Looking at Spacings to Assess Streakiness

Jim Albert
Department of Mathematics and Statistics
Bowling Green State University

September 2013

The Problem

- Collect hitting data for all players in a season.
- Focus on patterns of hot and cold hitting for all players.
- Is there evidence that players are truly streaky?
- Or maybe players are not truly streaky ...
- Patterns we see are similar to the patterns one would observe by flipping a collection of coins.

Binary Sequence

- Observe sequence of hitting data for a particular baseball player.
- For each opportunity, observe a hit (1) or a out (0).
- Interested in the pattern of streaks and slumps
- Often interested in the *spacings*, the number of failures between consecutive successes.
- The media reports *large* values of these spacings.
- In 2012 season, Josh Reddick had a "0 for 30" streak.

Simple example

- Observe the following data for a player in 15 at-bats:
- 0 0 1 1 1 0 0 0 0 1 0 1 0 0 1
- Spacings are 2, 0, 0, 4, 1, 2
- The values of these spacings are informative about the player's success probabilities.

How to Measure Streakiness?

- Long runs of successes or long runs of failures.
- Permutation test – measure deviation from randomness by a p value
- Focus here is on a Bayes test statistic – compare two models

A Geometric Model

- Let $y_1, ..., y_n$ denote observed spacings for a particular player
- Assume $\{y_i\}$ independent, where y_j is Geometric (p_j):

$$f(y_j|p_j) = p_j(1 - p_j)^{y_j}, \ \ y_j = 0, 1, 2, ...$$

- Talk about *true* streakiness by considering models on probabilities $p_1, ..., p_n$.

The Consistent Model: Model M

- If hitter is *not* streaky or truly consistent, one believes hitting probabilities are constant:

$$p_1 = ... = p_n = p.$$

- Assume we have little information about location of p.
- Assign p a noninformative prior

$$g(p) = \frac{1}{p(1-p)}.$$

The Streaky Model: Model M_K

- If hitter is truly streaky, one believes hitting probabilities vary across the season.

- Assume $p_1, ..., p_n$ is distributed according to a beta density

$$g(p) \propto p^{a-1}(1-p)^{b-1}, \ 0 < p < 1.$$

- Parameterize in terms of mean and precision

$$\eta = \frac{a}{a+b}, \ K = a + b.$$

- Fix value of K and assign η noninformative prior

$$g(\eta) = \frac{1}{\eta(1-\eta)}.$$

A Test Statistic

- A Bayes factor.
- Ratio of the marginal probabilities of the observed data y under the two models M_K and M:
- Bayes factor in support of true streakiness is

$$BF_K = \frac{f(y|M_K)}{f(y|M)}.$$

- Values of $BF_K > 1$ support the streaky model, and values $BF_K < 1$ support the consistent model.

Use a Subjective Approach for Obtaining K

- Think about the variability of success probabilities $\{p_j\}$ if the player was truly streaky.
- Specify a standard deviation of the probabilities.
- This can be used to assess a value of K.

Look at All Players in the 2012 Season

- Consider all players with at least 200 at-bats. (Want to exclude pitchers.)
- For each player, use Bayes factor with $\log K = 5$ to test for streakiness.
- Measure Bayes factor on log scale (positive values of $\log BF_K$ support streakiness)
- Plot log Bayes factors against the sample size (number of at-bats).
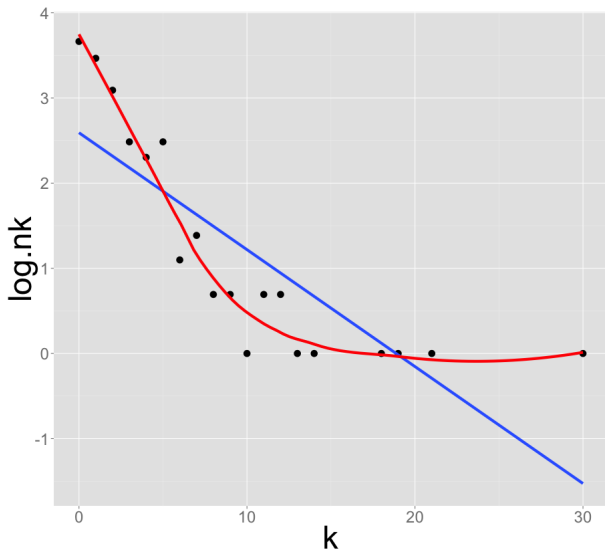
Bayes Factors for all 2012 Players
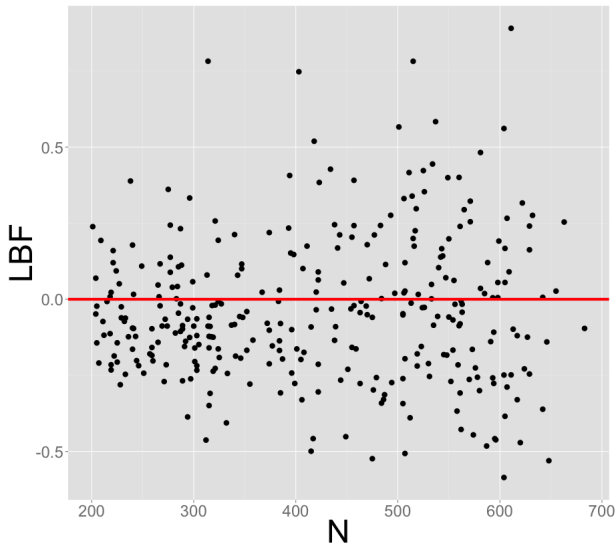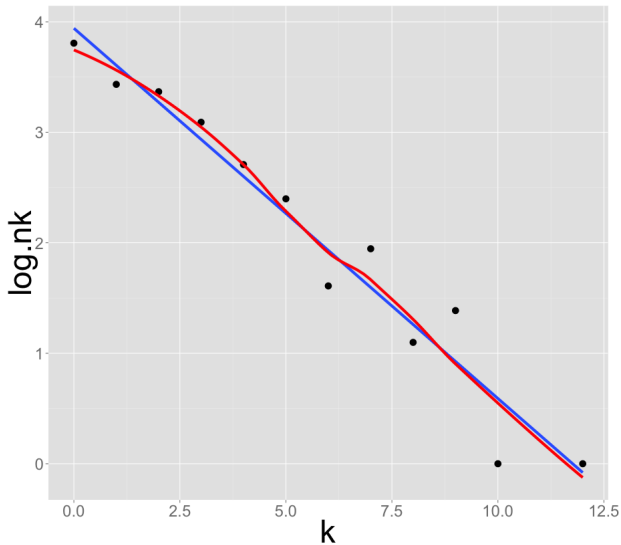
Label 2 Interesting Outliers
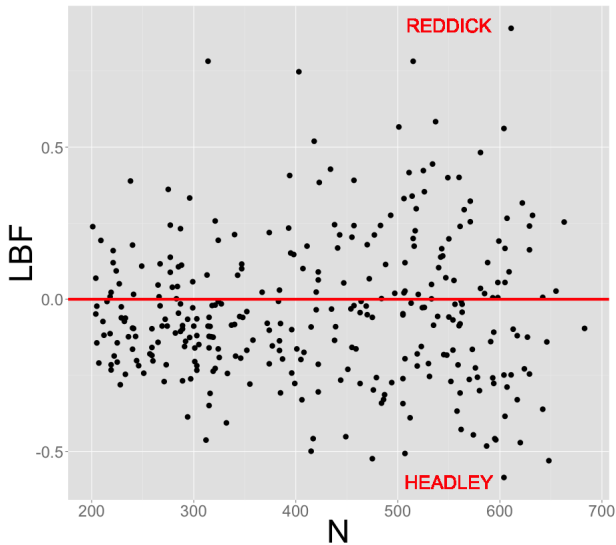
A Geometric Plot

- Tally spacings - get $\{y, f_y\}$
- Plot $\log f_j$ (vertical) against j (horizontal)
- If geometric, plot will be linear
- Graph shows deviations from geometric

Geometric Plot for Reddick

Geometric Plot for Headley

Back to 2012 Data

- Evidence for streakiness ($\log BF > 0$) for 120 (38%) of the players.
- Observe a very streaky player (Reddick) and a very consistent hitter (Headley)
- But maybe players are truly consistent and we are observing streakiness due to multiplicity.
- Are these streaky outcomes a result of a "reasonable" consistent model for hitting for all players?

An Exchangeable Consistent Model for Hitting

- Observe data $\{(y_j, n_j)\}$, where y_j is the number of hits in n_j at-bats for jth player.
- Assume y_j is binomial(n_j, P_j) (consistent model)
- Assume success probabilities $P_1, ..., P_N$ are exchangeable.
- Posterior estimate of P_j shrinks (adjusts) observed batting average y_j/n_j towards the overall average.

What Streaky Data Do We See from a Consistent Model?

- Estimate hitting probabilities $P_1, ..., P_N$ from the exchangeable model
- Simulate individual at-bat results $\{w_1, ..., w_{n_j}\}$ from independent Bernoulli distributions with constant probability $\hat{P}_j$.
- For each sequence of hitting data, compute Bayes test in support of streakiness $\log BF_K$.
- Find the number of players where there is support for streakiness ($\log BF_K > 0$).
- Repeat this process 200 times

Predictive Distribution of Number of Streaky Players
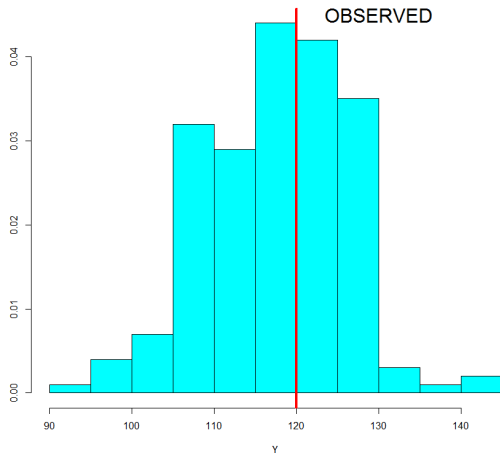
Summing Up ...

- Consistent model (one hitting probability for each player) seems useful in predicting pattern of streakiness of all players in a season.
- Players may exhibit streakiness for a particular season, but little evidence to suggest that players have streaky tendencies
- End of story?

But There are Other Definitions of "Success"

- Different definitions of batting "success", such as "hit", "strikeout", "home run".
- Strikeout and home run rates are more informative about the player abilities
- See this by fitting a random effects model.

A Random Effects Model

- Observe success rates y_1/n_1, ..., y_N/n_N of N players.
- Give $p_1, ..., p_N$ a distribution $g(p)$ (random effects distribution)
- Estimate $g(p)$ from the data.
- Learn about the fraction of the variability in the rates that is due to (1) variation between player abilities (the p_j), and (2) binomial variation (luck).
- On this scale, strikeout rates exceed home run rates which exceed batting averages

Look at 53 Seasons of Batting Data

- Looked at batting sequences for all players with at least 200 AB for seasons 1960 through 2012
- Considered three definitions of "success", strikeout, home run, and hit
- For each season, performed the same predictive analysis – simulated predictive distribution of the number of streaky hitters ($\log BF > 0$) assuming consistent model
- Focused on predictive p-value of observed number of streaky hitters

Histogram of Predictive P-Values for H, HR, SO
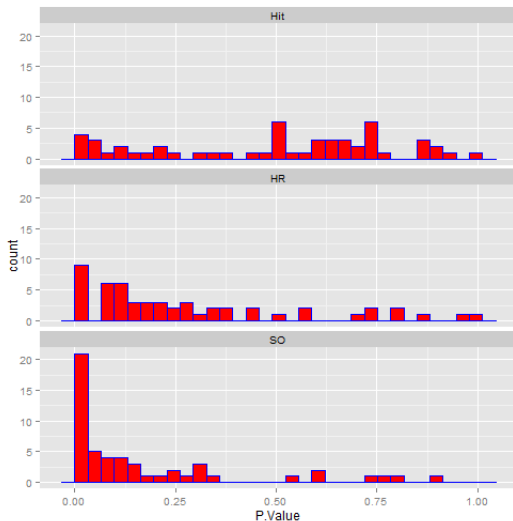
This is Interesting

- For home run rates and strikeout rates, there is more observed season streakiness among players that one would predict based on a consistent model
- There is a long right-tail in the streakiness distribution
- So there are "interesting" patterns of streakiness
- Focus on home run hitting – do there exist streaky home run hitters?

What Were the Streakiest Seasons of Home Run Hitting?

- 1998 Ivan Rodriguez, 1967 Lou Brock, 1965 Paul Schaal, 2007 Rickie Weeks
- What do these streaky seasons look like?
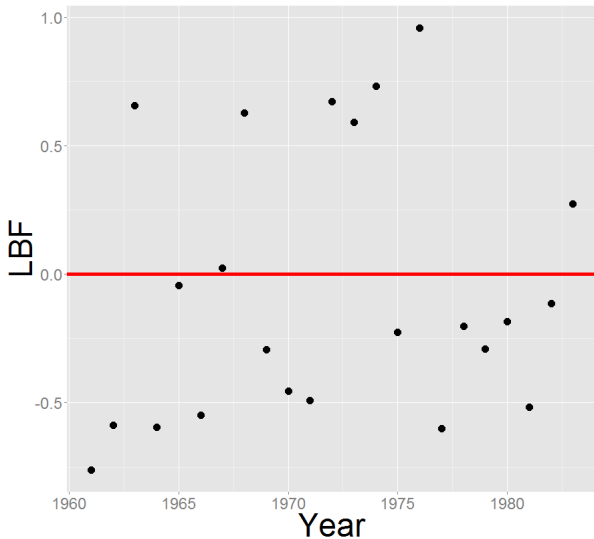- Focus on the 1998 Ivan Rodriquez

Graph of 1998 Ivan's Home Runs

Are There Players Who Tend to Be Streaky?

- Suppose one is "streaky" if $\log LBF > 0.5$
- Focus on the players who had the most streaky seasons
- Carl Yastrzemski had six streaky seasons in his career (1961 - 1983)
- Occurred in seasons 1963, 1968, 1972, 1973, 1974, 1976
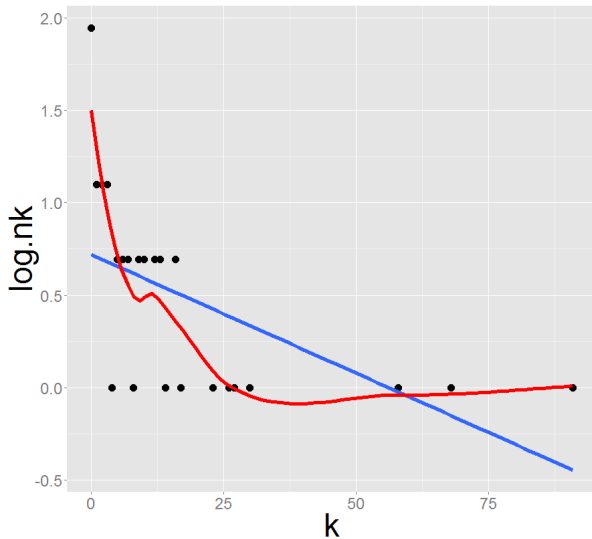
Carl Yastrzemski Streaky Measures

What Were the Streakiest Hitters Among the Sluggers (40+ HR)?

- 2002 Shawn Green, 2009 Albert Pujols, 1962 Willie Mays
- Focus on the 2002 Shawn Green
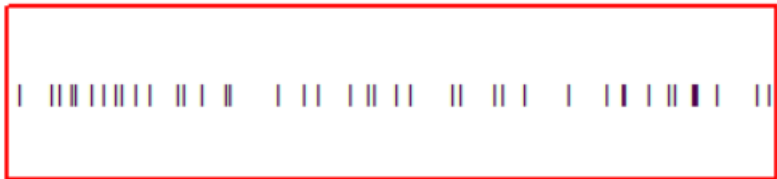
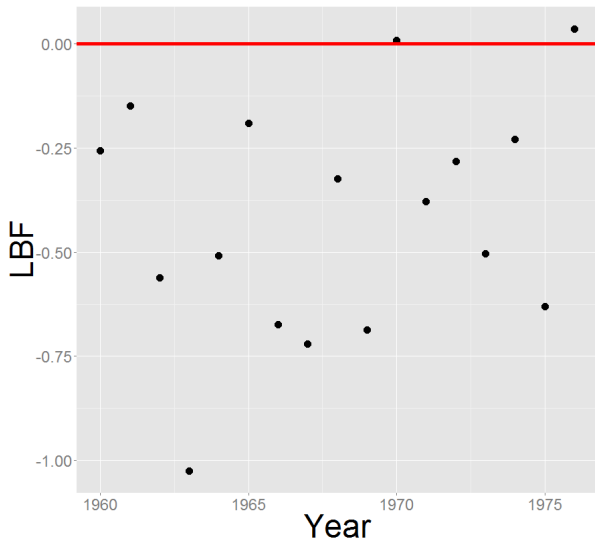# 2002 Shawn Green

Geometric Plot for Green

What Was the Most Consistent Hitter/Season Among the Sluggers (40+ HR)?

- 1963 Hank Aaron
- In fact, for practically all of Aaron's seasons, $\log BF < 0$
- Aaron was a remarkably consistent home run hitter
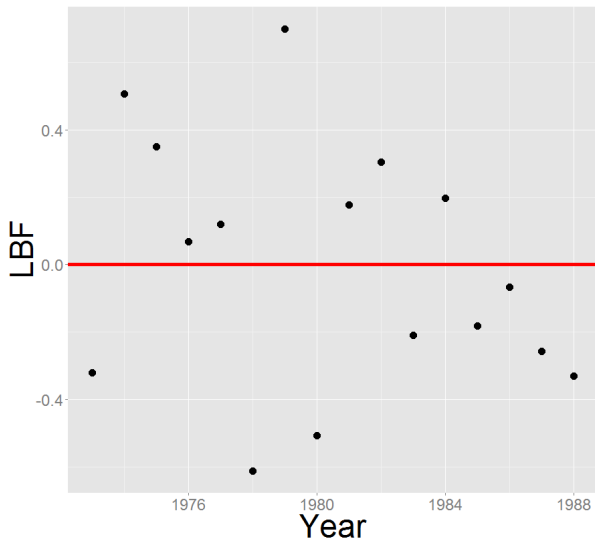- Was it due to his style of hitting?

1963 Hank Aaron

Career Hank Aaron

Mike Schmidt

- Great slugger for the Phillies
- Wrote a paper some years ago exploring timing of his 548 home runs
- Looked at spacings
- Mike was streaky early, but more consistent pattern of home run hitting later in his career
- Confirmed his patterns using this approach

Career Mike Schmidt

Closing Comments

- Bayes factor is a useful measure of streakiness
- Posterior predictive analysis helpful for seeing if there is "interesting" streakiness
- Hit/out is a different story than HR or Strikeout sequences
- Some players have interesting patterns of streakiness or consistency

Joe DiMaggio

The Great Streak in 1941

- 56-game hitting streak
- Recently, Retrosheet has some play-by-play records available for 1941 season
- Obtain DiMaggio's sequence of hits and outs for all at-bats
- Is there evidence that Joe was streaky in the 1941 season?

Was Joe Streaky in '41?

- No ... $\log BF \approx 0$
- No evidence to support streaky or consistent model
- What does this mean?
- Actually, great hitters are unlikely to be streaky during a season

References / Software

- Albert, J. (2008), "Streaky Hitting in Baseball", *Journal of Quantitative Analysis of Sports*, vol. 4.
- Albert, J. (2013), "Looking at Spacings to Assess Streakiness", *Journal of Quantitative Analysis of Sports*, vol. 9.
- Albert, J. (2013), "Was Joe DiMaggio Streaky?", *Baseball Prospectus* website
- R package BayesTestSpacings, http://bayes.bgsu.edu/spacings (download package and look at examples)