 0 1 1 0 1 0 1 1 0 0 0 0 0 1 1 1 1 0 0 1 0 1 0 0 1 1 0 0
0 1 0 1 0 0 0 1 1 1 0 1 1 1 1 0 0 1 1 1 1 0 0 1 0 0 1 0 1
0 0 0 1 0 1 0 0 0 0 0 1 1 1 0 0 1 0 0 0 0 0 0 1 1 1 0 1 0
0 0 0 1 1 1 1 0 1 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 1 1 0 1 0
0 0 0 0 0 0 0 0 1 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0
0 1 0 0 0 0 1 1 0 1 0 0 1 0 0 0 1 0 1 0 0 0 0 0 0 0 0 1 0
1 0 0 1 1 1 1 0 0 0 0 0 1 1 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 1 0 0
0 1 1 0 0 1 0 0 0 1 0 0 0 0 1 0 0 0 0 1 1 0 1 1 1 0 1 0 0
0 0 0 1 0 1 1 0 0 0 1 0 0 0 1
```

There are many ways of measuring patterns of streakiness in this binary sequence. One general way of detecting streakiness, described by Albright (1993) and Albert and Pepple (2001), is by the patterns of *runs* or consecutive streaks of the same outcome. If we look at the pattern of hits (1's) and outs (0's) in Guillen's first group of at-bats

```
0 1 0 1 1 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 1 1 1 0 0 0 0 0 0
```

we see that he began with a run of one out, a run of one hit, a run of one out, a run of two hits, a run of three outs, and so on. One can measure streakiness in a player's sequence for an entire season by using different run statistics:

- the length of the longest run of successes
- the length of the longest run of failures
- the mean length of the lengths of runs of success
- the mean length of the lengths of runs of failures
- the total number of runs in a sequence

For Guillen's data, we compute

- the length of the longest run of outs is 19
- the length of the longest run of hits is 4

- the average length of runs of outs is 3.24
- the average length of runs of hits is 1.53
- the total number of runs is 140.

A player who is streaky may exhibit an unusually long run of successes or a long run of failures. He may tend to have long runs of failures and so the mean length of failure runs would be large. Also if the player tends to have long runs of successes or failures, then the total number of runs in the sequence would be small.

A second way to quantify streakiness in a binary sequence is based on moving averages. Suppose one chooses a span of w at-bats and computes the set of moving averages $\{m_j, j = 1, ..., n - w + 1\}$ where m_j is the proportion of hits in the at-bats from j to $j + w - 1$:

$$m_j = \frac{\sum_{i=j}^{j+w-1} y_i}{w}.$$

By plotting the moving averages $\{m_j\}$ across time, one sees the volatility of the player's success rates in short time intervals. (This approach is described in Albert and Pepple (2001) and Chapter 5 of Albert and Bennett (2003).) Figure 2 shows the moving averages of Guillen's hit/out sequence using a window of 30 at-bats, corresponding to about a week of games. The horizontal line in the figure corresponds to Guillen's .320 batting average for the entire season. Looking at this figure, we see that Guillen had a mild hitting slump followed by a long hot hitting period that peaked at 100 at-bats, and two hot and two cold spells towards the end of the season. One can measure streakiness in a moving average plot by

- the range of the moving averages $R = \max_j m_j - \min_j m_j$
- the mean variation of the moving averages about the season average

$$B = \frac{1}{n - w + 1} \sum_{j=1}^{n-w+1} |m_j - \bar{y}|, \text{ where } \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

Streaky players will tend to have large values of R and B. For Guillen's data, one can compute

- the range $R = 0.567 - 0.067 = 0.5$

Table 1: Table of previous hitting outcome and present outcome for the 2005 Carlos Guillen.

	Current At-bat	
Previous At-bat	Out	Hit
Out	157	70
Hit	69	37

- the mean variation $B = 0.0977$

We will refer to B as the "black" statistic since it is proportional to the shaded area in the moving average plot in Figure 2.

A third way to quantify streakiness in the sequence is to look the relationship of the hitter's success with the success in the previous at-bats. (This approach is described in the basketball context by Wardrop (1995).) Suppose we divide the data into bins of length w and categorize the success in the ith and $(i+1)$th periods for all i. For example, if $w = 1$, we look at the relationship between the present outcome with the preceding outcome. For Carlos Guillen, we obtain Table 1.

Looking at the table, we see that there were $157+70 = 227$ at-bats where the previous outcome was an out; of these outcomes the proportion of hits in the next at-bat was $70/227 = .308$. In contrast, of the $69+37 = 106$ at-bats where he had a hit, the proportion of hits in the next at-bat was $37/106 = .349$. In this case, there is some evidence that Guillen performs better after a success than after a failure.

Instead of looking at periods of one at-bat, we can divide Guillen's 334 at-bats into 167 periods of two at-bats, and categorize the number of hits in the ith period with the $(i+1)$th period, obtaining Table 2. Looking at this table, one can compute the mean number of hits after 0, 1, and 2 hits in the previous period, as displayed in Table 3. From Table 2, we see that the mean number of hits is essentially the same after 0 and 1 hits in the previous at-bats. The number of hits after 2 hits is increased, but it is not significantly larger due to the small sample size.

From these two-way tables, one can measure dependence of the previous success and the current success by the Pearson chi-square statistic and the corresponding p-value of the test of independence. Large chi-square statistics

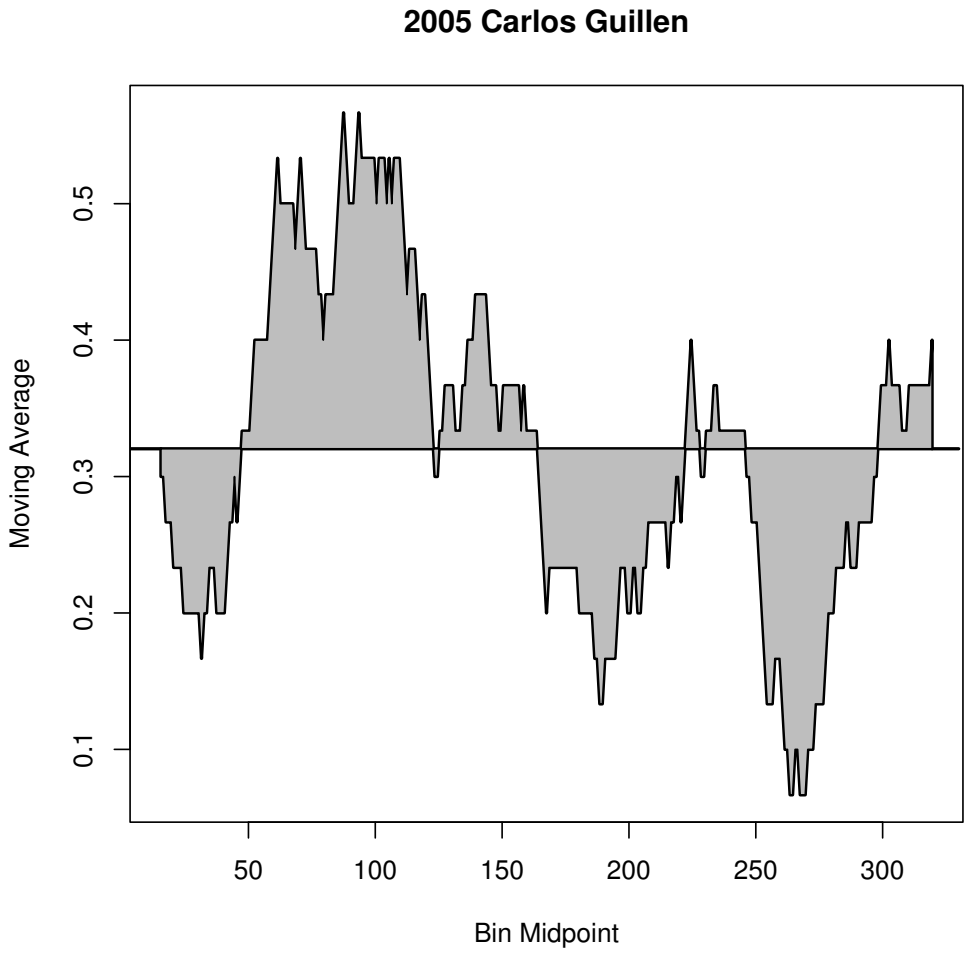

Figure 2: Moving average graph for Carlos Guillen.

Table 2: Table of hits in previous period and current period of two at-bats for the 2005 Carlos Guillen.

	# hits in current period		
# hits in previous period	0	1	2
0	39	32	7
1	33	30	7
2	6	8	4

Table 3: Mean number of hits in the current period after 0, 1, 2 hits in the previous period for the 2005 Carlos Guillen.

# hits in previous period	Mean # hits in current period
0	0.59
1	0.63
2	0.89

or small p-values suggests some relationship between a player's performance in the previous and current periods.

A fourth way of measuring streakiness is based on a more formal testing procedure based on the introduction of alternative streaky models. To define these models, suppose we first group the player's hitting data into bins of a given size – here we use bins of 20 at-bats, corresponding to a period of 4 games. We then have the grouped hitting data $x_1, ..., x_n$, where x_i is the number of successes in the ith period. Suppose that x_i is distributed binomial with probability of success p_i. The consistent p model, denoted by M_C, states that the player's hitting success is a constant value p over the entire season.

$$M_C : p_1 = ... = p_n = p.$$

To complete this model, we assume that the constant value p has the noninformative uniform prior. The streaky model, denoted by M_S, says that the player's hitting probabilities over the season $\{p_i\}$ vary according to a beta density of the form

$$g(p) = \frac{1}{B(K\eta, K(1-\eta))} p^{K\eta-1}(1-p)^{K(1-\eta)-1},\ 0 < p < 1.$$

In this beta density, the parameter η is the mean and K is a precision parameter. We fix the parameter K to a particular value, and assign the parameter η a uniform noninformative prior. As the precision K approaches infinity, the streaky model M_S approaches the consistent model M_C. So the parameter K can be viewed as a measure of streakiness where smaller values indicate a higher level of streakiness. The streaky statistic is defined to the Bayes factor in support of the streaky model M_S over the consistent model M_C. This Bayes factor, denoted by BF, is given by

$$BF = \frac{m(x|M_S)}{m(x|M_C)},$$

where $m(x|M)$ is the predictive density of the data x given the model M. (See Kass and Raftery (1995) for a general discussion of Bayes factors, and Kass and Vaidyanathan (1992) and Raftery (1996) for illustrations of Bayes factors for exponential family models. Albert (2007) uses this method for detecting streakiness for Derek Jeter's hitting data for the 2004 season.)

In this setting, we fix a value of the precision parameter $K = 100$ to represent "true" streakiness or variation in the hitting probabilities, and use

the corresponding Bayes factor BF to measure streakiness in the data. A large value of the Bayes factor indicates support for true streakiness. For Carlos Guillen, we observe the hit counts and number of at-bats data

```
5/20  5/20  7/20 10/20 10/20 10/20  6/20  9/20
4/20  4/20  6/20  7/20  4/20  2/20  6/20 12/34
```

For Guillen, we compute $BF = 1.32$, indicating a modest level of support for the streaky model M_S. Instead of computing the Bayes factor, an alternative method for getting a streaky statistic is to fit the streaky model M_S with unknown K to Guillen's data and estimate the value of the precision parameter K. The estimate of $\log K$ by fitting the model M_S is $\log \hat{K} = 3.30$. This estimate of K is a measure of the variation in hitting probabilities $p_1, ..., p_n$.

In this paper we focus on five measures of streakiness: (1) the mean length of runs of failures, (2) the longest run of failures, (3) the longest run of successes, (4) the black statistic B (using a window of 30 at-bats), and (5) the log Bayes factor $\log BF$using bins of 20 at-bats. We include the longest runs of successes and failures since these statistics receive much attention in the media. The mean length of runs of failures may provide more information than the longest run since it includes the lengths of all runs of failures during the season. The black statistic B seems to be a useful measure of the volatility of a player's success rate over time. The two Bayesian measures $\log BF$ and $\log \hat{K}$ are highly correlated, so we include only one measure here. In our preliminary work, the measures of streakiness based on the two-way contingency tables didn't seem very powerful in detecting streaky ability and so we exclude them here.

6 Streakiness in the 2005 Season

To gain an initial understanding about the streaky behavior of all regular players for the 2005 season, we compute the five measures for the hit/out data for all 287 players. Figure 3 presents histograms of the five statistics for all regular players. In each figure, the value for Carlos Guillen is indicated by a vertical line. Here are some general comments about the pattern of these streaky statistics across players.

1. The **average length of a run of outs** tends to be about 3.7 at-bats. There is one unusual streaky player who had an average run length

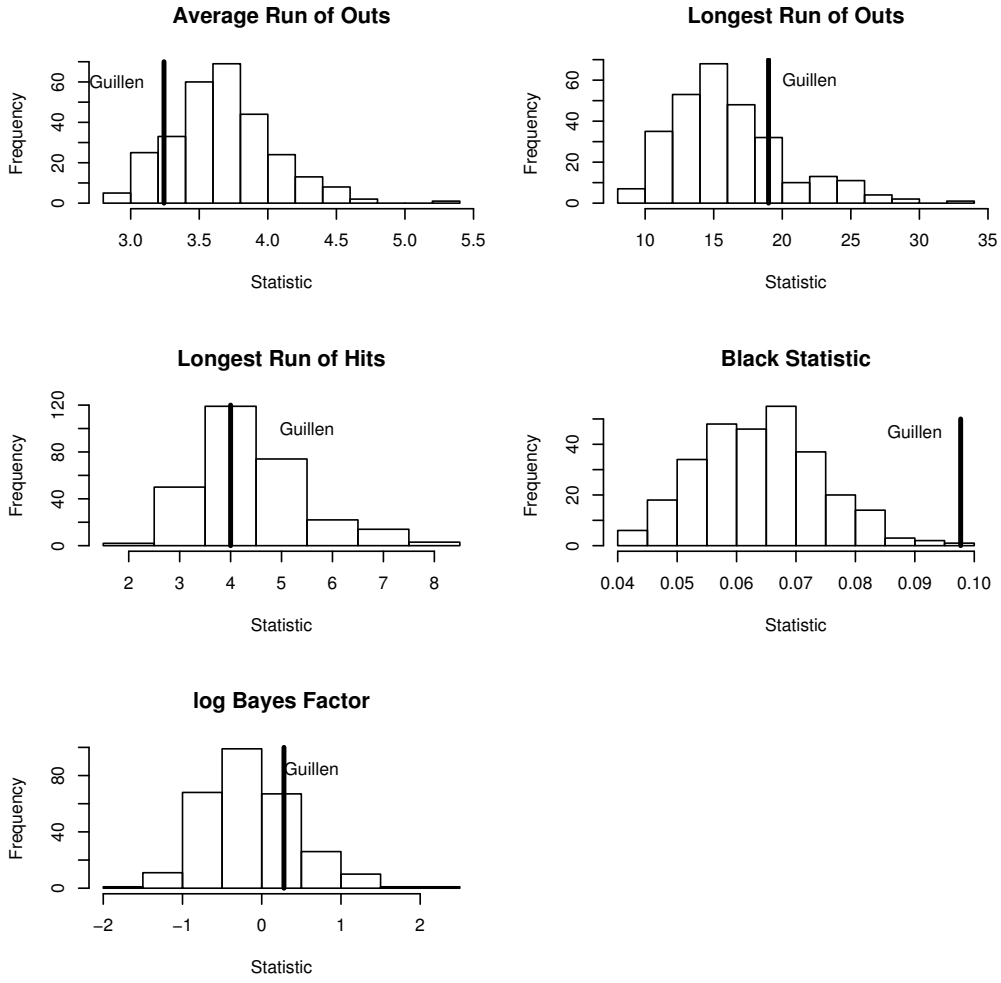

Figure 3: Histograms of five streaky statistics of hit/out data for all 2005 regular players.

of 5.3 and the most consistent player had an average run length of 3. With regards to this statistic, Guillen was relatively consistent.

2. The **longest run of outs** has a relatively long right tail, suggesting that there are some streaky players with respect to this statistic. One player had a long hitless streak of 34 at-bats. Guillen's longest hitless streak of 19 is relatively high among the regular players.

3. The **longest run of hits** statistic is concentrated on the five values 3, 4, 5, 6, and 7, and values of 2 and 8 were unusual. Guillen's longest streak of 4 hits is right in the middle of this distribution.

4. The **black statistic** is pretty symmetric about the value 0.065. Guillen's value is in the far right tail of this distribution.

5. The **log Bayes factor** statistics $\{\log BF\}$ are symmetric about the mean value of -0.13. Recall that values larger than zero indicate some support for streakiness, and values smaller than zero indicate support for consistency. The values range from -1.5 and 2.1 and Guillen's value of 0.28 indicates modest support for streakiness.

Table 4 presents a correlation matrix of these statistics. It is interesting that the measures are not strongly correlated even though all statistics are measures of streakiness. There is a strong correlation of 0.73 between the black statistic and the log Bayes factor. There is a moderately strong correlation (0.34) between the average run length of outs and the longest run of outs. Also there are moderately strong relationships between the black statistic and the lengths of the longest runs of outs and hits. There are some negative correlations. Players who have long streaks of hits tend to have small average run lengths of outs.

7 Adjustment

The histograms presented in Figure 3 are a first step in understanding the streaky patterns across players. For example, if a player gets five consecutive hits, we can say that this is not surprising since many players accomplished this feat during a season. But there are two confounding variables that make it difficult to interpret the size of these statistics. First, the observed streaky statistic value is dependent on the hitting ability of a player. For example,

Table 4: Correlation matrix of five streakiness statistics

	mean outs	long out	long hit	black	log Bayes factor
mean outs	1.00	0.34	−0.17	−0.04	0.14
long out	0.34	1.00	−0.18	0.25	0.31
long hit	−0.17	−0.08	1.00	0.24	0.17
black	−0.04	0.25	0.24	1.00	0.73
log Bayes factor	0.14	0.31	0.17	0.73	1.00

if one considers the longest run of outs, this statistic will tend to be larger for a weak batter with a small batting average. Second, the value of these statistics is dependent on the number of opportunities. A lead-off hitter with many at-bats during a season will have a better chance of having a long hitting streak or a long slump of failures.

To understand the size of a particular statistic for a player, s_{obs}, it is important to adjust this statistic for the player's hitting ability and his number of at-bats. We can perform this adjustment by considering the value s_{obs} among a hypothetical population of consistent hitters who have the same ability (as measured by the player's proportion of successes) and the same number of at-bats. A simulation adjustment method proceeds as follows.

1. Estimate the success probability p for the player by the proportion of successes $\hat{p}$.

2. Given the values of n and $\hat{p}$, simulate m sequences of binary data of length n, where the probability of successes is equal to $\hat{p}$.

3. For each sequence, compute the value of the streaky statistic, obtaining a collection of statistics $\{s_j\}$.

4. Compute the empirical p-value

$$P = \frac{1}{m} \sum_{j=1}^{m} I(s_j \geq s_{obs}),$$

where $I(A)$ is the indicator function that is equal to one if A is true and equal to 0 otherwise. This simulation process is equivalent to the

use of the nonparametric bootstrap procedure to estimate a sampling distribution as described in Efron and Tibshirani (1994).

The p-value P measures the extremeness of the player's streaky statistic in the context of all "consistent p" players having the same number of at-bats and same probability of success p. A p-value close to zero would suggest that this player is unusually streaky and a p-value close to one suggests that this player exhibits consistency beyond what would be predicted using the consistent p model.

We demonstrate these p-value calculations for Carlos Guillen. In the 2005 season, Guillen had 107 at-bats in 334 at-bats for a season batting average of $107/334 = .320$. Suppose we have a large group of hitters, each with a hitting probability of $p = .320$, and each has a season of $n = 334$ at-bats. We simulate a sequence of binary hit/out data for each player and compute the value of the streaky statistic. We collect the streaky statistics for 1000 players of this type.

Figure 4 illustrates these calculations. The first histogram in the top left portion of the figure shows the average run of outs for our 1000 hypothetical players who came to bat 334 times with a hitting probability of $p = .320$. Note that this histogram is different in location and shape from the histogram of the average run of outs of all 2005 players. The average run of outs tends to be smaller for our 1000 players since they came to bat only 334 times, in contrast to many of the regular players who had 500-700 at-bats. The vertical lines in the figure correspond to the values of the streaky statistic for Guillen. In 2005, Guillen's average runs of outs was 3.24. The p-value is the proportion of simulated players who had an average run of outs at least as large as 3.24. The p-value is computed to be 0.316 which is not sufficient evidence that Guillen was different from these consistent hitters. This same type of calculation was performed for each of the remaining four statistics. Note that Guillen was a bit unusual with respect to his long run of 19 outs and his black statistic value of 0.0977 – the corresponding p-values are 0.064 and 0.009, respectively. It is important to emphasize that these p-values cannot be computed from the histograms of all 2005 regulars (Figure 3), since the values of these streaky statistics are confounded with the number of attempts (at-bats) and the different probabilities of success.

These empirical p-values were computed for all regular players in the 2005 season using all five of our streaky statistics. Histograms of p-values for the statistics are presented in Figure 5. If the data were generated from

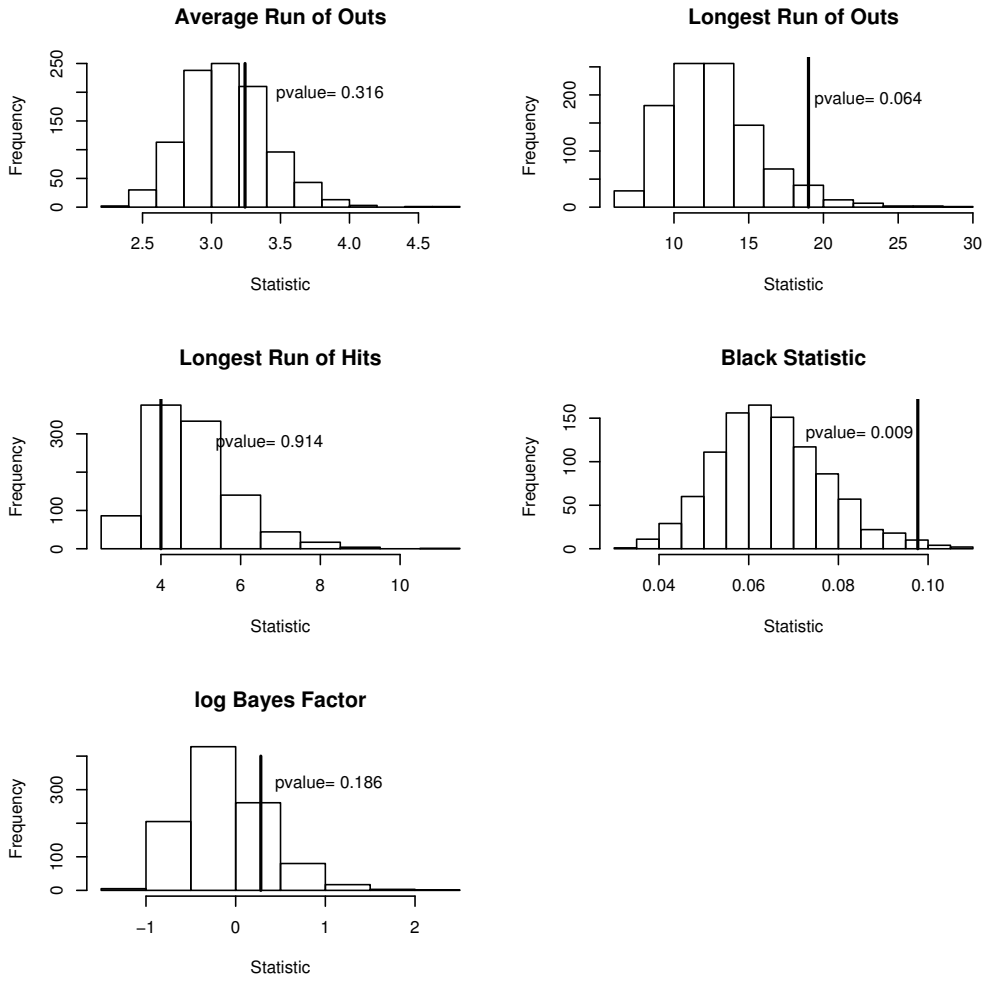

Figure 4: Illustration of p-value calculations for Ozzie Guillen for hit/out data.

consistent models with known probabilities, then one would see a uniform pattern in these p-values. This means, that one would observe that, say 10%, of the p-values would be smaller than .1, since this is the percentage that would be predicted from a uniform curve. However, our p-values are computed using estimated values for the probabilities and the patterns of these p-values show some interesting non-uniform shapes. (See Bayarri and Berger (2000) and Robins et al (2000) for discussion on the distribution of p-values for composite hypotheses.) Let's comment about each histogram in turn.

1. The p-values of the **average lengths of the runs of outs** have a bell-shaped appearance. This means that the players' average runs of outs tended to cluster more towards the mean assuming consistent p models and there were few players that exhibited streakiness using this measure. The average length of runs of outs does not appear to be a very helpful statistic for detecting streakiness since few players were extreme with respect to this measure.

2. In contrast, the p-values of the **longest run of outs** is more uniform in appearance. There is a cluster of values close to zero suggesting that there are more streaky players that expected using this statistic.

3. The p-values for the **longest run of hits** has a spike close to one. This is likely a consequence of the discreteness of this particular test statistic. For many players, the observed longest run of hits was a likely value and the probability of this value would inflate the p-value. We saw this phenomenon in the calculation of the p-value for Carlos Guillen in Figure 4.

4. The histogram of the p-values for the **black** and **Bayes factor** statistics resemble the histogram for the longest run of outs statistic. In both cases we see uniform shapes with a cluster of p-values close to zero.

It is difficult to judge the suitability of the consistent p models due to the nonuniform shapes of the p-values. It is hard to say if the nonuniform shape is due to some "true" streakiness or due to the inherent nature of these p-values using estimated probability values. In Section 9, we will compare the distributions of p-values with simulated distributions using "improved" estimates of the success probabilities.

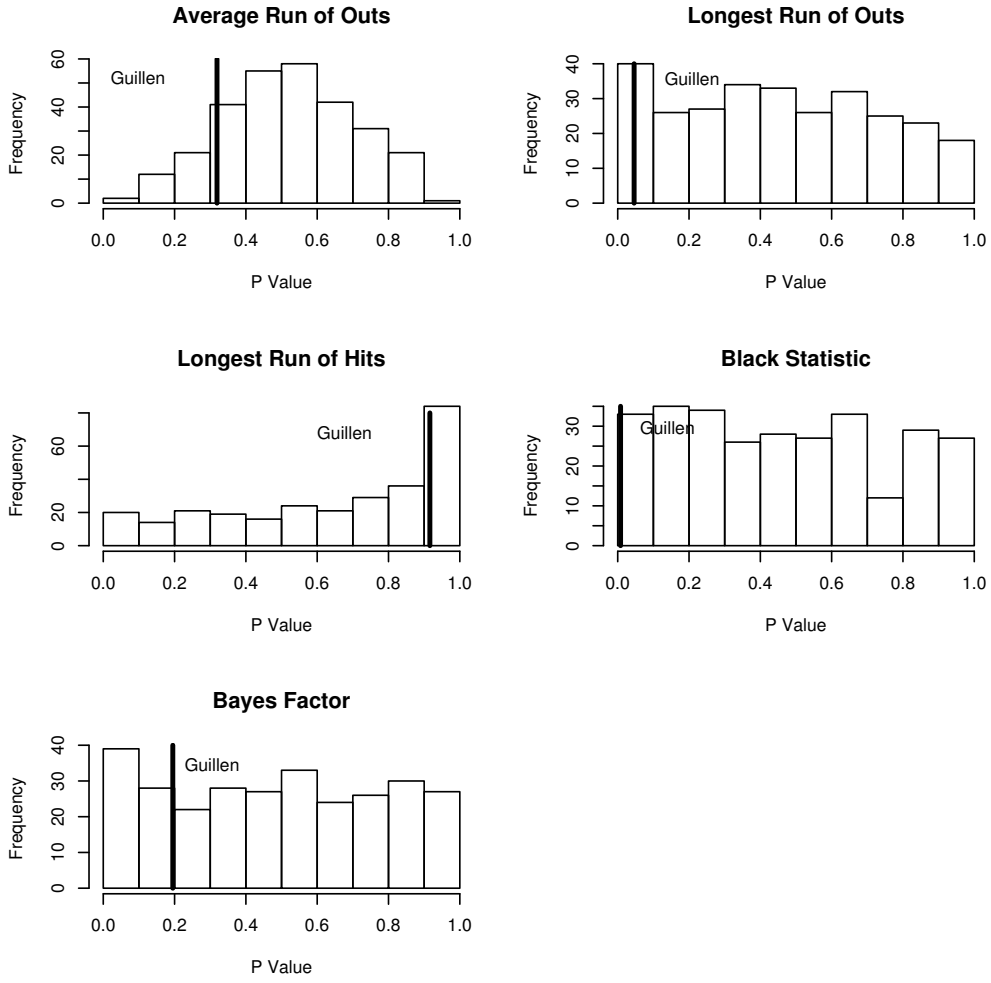

Figure 5: Histograms of p-values for five streaky statistics of hit/out for all 2005 regular players.

8 An Exchangeable Hitting Model

The previous section illustrated the idea of adjusting a player's streaky statistic for his hitting ability and the number of opportunities. But this adjustment procedure can be improved in several ways. We made the assumption that a player's hitting ability p can be estimated by his observed hitting rate during a season. However, since we have the hitting data for all regular players during a season, we can use an exchangeable model, similar to that used to measure streakiness, to better simultaneously estimate the abilities of all players. Efron and Morris (1975) was the first paper on the use of exchangeable modeling to estimate hitting probabilities and Everson (2007) gives a recent survey on the related topic of Stein estimation. Albert (2004, 2006) demonstrates the use of this model in learning about the abilities of baseball hitters and pitchers.

We are given hitting data for N players in a season, where the ith player has y_i successes in n_i attempts. If success is a hit then we observe the number of hits y_i in n_i at-bats, if success is a strikeout, then we observe the number of strikeouts y_i in n_i at-bats, and if success is a home run, then we observe the number of home runs y_i in n_i balls in play. We first assume the y_i's are independent, where y_i is binomial (n_i, p_i). We assume the hitting probabilities $p_1, ..., p_N$ are distributed according to the beta density

$$g(p) = \frac{1}{B(L\eta, L(1-\eta))} p^{L\eta - 1}(1-p)^{L(1-\eta)-1},\ 0 < p < 1.$$

We complete the model by assigning (η, L) the noninformative prior

$$g(\eta, L) = \frac{1}{\eta(1-\eta)} \frac{1}{(1+L)^2},\ 0 < \eta < 1, L > 0.$$

In this model, the parameter η is the mean of the random effects distribution of the success probabilities $p_1, ..., p_N$. The precision parameter L measures the spread of this random effects distribution.

Given the hitting data $(y_i, n_i), i = 1, ..., N$ for each definition of success, we fit this model and the estimates of the hyperparameters η and L are displayed in Table 5. By use of this model, the Bayesian estimate at a player's hitting probability p_j shrinks the observed rate y_j/n_j towards the average rate for all players. The Bayesian estimate is approximately given by

$$\frac{n_i}{n_i + \hat{L}} \frac{y_j}{n_j} + \frac{\hat{L}}{n_i + \hat{L}} \hat{\eta}.$$

For hit/out data, since $\hat{L}$ is large, these estimates shrink the observed batting averages strongly towards the overall batting average. The estimates of the probabilities for striking out and hitting a home run make smaller adjustments to the observed player strikeout rates and home run rates, respectively. (Albert (2005) fits this model to different measures of hitting and shows that some measures are more reflective of players' abilities.)

Table 5: Estimates of parameters η and L using beta binomial model.

Data	$\hat{\eta}$	$\hat{L}$
hitting	0.274	1121
strikeout	0.174	46.3
home run	0.0402	69.8

9 Checking the Exchangeable Model

By use of the exchangeable model, we obtained improved estimates at the group of hitting probabilities. In this section, we see if the pattern of streaky hitting during the 2005 season is consistent with this exchangeable model. Specifically, we check if the observed values of these streaky statistics for the 2005 season are consistent with the predicted values of these statistics from the model. Our method for checking the suitability of this model, based on the posterior predictive distribution, is described as follows. (The use of the posterior predictive distribution in Bayesian model checking is described by Rubin (1984) and in Chapter 6 of Gelman et al (2003).)

1. We simulate a value of (η, K) from the posterior distribution.

2. Given this draw of (η, K), we simulate hitting probabilities $p_1, ..., p_N$ from independent beta distributions.

3. Given these hitting probabilities, we simulate binary sequences $y_{i1}, ..., y_{in_i}$ for all players, where $\{y_{ij}\}$ are independent Bernoulli (p_i). In this step, the numbers of trials $n_1, ..., n_N$ are the actual number of opportunities for the 2005 players.

4. Based on a binary sequence for the ith player, we compute a streaky statistic S_i.

For each simulation (representing a season of baseball), we obtain a collection of streaky statistics $S_1, ..., S_N$.

In Section 7, we found it difficult to interpret the sizes of the streaky statistics due to the nonuniform shapes of the p-value distributions. But the use of the posterior prediction distribution does not have this disadvantage. We are simply checking if the observed streaky data is consistent with simulated draws of future streaky data generated from the exchangeable model. Also, the adjustment method of Section 7 implicitly assumed that the hitting probabilities are equal to the estimated values. In contrast, the simulated data from the posterior predictive method automatically adjusts for the uncertainty in the location of the hitting probabilities.

Suppose we use the adjustment method of Section 7 for the statistics $S_1, ..., S_N$ simulated from the posterior predictive distribution of the exchangeable model. For each simulated player, we simulate independent sequences of Bernoulli data using the estimated probability of success, and compute the p-value for the statistic S_i. The reference distribution for a set of p-values is the uniform distribution. Equivalently, if we transform the p-value by the inverse normal transformation

$$T(p{-}value) = \Phi^{-1}(p{-}value),$$

then the reference distribution for the transformed p-values will be a standard normal curve. Figure 6 plots density estimates of the transformed p-values for ten sets of players simulated from the posterior predictive distribution. The solid line in each figure represents a density estimate of the transformed p-values for the actual 2005 data.

Recall from Section 7 we observed some non-uniform shapes of the distribution of p-values. In particular, the p-values corresponding to the average run of outs had a mound-shaped distribution and the p-values for the longest run of hits had a peak near one. We see the same patterns in the shapes of the transformed p-values of the simulated data in Figure 6. Most of the probability for the average run of outs is concentrated about -1 and $+1$, in contrast to the normal curve where the probability is concentrated between -2 and 2. The distribution of transformed p-values corresponding to the longest run of hits is shifted right compared to the standard normal. The other three distributions of transformed p-values seem close to normal for

the simulated data. The transformed p-values for the actual 2005 players seems to resemble the simulated p-values for the consistent model for each statistic. The conclusion is that the shapes of the p-values seen in Section 7 are a by-product of the bootstrap procedure rather than any indication of streakiness in the data.

Using a different approach, suppose we focus on three measures for streaky hitting, the longest run of outs, the black statistic, and the log Bayes factor, and summarize each collection of statistics for all players by a mean $\bar{S}$ and a standard deviation s_S. If we repeat the posterior prediction simulation m times, we obtain a collection of means $\{\bar{S}_j\}$ and a collection of standard deviations $\{s_{Sj}\}$ corresponding to the exchangeable model. For our 2005 data, we compute the mean streaky statistic $\bar{S}_{obs}$ and the standard deviation s_{obs} across all regular players. To see if these observed values are consistent with the simulated draws from the predictive distributions, we compute the p-values

$$P_1 = \frac{1}{m}\sum_{j=1}^{m} I(\bar{S}_j \geq \bar{S}_{obs}),\ P_2 = \frac{1}{m}\sum_{j=1}^{m} I(s_{Sj} \geq \bar{s}_{obs}).$$

If either of these p-values is close to 0 or 1, then the pattern of streakiness in the 2005 data is not consistent with the pattern of streakiness predicted from the exchangeable model.

Tables 6, 7, and 8 present the values of the p-values using the three streaky statistics for the hit/out, strikeout, and home run data, respectively. From looking at these tables, we see

1. For the **hit/out** data, the p-values appear significant for the mean of the black statistic, and close to significant for both the mean and standard deviation of the log Bayes factor. (We are using "significant" in a loose sense here; we are not adhering to a strict definition that a p-value smaller than 0.05 is significant.)

2. For the **strikeout** data, both the mean and standard deviation appear significant for both the black statistic and the log Bayes factor.

3. For the **home run data**, the standard deviation appears significant for both the longest run of outs data and the log Bayes factor.

A significant (small) p-value for the mean indicates a general tendency of the players to be streaky and a significant p-value for the standard deviation

indicates that there is more variability in streakiness (perhaps, more streakiness among some players) than predicted from the exchangeable model. For the strikeout data, there appears to be the greatest evidence of streakiness as measured by both the black and Bayes factor statistics. For the home run data, there appears to be a small group of players that exhibit unusual streakiness (since the standard deviations are unusually large). There is less evidence of streakiness for the hit/out data, but there is evidence that players tend to be streaky with respect to the black statistic.

Table 6: Posterior predictive p-values for hit/out data.

Statistic	P_1	P_2
longest run of outs	0.211	0.317
black statistic	0.033	0.344
log Bayes factor	0.124	0.092

Table 7: Posterior predictive p-values for strikeout data.

Statistic	P_1	P_2
longest run of outs	0.185	0.429
black statistic	0.086	0.057
log Bayes factor	0.008	0.040

Table 8: Posterior predictive p-values for home run data.

Statistic	P_1	P_2
longest run of outs	0.304	0.045
black statistic	0.440	0.424
log Bayes factor	0.59	0.040

To confirm these p-value calculations, we focus on the strikeout data. We simulated twenty datasets (each dataset the binary sequences for all

regular players) from the posterior predictive distribution of the exchangeable model and computed the log Bayes factors for each dataset. Each of the distributions of the log Bayes statistic, representing the streaky statistics for all players, was summarized by the 5th, 50th and 95th percentiles and the light lines of Figure 6 show these distributions. The actual distribution of log Bayes factors for the 2005 players is plotted as a solid line. Comparing the 2005 data with the simulated data, the most visible difference is that the distribution of 2005 values appears to have a longer right tail than the simulated distributions. The interpretation is that there are some players that are unusually streaky in their patterns of strikeouts.

10 Streaky Players

The work in the previous section demonstrated that there was some lack-of-fit of the exchangeable model. Specifically, the distribution of streaky statistics seems somewhat more spread out than predicted under the model. So this motivates looking at players that appear unusually streaky. We rank the players by use of the p-values in the adjustment procedure of Section 7. Recall that a small p-value corresponded to a player that has an unusual streaky pattern. Since we have 284 regular places, one would expect there to be $284 \times .05 = 14$ "streaky" players even if the consistent-p model was true. Table 9 gives the number of streaky players for each of the three definitions of hitting success and the three streaky statistics. Note that we observe more than 14 streaky players only for the hit/out data and for the strikeout data using the Bayes factor criterion. This indicates that we should only look at a small number of the top streaky hitters.

Table 9: Number of streaky players, where "streaky" corresponds to a p-value smaller than 0.05. Since there are 284 players, one would expect there to be $284 \times .05 = 14$ significant players by chance.

Statistic	Hit/out data	Strikeout data	Home run data
longest run of outs	24	14	11
black statistic	20	15	5
log Bayes factor	24	23	14

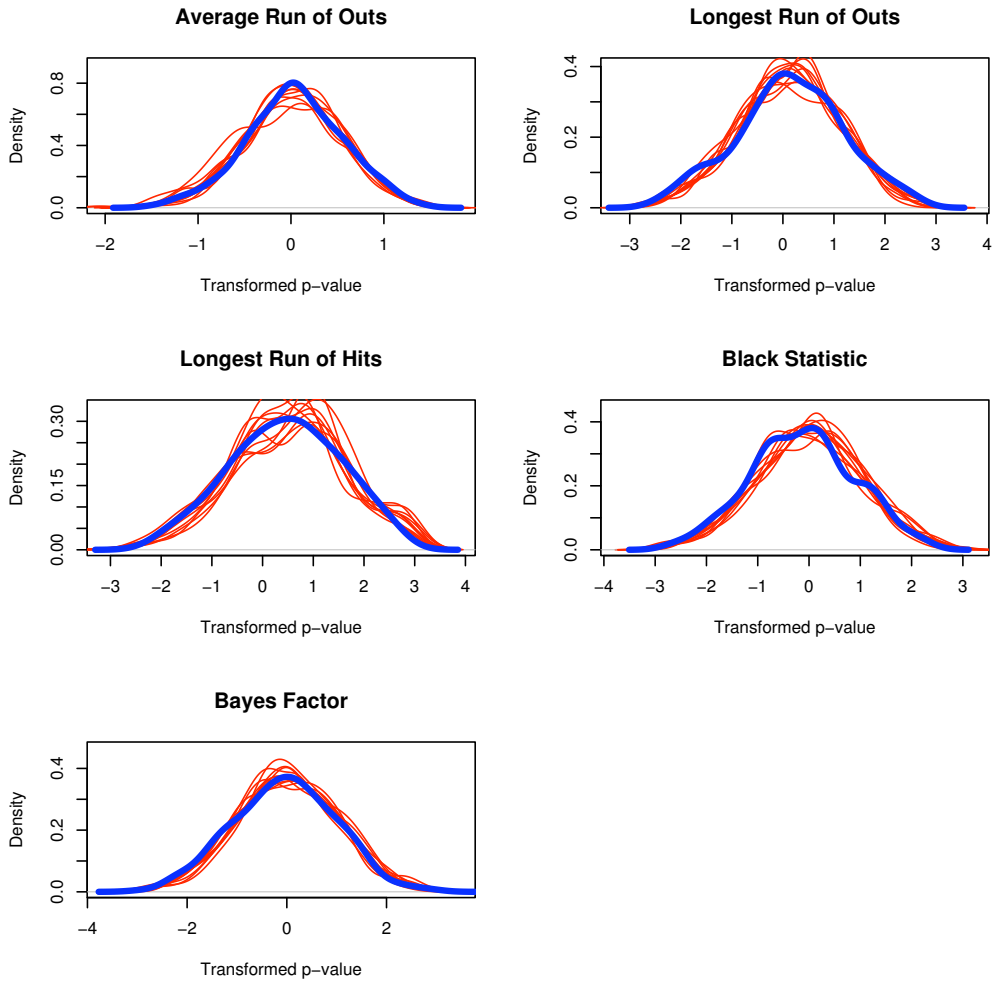

Figure 6: Distribution of transformed p-values of five statistics simulated from beta/binomial exchangeable model. The ten lines represent the transformed p-values for ten simulations from the model and the solid line represents the transformed p-values for the 2005 regular players.

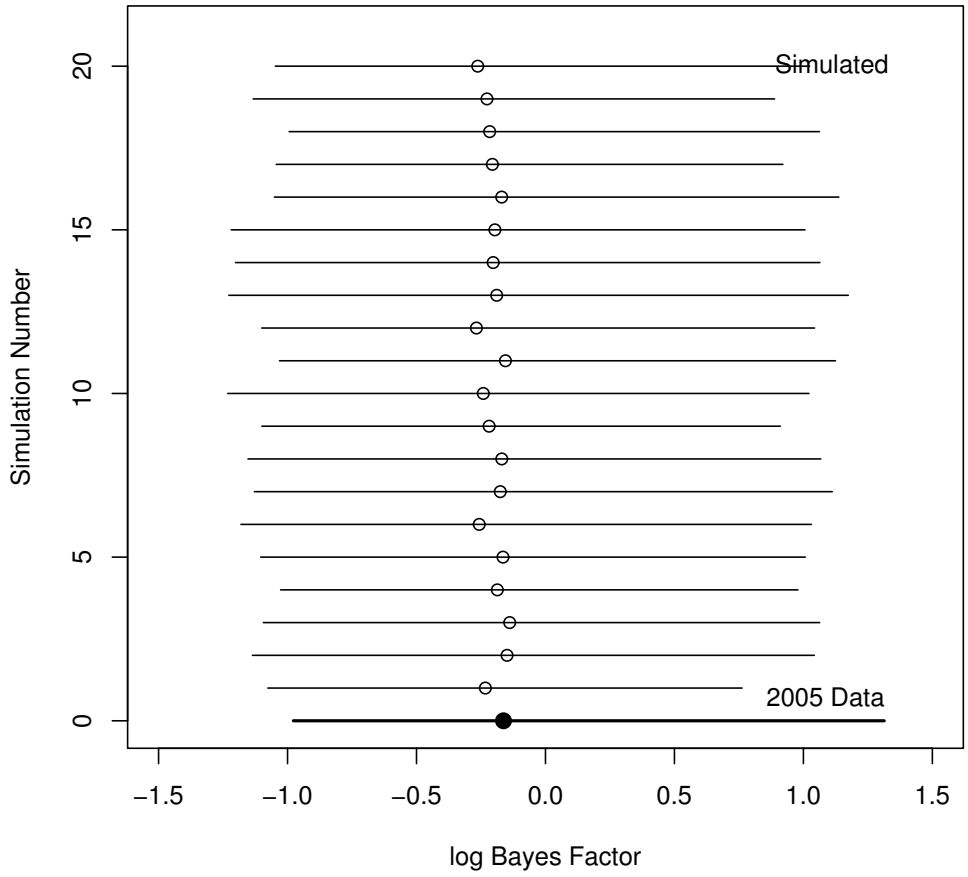

Figure 7: Distributions of twenty simulated distributions of log Bayes factors for strikeout data assuming beta/binomial model. Each line shows the 5th, 50th, and 95th percentiles. The bold line is the distribution of log Bayes factors for the 2005 regular players.

Tables 10 and 11 give lists of the top ten streaky hitters for the three types of hitting and the black statistic and the Bayes factor. A couple of observations can be made from this table. First, there is little overlap of players across types of hitting. That means that a player who appears unusually streaky in hits doesn't generally appear streaky in strikeouts and home runs. Also there is a small amount of overlap between the lists for the Bayes factor and black statistics. So the definition of a streaky player depends on the particular definition of streakiness that we use.

Table 10: Top ten streaky players with respect to hits/outs, strikeouts, and home runs using Bayes factor statistic. The players are ranked with respect to decreasing p-value in the adjustment procedure of Section 6.

Hit/Outs	Strikeout	Home run
Jorge Posada	Eric Hinske	Bobby Abreu
Chipper Jones	Brian Roberts	Clint Barmes
Mike Piazza	J.J. Hardy	Hee Choi
Eric Byrnes	Todd Hollandsworth	Gary Matthews
Neifi Perez	Ruben Gotay	Bernie Williams
Victor Martinez	Shawn Green	Rod Barajas
Jason Bay	Shea Hillenbrand	Edgardo Alfonzo
Todd Helton	Luis Matos	Joe Randa
Cesar Izturis	Alex Rodriguez	Chris Burke
Jeff Kent	Rondell White	Ichiro Suzuki

11 Final Comments

What insight have we gained about streaky hitting on this analysis? First, it is difficult to judge the streaky ability of a single player on the basis of his streaky performance during a single season. Since there are so many players in baseball playing over many seasons, it is very possible to observe long streaks or extreme short-term batting performances even if all players are consistent-p hitters. This multiplicity problem is similar to the interesting streaky behavior observed if many students simultaneously toss coins.

Table 11: Top ten streaky players with respect to hits/outs, strikeouts, and home runs using black statistic. The players are ranked with respect to decreasing p-value in the adjustment procedure of Section 6.

Hit/Outs	Strikeout	Home run
Mike Sweeney	Hee Choi	Jose Guillen
Juan Uribe	Brian Roberts	Kevin Millar
Carlos Guillen	Todd Hollandsworth	Hank Blalock
Jeff Kent	Jacque Jones	Hee Choi
Chipper Jones	Matt LeCroy	Bobby Abreu
Cesar Izturis	Julio Lugo	Mark Bellhorn
Matt Lawton	Justin Morneau	Ichiro Suzuki
Felipe Lopez	Jeff Conine	Bret Boone
Victor Martinez	Gary Matthews	Bernie Williams
Eric Byrnes	Rondell White	Mark Teixeira

When we controlled the multiplicity problem by looking at the streaky performances of all regular players, new problems were encountered. The size of a streaky statistic, say the length of the longest run of successes, depends on the player's ability and his number of opportunities during a season. We proposed a simple simulation adjustment method that controlled for ability and number of opportunities. When we perform this adjustment method for a player, we obtain a p-value that is the probability that a consistent-p hitter with the same ability and opportunities would be as streaky as the observed player. We improved this adjustment method by use of an exchangeable model that provided better estimates of the players' hitting abilities. We saw that the bootstrap procedure for computing a p-value can lead to p-values that are not uniformly distributed even when the consistent model is true.

One interesting conclusion from this study is that the exchangeable model assuming that all players are consistent-p hitters actually explains most of the variation of streaky hitting that we observe in a single season. Does this imply that all players are consistent-p hitters? Of course not. The probability that a player succeeds in hitting should depend on many variables such as the opposing pitcher, the ballpark, and the game situation. But this exchangeable model that ignores these extra effects seems to be useful in

understanding the patterns in streaky hitting in baseball. It is doubtful that much more could be learned about streakiness through the addition of new covariates.

Although the model explains most of the streaky hitting in baseball, some model misfit is present. Among all regular players, there are more streaky hitters than one would predict on the basis of this exchangeable model. On the basis of this observation, we identified the players who were unusually streaky in the patterns of getting hits, home runs, and strikeouts. But there is little evidence on the basis of this study that particular players are generally streaky – players that appeared streaky in the pattern of hits/outs did not appear streaky in the pattern of strikeouts. This suggests that it may be hard to find players that exhibit streaky behavior across seasons, although this would be an interesting study for future work.

References

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

Albert, J. and Pepple, P. (2001), "Using Model/Data Simulations to Detect Streakiness," *The American Statistician*, 55, 41-50.

Albert, J. (2004), "Streakiness in Team Performance," *Chance*, Vol. 37-43.

Albert, J. (2005), "A Batting Average: Does it Represent Ability or Luck?," *STATS*, 44, Fall-Winter issue.

Albert, J. (2006), "Pitching Statistics, Talent and Luck, and the Best Strikeout Seasons of All-Time," *Journal of Quantitative Analysis of Sports,* Vol. 2.

Albert, J. (2007), *Bayesian Computation with R*, New York: Springer.

Albert, J. and Bennett, J. (2003), *Curve Ball*, Copernicus Press.

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

Bar-Eli, M., Avugos, S., and Raab, M. (2006), "Twenty Years of 'Hot Hand' Research: Review and Critique, *Psychology of Sport and Exercise*, 7, 525-553.

Bayarri, M. and Berger, J. (2000), "P-Values for Composite Null Models," *Journal of the American Statistical Association*, 95, 1127-1142.

Berry, S. (1991), "The Summer of '41: A Probability Analysis of DiMaggio's Streak and Williams' Average of .406," *Chance*, 4, 8-11

Raftery, A. (1996), "Approximate Bayes Factors and Accounting for Model Uncertainty in Generalised Linear Models," *Biometrika*, 129-144.

Dorsey-Palmateer, R. and Smith, G. (2004), "Bowlers' Hot Hands," *The American Statistician*, 38-45.

Efron, B. and Morris, C. (1975), "Data Analysis Using Stein's Estimator," *Journal of the American Statistical Association,* 70, 311-319.

Efron, B. and Tibshirani, R. (1994), *An Introduction to the Bootstrap,* CRC Press.

Everson, P. (2007), "Stein's Paradox Revisited," *Chance*, 20, 49-56.

Gelman, A., Carlin, J. B., Rubin, D. B., and Stern, H. S. (2003), *Bayesian Data Analysis*, CRC Press.

Gilovich T., Vallone, R. and Tversky, A. (1985), "The Hot Hand in Basketball: on the Misperception of Random Sequences," *Cognitive Psychology*, 17, 295-314.

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

Kass, R. and Raftery, A. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773-795.

Kass, R. and Vaidyanathan, S. (1992), "Approximate Bayes Factors and Orthogonal Parameters, with Application to Testing Equality of Two Binomial Proportions," *Journal of the Royal Statistical Society,* Series B, 129-144.

KLAASSEN, F. and MAGNUS, J. (2001), "Are Points in Tennis Independent and Identically Distributed? Evidence from a Dynamic Binary Panel Data Model" *Journal of the American Statistical Association*, 96, 500-509.

ROBINS, J., VAN DER VAART, A. and VENTURA, V. (2000), "Asymptotic Distribution of P Values in Composite Null Models," *Journal of the American Statistical Association*, 95, pp. 1143-1156.

RUBIN, D. (1984), "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician, *Annals of Statistics*, 12, 1151-1172.

STERN, H. (1997), "Judging Who's Hot and Who's Not," *Chance,* 10, 40-43.

TVERSKY, A. and GILOVICH, T. (1989),"The Cold Facts about the 'Hot Hand' in Basketball, *Chance,* 2, 16-21.

WARDROP, R. (1999),"Statistical Tests for the Hot Hand in Basketball in a Controlled Setting," Technical report, Department of Statistics, University of Wisconsin-Madison.

WARDROP, R. L. (1995), Simpsons Paradox and the Hot Hand in Basketball, *American Statistician*, 49, 2428.