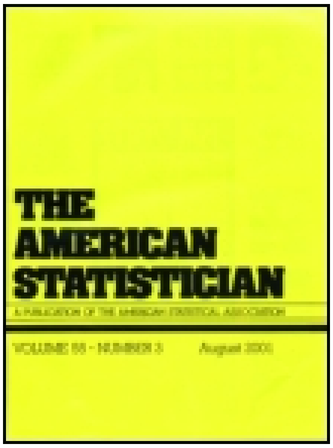

THE AMERICAN STATISTICIAN

The American Statistician
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/utas20

Using Model/Data Simulations to Detect Streakiness

Jim Albert[a] & Patricia Williamson[a]

[a] Jim Albert is Professor, Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH 43403 . Patricia Williamson is Associate Professor, Department of Mathematical Sciences, Virginia Commonwealth University, Richmond, VA 23284-2014.

Published online: 01 Jan 2012.

To cite this article: Jim Albert & Patricia Williamson (2001) Using Model/Data Simulations to Detect Streakiness, The American Statistician, 55:1, 41-50, DOI: 10.1198/000313001300339923

To link to this article: http://dx.doi.org/10.1198/000313001300339923

Using Model/Data Simulations to Detect Streakiness

Jim ALBERT and Patricia WILLIAMSON

A simulation-based approach is proposed for approximating a Bayesian analysis. Parameters and data are simulated from a Bayesian model and inference about a parameter is performed by exploring the set of simulated parameter values conditional on a set of values of a simulated statistic. The approach is used to learn about parameters of a streaky model on the basis of a statistic used to measure streakiness. The method is illustrated to detect streakiness in baseball hitting data and basketball shooting data.

KEY WORDS: Bayesian inference; Coin-tossing; Hot-hand; Markov switching model; Overdispersion, Quasi-binomial.

1. INTRODUCTION

There has been much recent interest in the detection of streakiness or the "hot hand" in the performance of athletes in baseball, basketball, and other sports. Gilovich, Vallone, and Tversky (1985) and Tversky and Gilovich (1989) discussed the existence of the hot hand in basketball and concluded that any observed streakiness in data is simply one's misperception of the patterns inherent in random sequences. Larkey, Smith, and Kadane (1989), in their analysis of basketball shooting data from a number of professional players, took a different view. They concluded that there is some statistical evidence for the hot hand. Albright (1993) performed a number of statistical tests to detect streakiness in hitting data for a large group of professional players. Although some players appeared to exhibit streaky behavior for particular seasons, Albright found little support for the hot hand theory in baseball. In their comments on Albright's article, Stern and Morris (1993) criticized some of the tests that were conducted by Albright, and Albert (1993) presented a hidden Markov switching model that could be used to model streakiness. Wardrop (1999) also discussed the hot-hand debate. Specifically, Wardrop criticized particular tests used by Gilovich et al. (1985) and Tversky and Gilovich (1989) to judge if a player has the hot hand. See Stern (1997) for a recent discussion of the statistical detection of streakiness.

In the basketball and baseball settings, one observes a sequence of binary observations $y_1, \ldots, y_n$, where $y_i = 1$ if a successful (baseball) hit or (basketball) shot is observed, and $y_i = 0$ otherwise. The basic *coin-tossing model* assumes that the y_i are independent Bernoulli trials with a constant probability of success p. The *hot hand* or *streaky hypothesis* represents a deviation from the coin-tossing model. In this alternative hypothesis, there is either *nonstationarity* where the probability of success does not stay constant over the trials, or *autocorrelation* where the probability of success on a given trial depends on the player's success in recent trials. We refer to streakiness as a presence of nonstationarity and/ or autocorrelation in the sequence.

In the usual approach to searching for the hot hand, a value of a statistic T is observed, such as a long run of successes, which seems to support the hot hand theory. Then one computes a p value of this statistic assuming the coin-tossing model. If this p value is small, then one rejects the coin-tossing model.

However, this approach generally is not effective in detecting the hot hand theory. One problem is that it is difficult to gauge the smallness of the p value. One will typically find significant p values if a large number of tests are made, or if the test is performed on a sample sequence with many observations. A more serious criticism of this approach is that there is little investigation into the ability of the test to detect deviations from the coin tossing model. Often many tests are performed without understanding the power of the tests to detect the type of streaky patterns that are of interest. By doing a power study, Wardrop (1999) demonstrated that particular "streaky statistics" used by Gilovich et al. (1985) are almost useless in detecting deviations from coin-tossing. One reason why power calculations are not made is that it can be hard to construct alternative models that represent a belief in a hot hand theory. Barry and Hartigan (1993) and Albert (1993) represented two of the relatively short list of articles that used particular hot hand models to analyze sequences of win/ loss or hit/ not-hit data.

This article proposes a simple Bayesian method, based on simulation, for detecting streakiness in sports data. Suppose that one can construct a streaky model that can possibly generate the sequence of successes and failures. We assume that this model is parameterized by a vector of parameters θ, and initial beliefs about θ are represented by a prior distribution. Suppose that a streaky statistic $T = t^{\text{obs}}$ is observed—what have we learned about the parameter θ? Section 2 describes a simple scheme proposed by Albert (1997) in a teaching context, which simulates from the joint distribution of the parameter and the data. By exploring the set of simulated parameter values conditional on a set of values of the simulated statistic T, we perform inferences on the streakiness parameter θ.

Section 3 introduces the use of the model/ data simulation scheme for learning about possible nonstationarity in home run data for Mark McGwire. An overdispersion model is used to model nonstationarity, and the approxi-

Jim Albert is Professor, Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH 43403 (E-mail: albert@bgnet.bgsu.edu). Patricia Williamson is Associate Professor, Department of Mathematical Sciences, Virginia Commonwealth University, Richmond, VA 23284-2014.

mate posterior distribution on the overdispersion parameter based on the proposed simulation method is compared with the exact posterior distribution. Section 4 illustrates the fitting of a Markov switching model to detect streakiness in day-to-day baseball hitting data for a particular player. A single parameter λ is introduced to measure the distance from coin-tossing toward a streaky model, and we investigate the use of six "streaky statistics" in learning about this parameter. We measure the usefulness of a statistic in inference by means of the correlation of the simulated pairs of parameter and statistic values. Section 5 considers the issue of streakiness for basketball shooting data for a particular college player. A quasi-binomial distribution is used to model shooting proportions in groups of five shots—this distribution can model variability in these proportions that is greater or smaller than what is predicted from a binomial distribution. In a plot of dispersion parameters across time, the shooter appears to be more consistent over time, and a model/ data simulation algorithm is used to assess the significance of the observed pattern in the data.

2. SIMULATING MODELS AND DATA

Suppose that one observes data $y_1, \ldots, y_n$ from a sampling model $f(y|\theta)$ and beliefs about the parameter θ are described by means of a proper prior distribution $g(\theta)$. (In the following, y will denote the vector of observations and θ is a vector of parameters.) The object is to learn about θ after the data $y = y^{\text{obs}}$ are observed, and, by Bayes's rule, this learning is accomplished by computation of the posterior distribution

$$g(\theta|y^{\text{obs}}) \propto \prod_{i=1}^{n} f(y_i^{\text{obs}}|\theta)g(\theta).$$

Now suppose that this posterior density is awkward to compute directly due to the complicated form of the likelihood. However, there exists a statistic $t(y)$ that we believe summarizes most, if not all, of the information contained in the data y. (If $t(y)$ is sufficient, then one can show that the posterior distribution will depend on y only through the value of the sufficient statistic, and so t will summarize *all* of the information about θ.) In this case, the posterior density can be approximated by the conditional density

$$g(\theta|t^{\text{obs}}) \propto f(t^{\text{obs}}|\theta)g(\theta),$$

where $f(t|\theta)$ is the sampling density of the statistic t.

In our simulation approach, it will be difficult to obtain a large sample of simulated values of θ conditional on a specific value of t, since t takes on many discrete values. However, one can approximate the likelihood

$$f(t^{\text{obs}}|\theta) \approx \Pr(t^{\text{obs}} - \epsilon < t < t^{\text{obs}} + \epsilon)\frac{1}{2\epsilon}$$

for a small value of ϵ, and it follows that the conditional density of interest is approximated by

$$g(\theta|t^{\text{obs}}) \approx g(\theta|t^{\text{obs}} - \epsilon < t < t^{\text{obs}} + \epsilon).$$

To summarize, if the statistic t is nearly sufficient for θ, then we can approximate the posterior density of interest by the density of θ conditional on the statistic t falling in a small interval about the observed value t^{obs}.

This approximation method is straightforward to apply in practice. In the following algorithm, we use the prior and sampling density to simulate sets of the parameter θ and the data y. From each simulated dataset y, the statistic t is computed. Then inference about θ is performed by focusing on the simulated values of θ where the simulated t values fall in an interval about the observed value t^{obs}.

2.1 The Basic Algorithm

1. **Simulate models and data.** Repeat the following three steps to obtain N model/ data simulated pairs $\{(\theta_j^S, t_j^S)\}$.

   (a) Simulate a value of the parameter θ from the prior density $g(\theta)$—call this simulated value θ^S.
   (b) Simulate data $y_1^S, \ldots, y_n^S$ independently from the sampling density $f(y|\theta^S)$.
   (c) Compute the value of the statistic $t^S = t(y_1^S, \ldots, y_n^S)$ based on the simulated data.

2. **Perform the inference.** If $t = t^{\text{obs}}$ is observed, one learns about the parameter θ by focusing on the set of simulated pairs for which t^S falls in the interval $(t^{\text{obs}} - \epsilon, t^{\text{obs}} + \epsilon)$, for a suitably chosen value of the interval half-width ϵ.

How should ϵ be chosen? One wishes to choose ϵ small so that one has a good approximation to the posterior density which conditions on the value $t = t^{\text{obs}}$. On the other hand, one needs to collect enough simulation draws $\{\theta^S\}$ so that one gets accurate summaries of the posterior distribution. In the examples, ϵ was chosen so that the interval $(t^{\text{obs}} - \epsilon, t^{\text{obs}} + \epsilon)$ represented a small fraction of the simulated values of t^S, and the algorithm was iterated until at least 500 simulated values of t in this range were collected.

3. FITTING AN OVERDISPERSION MODEL

We first illustrate our method for a problem where the exact posterior density is straightforward to obtain. Mark McGwire has been the most prolific home run hitter in recent baseball history. In each of the five seasons 1995–1999, McGwire's observed rate in hitting home runs has remained in the 9–10% range, where the home run rate is defined as the number of home runs divided by the number of plate appearances. It is interesting to see if McGwire's pattern of home run hitting in this five-year period can be represented by a simple binomial model with a constant probability of success.

For each of the 698 games that McGwire played in this five-year period, we observe the number of plate appearances, m_i, and the number of home runs, x_i. If McGwire was indeed a streaky home run hitter, one might expect the home run counts to exhibit some clustering. In this streaky scenario, McGwire might hit a large number of home runs during some weeks, and have other weeks where he hits few or no home runs. To detect this type of clustering pattern, we group the data using a window of five games, which corresponds approximately to a week of games. Let y_i and

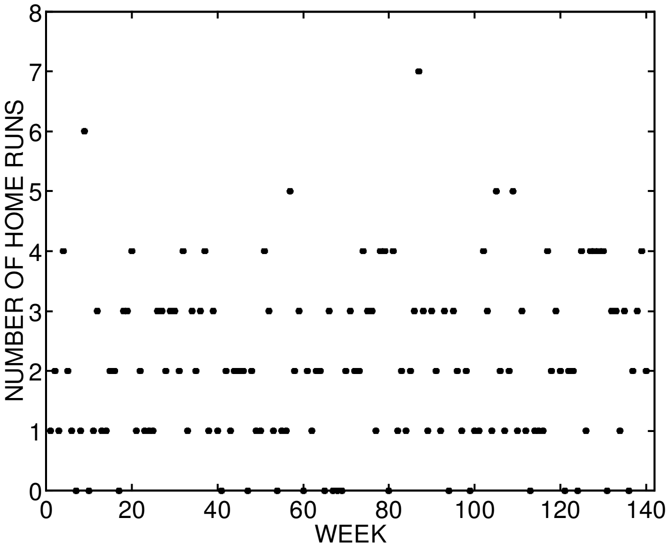

Figure 1. Graph of number of home runs hit by Mark McGwire in sequence of five-game periods during the seasons 1995–1999.

n_i denote, respectively, the number of home runs and the number of plate appearances in the ith group of five games, where $i = 1, \ldots, 140$. Figure 1 graphs the number of home runs y_i for McGwire as a function of the group number i.

A basic model for this data assumes that the counts $\{y_i\}$ are independently distributed from binomial distributions with respective sample sizes $\{n_i\}$ and a constant probability p. The maximum likelihood estimate for this data is $\hat{p} = \sum y_i / \sum n_i = .0976$. To assess the suitability of a binomial fit, home run data were simulated for 1,000 five-year periods using McGwire's actual plate appearances and the estimated binomial probability. Table 1 displays a frequency table of the observed counts $\{y_i\}$ and the mean simulated counts based on the binomial model. The third row of the table gives the Pearson residuals $\{(\text{observed} - \text{expected})/\sqrt{\text{expected}}\}$. We see some differences between the two sets of counts. Specifically, the numbers of observed periods with 0 and 4 home runs are somewhat higher, and the number of periods with 2 and 5 home runs are somewhat lower than what are predicted from the binomial model.

This brief analysis suggests that McGwire's home run counts may be more dispersed than predicted from a binomial model, and a binomial/ beta distribution will be used to model the possible overdispersion. Suppose that the counts y_i are independently distributed from binomial distributions with parameters n_i and p_i. We suppose that the probabilities $p_1, \ldots, p_{140}$ are a random sample from a beta distribution with mean η and precision parameter K.

$$g(p_i) = \frac{1}{B(K\eta, K(1-\eta))} p_i^{K\eta-1}(1-p_i)^{K(1-\eta)-1},$$

where $B(a, b)$ is the beta function. Uncertainty about the Beta hyperparameters (η, K) is modeled by means of a prior density $g(\eta, K)$.

In this binomial/ beta distribution, the hyperparameter η represents an average value of the probabilities $\{p_i\}$ and overdispersion is measured by the precision parameter K. As K approaches infinity, the model approaches a binomial model with common mean p. Here we place a uniform prior on the mean parameter η and assign K the proper, but vague, prior $(1 + K)^{-2}$.

A straightforward calculation shows that the posterior distribution of the hyperparameters (η, K) is given by

$$g(\eta, K|\text{data})$$
$$\propto \prod_{i=1}^{N} \frac{B(K\eta + y_i, K(1-\eta) + n_i - y_i)}{B(K\eta, K(1-\eta))} g(\eta, K),$$

where N is the number of periods. One makes inferences on the overdispersion parameter K by means of its marginal posterior

$$g(K|\text{data}) = \int_0^1 g(\eta, K|\text{data}) d\eta.$$

To learn about overdispersion using the model/ data simulation scheme, we first think of statistics that can be used to estimate the parameters η and K. Obvious choices for these statistics are the mean $\bar{y}$ and the standard deviation s of the y_i. The statistic s, in particular, measures the "larger than binomial" variation that we see in the observed counts. Then we simulate one set of models and data as follows:

1. (**Simulate models.**) Simulate values of the binomial probabilities $\{p_i\}$ from the prior model, by simulating values of (η, K) from the prior $g(\eta, K)$, and then simulating $\{p_i\}$ independently from a beta(η, K) distribution. Call the simulated probabilities $p_1^S, \ldots, p_N^S$.
2. (**Simulate data.**) Simulate data $y_1^S, \ldots, y_N^S$ independently from binomial distributions with respective sample sizes $n_1, \ldots, n_N$ and probabilities $p_1^S, \ldots, p_N^S$.
3. (**Compute statistics.**) Compute the mean $\bar{y}^S$ and the standard deviation s^S of the simulated data $\{y_i^S\}$.

We repeat Steps 1–3 of the algorithm a large number of times, obtaining the model/ data simulated sequences $(\eta^S, K^S, \bar{y}^S, s^S)$.

The mean and standard deviation values from the sample are $\bar{y}_{\text{obs}} = 2.029$ and $s_{\text{obs}} = 1.414$. We make inferences about the parameters (η, K) by restricting attention to the set of simulated statistic pairs $(\bar{y}^S, s^S)$, where $\bar{y}^S$ is in the interval $(\bar{y}_{\text{obs}} - .2, \bar{y}_{\text{obs}} + .2)$ and s^S is in the interval $(s_{\text{obs}} - .05, s_{\text{obs}} + .05)$.

To assess the accuracy of this method, Figure 2 shows the histogram of the simulated $\log K$ values and the exact marginal posterior of $\log K$ is overdrawn on the histogram.

Looking at Figure 2, the histogram of simulated values matches up well with the exact posterior density of $\log K$.

Table 1. Observed Five-Game Home Run Counts, Expected Counts Under a Binomial Model, and Pearson Residuals for McGwire Data

	Number of home runs						
	0	*1*	*2*	*3*	*4*	*5*	*6+*
Observed	19.0	38.0	33.0	28.0	17.0	3.0	2.0
Expected	16.9	37.5	39.5	26.6	12.9	4.8	1.9
Residual	.51	.08	−1.03	.27	1.14	−.82	.07

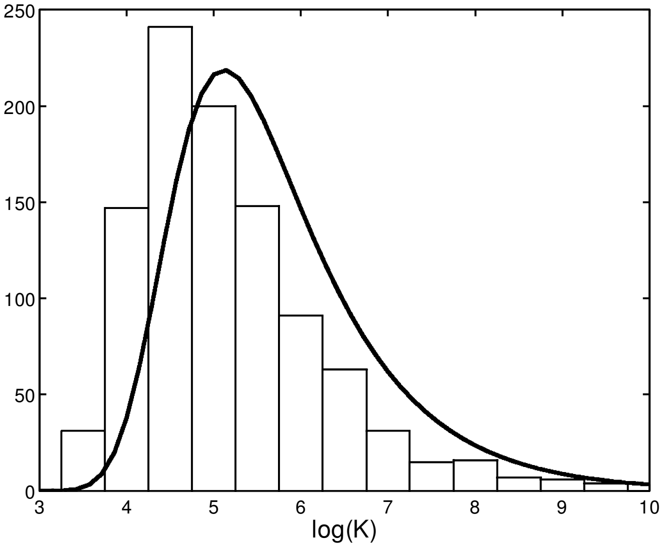

Figure 2. Posterior histogram of log K conditional on s in (1.364, 1.464) and ȳ in (1.829, 2.229). The posterior density of log K is drawn on top of the histogram.

The histogram is shifted to the left from the exact posterior. (The mean and standard deviation of the simulated draws are given by 5.17 and 1.10, respectively, which can be compared with the exact values of 5.69 and 1.09.) This behavior is due partly to the fact that the mean $\bar{y}$ and the standard deviation s are not sufficient for the parameters in this setting. Also, the fact that s is in the set $(s_{\text{obs}} - .05, s_{\text{obs}} + .05)$ is not as informative as knowing the exact value of s in the posterior. However, inferences based on the simulated sample will provide a good approximation to inferences based on the exact marginal posterior density.

Since the estimate of the beta hyperparameter K is large, there is little evidence that McGwire's recent pattern of home run hitting differs from a coin-tossing model. Suppose that one is interested in estimating McGwire's home run probability during the ith period p_i. In this binomial/ beta model, this probability can be estimated by

$$\hat{p}_i = \frac{\hat{K}}{n_i + \hat{K}}\hat{\eta} + \frac{n_i}{n_i + \hat{K}}\frac{y_i}{n_i},$$

where $\hat{\eta}$ is an estimate at McGwire's home run probability in recent years and $\hat{K}$ is an estimate at the beta precision parameter. Here $n_i \approx 20$ and $\hat{K} \approx \exp(5.69) = 296$, which indicates that the estimate at the ith period home run probability is approximately equal to the global estimate $\hat{\eta}$.

4. HOW STREAKY WAS JAVY LOPEZ IN 1998?

For a second example, consider the hitting data for Javy Lopez for the 1998 major league baseball season. We observe the number of hits and the number of at-bats for each of the 132 games he played during the season. Figure 3 graphs the number of hits and at-bats as a function of the game number.

Lopez was generally considered by the Atlanta Braves to be a streaky hitter during this season. In an Internet article published on April 17, 1998, Bill Zack commented that "If there's a more streaky hitter in baseball than Javier Lopez, the Braves don't want to meet him." The reason for this belief in streakiness was based on his poor and good performance in relatively short periods. At the time the article was published, Lopez was in a 1-for-17 streak. It was said in the article that Lopez can hit .400 one week and .200 the next week.

We will address three basic questions. First, what is a reasonable way of measuring the degree of streakiness or non-homogeneity in Lopez's hitting data? Second, what is a realistic "streaky model" that can represent the type of streakiness that we might expect for this baseball hitter? Last, given that we have found a suitable streaky statistic, what is the evidence provided by our statistic in support of the streaky model?

4.1 Construction of a Streaky Statistic

From the baseball article, it appears that the belief in streakiness is supported by the extreme performance of Lopez during short sequences of games. To look at this performance graphically, one can plot the moving batting average of Lopez against the game number. Figure 4 displays a moving average plot of Lopez's hitting using a window width of 10 games. A plotted point (a, b) represents the batting average b of Lopez in a neighborhood of ten games about the game number a. His batting average for the entire 1998 season is graphed using a dashed horizontal line. This graph shows interesting patterns. Lopez's hitting is pretty consistent for the first 40 games, he is hot for an interval about game 60, and then he oscillates between periods of poor hitting and good hitting at the end of the season.

One can quantitatively measure streakiness by using "interesting" statistics from this moving average plot. Specifically, streakiness is indicated by the following statistics.

- T_1 = difference between the largest and smallest moving average. Streakiness is indicated by a large value of T_1.

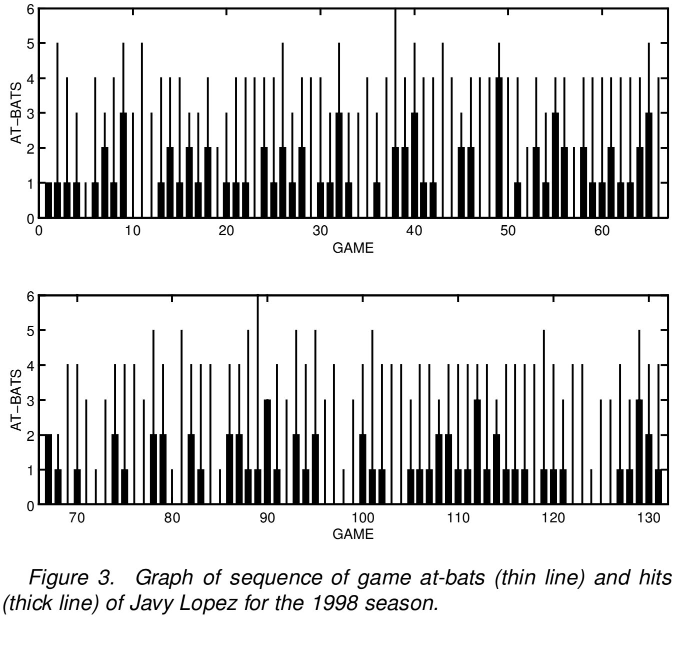

Figure 3. Graph of sequence of game at-bats (thin line) and hits (thick line) of Javy Lopez for the 1998 season.

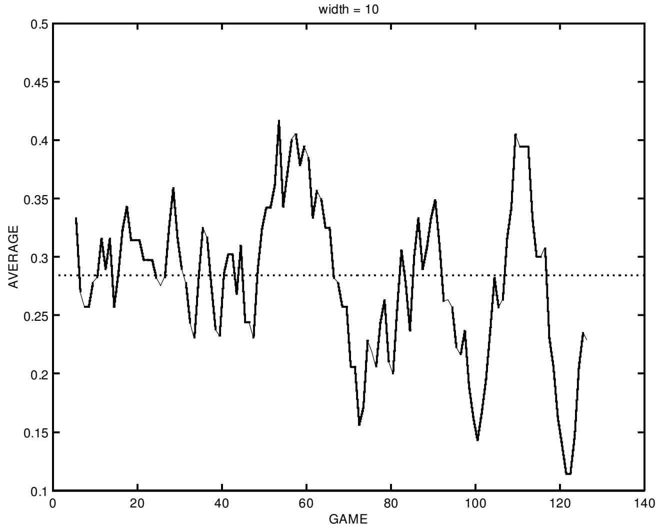

Figure 4. Graph of moving batting average of Javy Lopez using a window width of 10 games.

- $T_2 =$ sum of the absolute differences between the moving averages and the season batting average over all games. Large values of T_2 indicate streakiness with a large number of cold and hot spells.

Suppose that each game is classified as "hot" (1) or "cold" (0) if the corresponding moving average (corresponding to a window width w) is larger or smaller than the hitter's season average. Then the data are reduced to the following binary sequence and one can measure streakiness by interesting patterns in this sequence.

```
1000001110111111111000111100001
1000111010001111111111111111110
0000000000000001001111111000000
00000000011111111110000000000
```

We will consider the use of two statistics based on runs of hot or runs of cold in the binary sequence:

- $T_3 =$ number of runs (either runs of 0's or runs of 1's) in the sequence; and
- $T_4 =$ length of the longest run (either a run of 0's or a run of 1's) in the sequence.

One would expect streaky hitters to display a relatively small number of runs and occasionally have long runs of good or poor games.

A third set of statistics is based on the observed autocorrelation structure in the sequence of hitting data. If a hitter is streaky, then one would expect his hitting performance on a given day to depend on his hitting in the previous k games. We observe y hits in n at-bats in a particular game with hitting probability p. Suppose that the batting average of the player in the previous k games is given by x_k. The hitting probability of the present game can be modeled as a function of x_k using the logistic function

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_k.$$

We can use the slope estimate

$$T_5^{(k)} = \hat{\beta}_1$$

as an estimate of the dependence of the current day's hitting on the hitting on the previous k days. If this slope estimate is positive, then it provides some evidence that Lopez's hitting probability is higher when he has been successful in recent games.

The final statistic focuses on the homogeneity in the hitting probability of the player across games. Suppose that the games are subdivided by time into N groups, each of size b. If the true hitting probability does change over a season, one would expect significant variation in the observed group batting averages $\hat{p}_1, \ldots, \hat{p}_N$. So one measure of the non-homogeneity of the sequence would be

$$T_6^{(b)} = \text{standard deviation of} \\ \text{subgroup batting averages}\{\hat{p}_j\}$$

Table 2 lists values of these six streaky statistics for Javy Lopez's data.

Note that some of these statistics, by themselves, give some evidence about streakiness. The regression coefficients of the logistic regression on the previous five and ten games ($T_5^{(5)}, T_5^{(10)}$) are negative in sign, indicating a negative association between Lopez's hitting and his hitting in recent games.

4.2 A Streaky Model

Here we focus on the use of one particular model, the Markov switching model (Albert 1993), for describing streakiness, since it seems to be a reasonable representation of the streakiness that is discussed in the Internet article. Suppose that we represent Lopez's hitting as a sequence of independent Bernoulli trials with hitting probabilities that are game dependent. Suppose that Lopez has two possible hitting states for each game. He is either a hot hitter with hitting probability p_H, or a cold hitter with hitting probability p_C, where $p_H > p_C$.

To complete the model, we assume that Lopez switches between the hot and cold hitting states for different games according to a Markov chain. We suppose that, given he is in a hot or cold state in a game, the probability that he remains in the same state (either hot or cold) for the next game is a, and so the probability that he switches from cold to hot (or hot to cold) in the next game is $1 - a$. By choosing the staying probability a large, we are inducing

Table 2. Values of Six Streaky Statistics for Javy Lopez's Data

Name	*Value*
Range of moving averages	T_1 = .302
Sum of abs. diff. of moving averages	T_2 = 6.27
Number of runs	T_3 = 22
Length of longest run	T_4 = 18
Logistic regression slope	$T_5^{(5)}$ = −.861
Logistic regression slope	$T_5^{(10)}$ = −1.217
Standard deviation of subgroup bavg	$T_6^{(10)}$ = .0472
Standard deviation of subgroup bavg	$T_6^{(5)}$ = .1005

positive autocorrelation in the sequence of hitting probabilities across games, and this will induce streakiness in the observed hitting sequence.

4.3 Prior Distribution on Parameters of the Streaky Model

The Markov switching model is parameterized by three numbers: the hot and cold probabilities, p_C and p_H, and the staying probability a. The basic coin-tossing model (independent trials and constant probability of success) is a special case of the Markov switching model where $p_H = p_C$. In this situation, we let the staying probability a be equal to .5—this indicates that there is no propensity to stay or leave from the same hot or cold state. At the other extreme, "significant" streakiness implies that there will be a large difference d between the hot and cold probabilities and there is a large positive probability of staying in a given state. If we regard p as an average probability, we can parameterize this streaky model by $p_H = p + d/2, p_C = p - d/2$, and the staying probability will have a large positive value, say $a = a^*$.

To define a continuum of models between coin tossing and significant streakiness, we define a compromise model M_λ which is defined by parameters that are weighted averages of the parameters in the coin tossing and streaky models:

$$M_\lambda = \{p_C = p - \lambda d/2,\ p_H = p + \lambda d/2,\ a = .5 + \lambda(a^* - .5)\},\ 0 \leq \lambda \leq 1.$$

Note that M_0 corresponds to coin-tossing, M_1 corresponds to high streakiness, and the degree of streakiness is defined by the parameter λ. Values of λ close to one imply significant streakiness, and values of λ near zero are similar to independent coin tossing. We will call λ the streaky parameter, since it defines the distance away from coin-tossing towards the streaky alternative hypothesis.

In this example, we first define what we feel is significant streakiness for Lopez. We believe that Lopez is indeed streaky if his hot and cold probabilities are 100 points higher and lower, respectively, from his overall batting average .284 (i.e., $d = .1$). In addition, we view Lopez as very streaky if his staying probability is equal to $a^* = .8$. To complete the model, we place a uniform prior on the compromise parameter λ, reflecting little prior information about where Lopez falls on the streakiness continuum. This prior could easily be adjusted to model alternative beliefs about streakiness. For example, a strong belief in coin-tossing behavior would be represented by a prior that placed a large probability about the value $\lambda = 0$.

Table 3. Correlation between simulated values of the streaky parameter λ and the statistic T for 8 choices of streaky statistics

Name	*Correlation*
T_1	.30
T_2	.33
T_3	−.14
T_4	.07
$T_5^{(5)}$	.22
$T_5^{(10)}$	.16
$T_6^{(10)}$	.28
$T_6^{(5)}$	.34

4.4 Simulating Models and Data

Suppose that we are interested in doing inference about the streakiness parameter λ based on a particular streaky statistic T. We perform the model/ data simulation as follows:

1. We simulate a value of λ, λ^S, from the uniform prior. This value of λ defines values of the hot and cold probabilities and the staying probability. Call these simulated values p_C^S, p_H^S, and a^S.

2. Based on the simulated values p_C^S, p_H^S, a^S, we simulate the states (hot and cold) for the hitter for all of his games. This is done by flipping a coin to decide the state of the first game and then simulating the Markov chain to get the states of the remaining games. These states determine the set of hitting probabilities for all games.

3. We simulate binomial data based on the set of hitting probabilities.

4. We compute the value of the streaky statistic T from the simulated binomial data.

We performed this model/ data simulation for 10,000 iterations. In Step 4 of the algorithm, we computed values of each of our proposed streaky statistics $T_1, \ldots, T_6$. When we plotted values of the simulated model/ data pairs (λ^S, T_i^S) for each i, we found that the parameter and statistic were approximately linearly related. So we can summarize the relationship between the streaky parameter and the streaky statistic by means of the correlation of the model/ data pairs. Looking at the correlations in Table 3, we see that there is a wide range of values. Values of the correlation that are close to zero correspond to statistics that are not informative about the value of the streaky statistic λ. In contrast, large absolute values of the correlation correspond to statistics that are useful in learning about the degree of streakiness in the Markov switching model.

To help interpret the size of the correlations in Table 3, consider the simple binomial/ beta situation, where data y is binomial(n, p), and p has a beta distribution proportional to $p^{K\eta-1}(1-p)^{K(1-\eta)-1}$, as used in Section 3. Here the sample proportion $\hat{p} = y/n$ is used to learn about the probability p. A straightforward calculation gives that the correlation between $\hat{p}$ and p is

$$\mathrm{corr}(\hat{p}, p) = 1/\sqrt{1 + K/n}.$$

Suppose that a uniform prior is chosen for p where $K = 2$. In this case, the correlation between $\hat{p}$ and p increases rapidly as a function of n—the correlation is equal to .85, .91, and .95 for $n = 5, 10, 20$, respectively. So the correlations between statistics and parameter in Table 3 appear relatively small to those in the binomial/ beta model.

We first look at the statistic T_3, the number of runs. The correlation of T_3 and λ is a small negative number, indicating that this statistic is not especially helpful in detecting streakiness. This point is amplified by Figure 5 which dis-

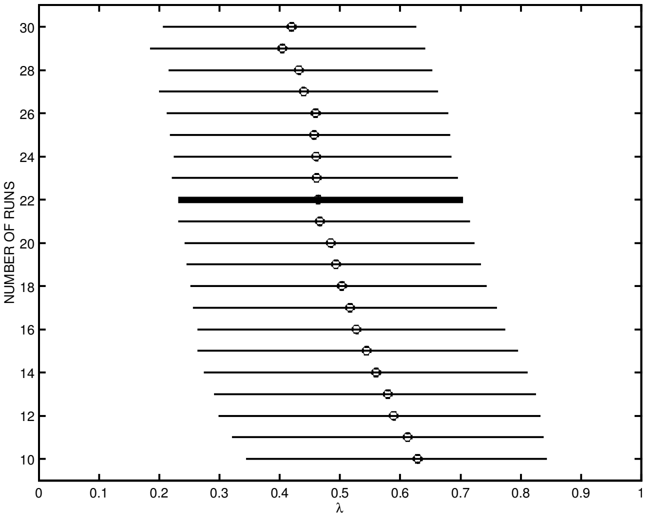

Figure 5. Posterior median and quartiles of streakiness parameter λ given value of number of runs statistic T_3.

plays the posterior median and quartiles of λ for a range of values of T_3. Note that the location of the middle 50% of the posterior distribution barely changes as the number of runs in the data increases. For Lopez's data, the number of runs is 22, and the posterior 50% probability interval of (.23, .70) is shown by a bold line. This interval is similar in location to the 50% prior probability interval of (.25, .75), which says that we have learned little about streakiness from this particular statistic.

Figure 6 displays posterior probability intervals for a more useful statistic, $T_6^{(5)}$, the standard deviation of subgroup batting averages. In contrast to Figure 5, the posterior median of λ increases as a function of the standard deviation, which shows that the subgroup standard deviation is informative for detecting streakiness. In our example, Lopez's value of $T_6^{(5)} = .1005$. The location of the 50% posterior probability, indicated on the figure by a bold line, is (.21, .67), which can be compared to the prior probability interval of (.25, .75).

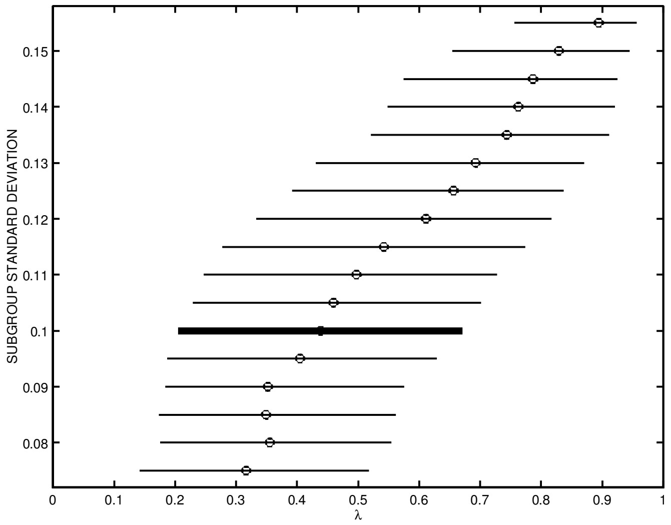

Figure 6. Posterior median and quartiles of streakiness parameter λ given value of standard deviation statistic $T_6^{(5)}$.

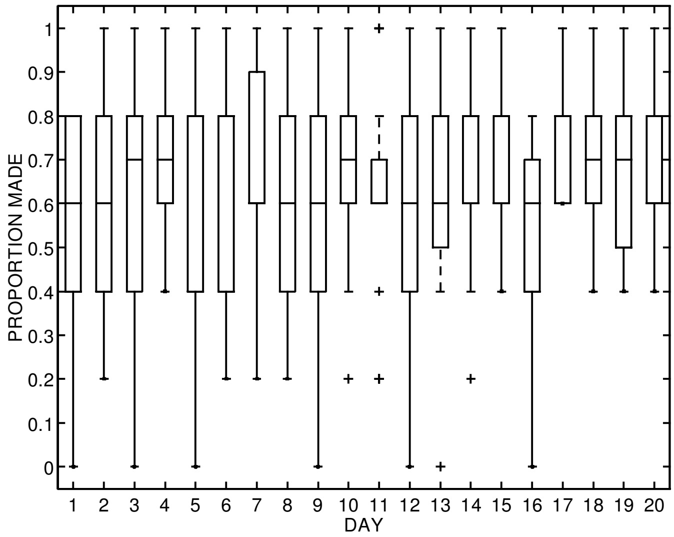

Figure 7. Boxplots of proportions of made shots grouped by practice number.

Was Lopez a streaky hitter in 1998? Using both of the statistics T_3 and $T_6^{(5)}$, the posterior distribution for the streaky parameter λ was essentially the same as the prior distribution, indicating that the data have provided little evidence in support of or against streakiness. In other words, none of the streaky statistics proposed here appear to provide any support to the belief that Lopez was a streaky hitter during this season.

5. DETECTING STREAKINESS IN FREE-THROW SHOOTING DATA

Wardrop (1999) described a controlled experiment to detect streakiness in basketball shooting. A college player took a sequence of shots from beyond the three-point line for 20 practices during the summer. On each of the 20 practices, the player attempted 100 shots and the results of the shots were recorded. The data can be represented as a matrix of 20 rows and 100 columns of 0's and 1's, where the rows correspond to practices, and a data value of 1 indicates a made shot and 0 indicates a miss.

These data resemble the baseball hitting data discussed in the earlier section and one could fit a Markov switching model to represent streakiness. We will instead fit a dispersion model to these data to illustrate the generality of the algorithm of Section 2.

If the player is indeed streaky, then one would expect the data to be somewhat clustered. That is, the player should have a larger number of cold and hot spells in shooting than what would be predicted from a binomial model. As an initial investigation, we group the 100 shots for each practice into 20 groups of size 5. For each group, the proportion of made shots is recorded. The data are summarized by the proportions $\{\hat{y}_{ij}\}$, where i indicates the practice number and j denotes the group within practice ($1 \leq i \leq 20, 1 \leq j \leq 20$). Figure 7 displays boxplots of the proportions $\{\hat{y}_{ij}\}$ grouped by practice number. Looking at the graph, we see that the spread of the boxplots tends to decrease over time, indicating that the player is becoming more consistent in her shooting performance.

Consider a set of group proportions for a given practice, say $\hat{y}_{i1}, \ldots, \hat{y}_{i20}$. If the shots were independent Bernoulli trials with constant probability p, then the group proportions would be independent with common mean p and standard deviation $\sqrt{p(1-p)/5}$. This binomial standard deviation can be estimated by $\sqrt{\hat{p}(1-\hat{p})/5}$, where $\hat{p}$ is the proportion of made shots for this practice.

We can assess this binomial assumption by comparing the actual standard deviation of the proportions $\{\hat{y}_{ij}\}$ with the estimated binomial standard deviation by means of a ratio

$$\frac{\text{standard deviation}\{\hat{y}_{ij}\}}{\sqrt{\hat{p}(1-\hat{p})/5}}.$$

Suppose we express the actual standard deviation as

$$\text{standard deviation}\{\hat{y}_{ij}\} = \sqrt{\hat{p}(1-\hat{p})/n_i^o},$$

where n_i^o is the "observed binomial sample size." Then the above ratio of standard deviations is equivalent to the ratio of sample sizes

$$\frac{\text{standard deviation}\{\hat{y}_{ij}\}}{\sqrt{\hat{p}(1-\hat{p})/5}} = \sqrt{\frac{5}{n_i^o}}.$$

For a particular practice, the statistic n_i^o measures how close the data resemble a binomial model. A value of n_i^o smaller than 5 indicates data that show more dispersion than predicted under the binomial model (called overdispersion), and a value of n_i^o larger than 5 indicates underdispersion.

Figure 8 plots the values of the logarithms of the sample size statistics, $\log n_i^o$, against the practice number i. The horizontal line at the value $\log 5$ is shown for reference. After some initial instability for practices 1–4, note that the sample sizes generally increase as a function of the practice number. This indicates that the data are generally overdispersed for early practices and underdispersed for later practices. A least-squares regression on the data for practices 5–20 gives the fit

$$\log n^o = 1.04 + .0447 \times \text{ (practice number)}.$$

One way to confirm this visual impression is to separately fit least-squares lines to the two datasets corresponding to practices $1, \ldots, c$, and practices $c+1, \ldots, 20$, respectively, and then choose the value of c which minimizes the sum of squared residuals. Using this procedure, we confirm that $c = 4$ is a suitable breakpoint for these data.

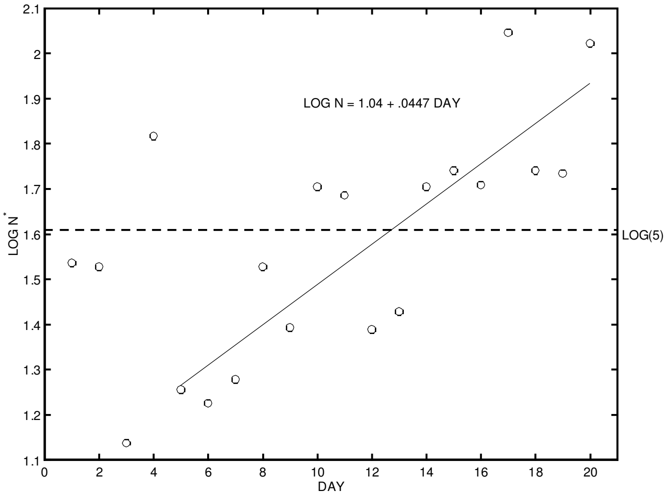

Figure 8. Plot of logarithms of sample sizes n_i^o as function of practice number. A least-squares line to the data for practices 5–20 is displayed.

The observed variation in spread that we see from Figure 8 motivates the consideration of a family of distributions that can accommodate more or less spread than a binomial. A convenient choice is the quasi-binomial density. [A general class of quasi-likelihood densities was described by McCullagh and Nelder (1989, chap. 9).] If y denotes the proportion of successes in n trials, then the quasi-binomial density with parameters n, p, n^* is given by

$$f(y|n, p, n^*) = \frac{C\Gamma(n^*+1)}{\Gamma(n^*y+1)\Gamma(n^*(1-y)+1)} p^{n^*y}(1-p)^{n^*(1-y)}, \quad y = 0, 1/n, 2/n, \ldots, 1,$$

where C is a constant to ensure that the density sums to one. When $n^* = n$, the density reduces to the binomial density. A value of $n^* < n (n^* > n)$ gives a density with greater (smaller) variability than the binomial.

We can model a pattern in the spread of the proportions by means of a regression model on the quasi-binomial sample sizes. Let y_{ij}, the proportion of shots made in the jth subgroup of the ith practice, have a quasi-binomial density with parameters $(5, p_i, n_i^*)$, where n_i^* measures the spread of the group proportions. Assume, for generality, that there is an initial period of practice time, $i \le c$, where the dispersion parameters follow the regression relationship

$$\log n_i^* = \beta_0^{(1)} + \beta_1^{(1)}(i - I^{(1)}),$$

where i is the practice number $(1 \le i \le c)$ and $I^{(1)}$ is the mean of the practice numbers. Likewise, after the practice time is over $(i > c)$, the dispersion parameters follow the new model

$$\log n_i^* = \beta_0^{(2)} + \beta_1^{(2)}(i - I^{(2)}).$$

To complete the model, prior distributions need to be placed on the success probabilities $p_1, \ldots, p_N$, the regression parameters $\beta_0^{(j)}, \beta_1^{(j)}, j = 1, 2$, and the breakpoint c. We suppose that the probabilities of making a shot can vary between practices and we assume that the p_i are a random sample from a $\text{beta}(26, 14)$ distribution. This prior assumes that the shooting probabilities fall, with probability .95, in the interval (.52, .77). We assume $\beta_0^{(1)}, \beta_1^{(1)}, \beta_0^{(2)}, \beta_1^{(2)}$ independent with each assigned a normal distribution with mean 0 and standard deviation .1. This prior is relatively noninformative and allows for increasing and decreasing relationships between the dispersion parameter and the practice number. Finally, we assign a uniform prior for the breakpoint c on the times $\{0, 1, \ldots, 9\}$, reflecting little prior knowledge about the length of the practice time.

One iteration of the model/data simulation for this example proceeds as follows.

1. Simulate a breakpoint c^S from the uniform prior on $\{0, 1, \ldots, 9\}$.
2. Simulate regression parameters $\beta_0^{(1)S}$, $\beta_1^{(1)S}$, $\beta_0^{(2)S}$, $\beta_1^{(2)S}$, and success probabilities $\{p_i^S\}$ from the above prior distributions.
3. Compute dispersion parameters $\{n_i^{*S}\}$ using the regression relationships.
4. Simulate group proportions $\{y_{ij}^S\}$, where y_{ij}^S is simulated from the quasi-binomial$(5, p_i^S, n_i^{*S})$ distribution.
5. Compute the estimated sample sizes $\{n_i^{oS}\}$, where as described earlier, n_i^{oS} is found by computing the ratio of the standard deviation of the $\{y_{ij}^S\}$ to the estimated binomial standard deviation.
6. Regress the log n_i^{oS} on the practice numbers using the procedure described earlier. Two separate least squares lines are fit using all possible breakpoints, and the optimal breakpoint $\hat{c}^S$ is found which minimizes the sum of squared residuals. A regression slope estimate $b_1^{(2)S}$ is found for the later practices using this optimal regression procedure.

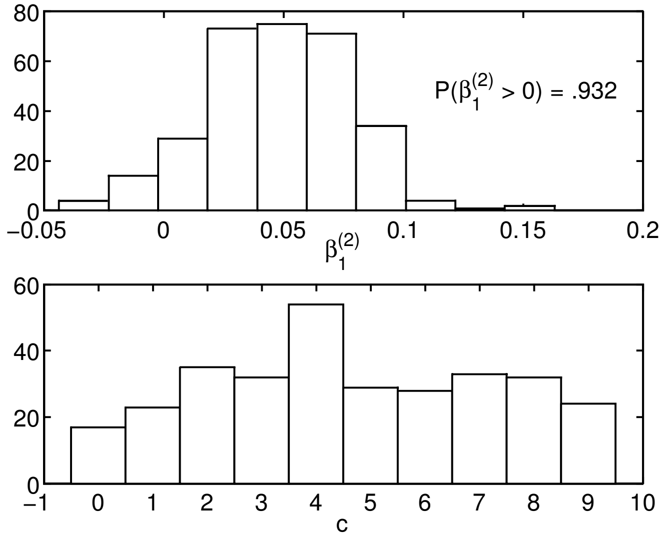

Figure 9. Histogram of simulated values of true breakpoint c and regression slope $\beta_1^{(2)}$ conditional on $\hat{c}$ = 4 and $b_1^{obs} \in (.0447 - .01, .0447 + .01)$.

This algorithm was repeated 1,000 times, resulting in a simulated sample of model/data sequences $(c^S, \beta_1^{(2)S}, \hat{c}^S, b_1^{(2)S})$. For the observed data, we observe $\hat{c} = 4$ and $b_1^{(2)} = .0447$. To perform inference about the slope parameter $\beta_1^{(2)}$, we focus on the simulated values $\beta_1^{(2)S}$ conditional on $\hat{c}^S = 4$ and the estimated slope $b_1^{(2)S)}$ in a small interval centered about the observed value .0447.

Figure 9 displays histograms of the posterior distributions of the regression slope $\beta_1^{(2)}$ (top) and the true breakpoint c (bottom) that are based on the simulated values produced using the above algorithm. Note that the histogram of the true breakpoint c is relatively diffuse, indicating that we have learned little about the location of the true breakpoint in these data. However, looking at the location of the histogram, we see that the posterior density of $\beta_1^{(2)}$ is centered about .05 and most of the probability lies on positive values. (The probability that the true slope exceeds 0 is approximately 93%.) Thus, there appears to be good evidence that the quasi-binomial sample size increases as a function of the practice number. In other words, the true variation in the shooter's day performance from practices 5–20 appears to decrease over time, which means that the shooter performance is more consistent with repeated practice. Wardrop (1999), using methods that did not involve grouping data, also found significant evidence against coin-tossing in these data.

6. CLOSING REMARKS

The model/data simulation algorithm has several desirable features. First, when a complete Bayesian analysis is difficult to implement and there exists a statistic t which is believed to summarize most of the information about the parameter θ, this simulation algorithm provides a convenient way to obtain an approximate Bayesian analysis. Second, the examples illustrate that this algorithm allows much flexibility in the choice of alternative models, which is especially important in the detection of streakiness. Last, one obtains simulated ordered pairs of (parameter, statistic), and the relationship that one finds in these simulated pairs is instructive on how one learns about the parameter based on the statistic. We saw in Section 4 that the observed correlation coefficient between a streaky statistic and the streakiness parameter led to a better choice of a statistic to detect streakiness.

Although this methodology appears potentially useful in the detection of streakiness, the first two examples demonstrate that coin-tossing is a good model for data that appear streaky. There is little evidence for nonstationarity in McGwire's home run data and, although Javy Lopez might appear streaky in his pattern of hitting across games, the Markov switching model was unable to pick up any significant streakiness. The basketball data seem to deviate the most from coin-tossing. In this example, streakiness was manifested by a change in the variation of shooting behavior across time. This methodology is flexible to allow for innovative ways of detecting streakiness, such as illustrated in Section 5, and in the construction of plausible models to explain the observed streakiness.

[Received July 1999. Revised March 2000.]

REFERENCES

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

——— (1997), "Teaching Bayes's Rule: A Data-Oriented Approach," *Journal of the American Statistical Association*, 51, 247–253.

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

Barry, D., and Hartigan, J. A. (1993), "Choice Models for Predicting Division Winners in Major League Baseball," *Journal of the American Statistical Association*, 88, 766–774.

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.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models*, London: Chapman and Hall.

Stern, H. S. (1997), "Judging Who's Hot and Who's Not," *Chance*, 10, 40–43.

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

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.