Statistical Tests for the Hot-Hand in Basketball in a Controlled Setting

Robert L. Wardrop

March 1, 1999

Abstract

When I watch basketball on television, it is a common occurrence to have an announcer state that some player has the *hot-hand*. This raises the question: Are Bernoulli trials an adequate model for the outcomes of successive shots in basketball? This paper addresses this question in a controlled (practice) setting. A large simulation study examines the power of the tests that have appeared in the literature as well as tests motivated by the work of Larkey, Smith, and Kadane (LSK). Three test statistics for the null hypothesis of Bernoulli trials have been considered in the literature; one of these, the runs test, is effective at detecting one-step autocorrelation, but poor at detecting nonstationariy. A second test is essentially equivalent to the runs test, and the third is shown to be worthless. The LSK-motivated tests are shown to be effective at detecting nonstationarity. Finally, a case study of 2,000 shots by a single player is analyzed. For this player, the model of Bernoulli trials is inadequate.

KEY WORDS: Bernoulli trials, the hot-hand, power, simulation study, case study.

1 Introduction

In this paper I consider a statistical analysis of basketball shooting in a controlled (practice) setting, with special interest in the hot-hand. In Section 2, I review and critically examine the two seminal papers on this topic: Gilovich, Vallone, and Tversky (GVT) [5], and Tversky and Gilovich (TG1) [10]. A simulation study of power is presented in Section 3. Finally, in Section 4, a case study of 2,000 trials is analyzed.

In GVT and TG1, three additional topics appear which are beyond the scope of this paper:

1. Modeling game free throw shooting,

2. Modeling game shooting, and

3. Opinions and misconceptions of fans.

Readers interested in the first of these topics should refer to Wardrop [12] for a further analysis of the free throw data from the papers.

Several researchers have considered the second topic; the interested reader is referred to Larkey, Smith, and Kadane (LSK) [7], Tversky and Gilovich (TG2) [11], Hooke [6] and Forthofer [4]. For related work in baseball, see Albright [1], Albert [2], Stern and Morris [9] and Stern [8]. Topic 3 is considered briefly in Section 2 of this paper.

Finally, readers interested in statistical research in sports are referred to Bennett [3]. The chapters on basketball and baseball should prove to be of special interest to readers of this paper.

2 Review of Literature

GVT appeared in 1985 in a "psychology journal." Four years later the same research was restructured as TG1 and appeared in a "statistics journal." For the most part the papers present identical analyses and interpretations of the data, with the earlier paper generally providing more detail.

Twenty-six members of the Cornell University varsity and junior varsity basketball teams generated the data that are examined. The players are labeled M1 (for male one) through M14, and F1 through F12. Each player provided two sequences of shot attempts: the *shooting data* and the *prediction data*. I will begin with an examination of the shooting data.

The plan was for each player to provide a sequence of 100 shots, but three of the players, M4 (90 shots), M7 (75), and M8 (50), fell short of the target number.

Twenty-six null hypotheses are tested; namely, for each player the null hypothesis is that his or her shots satisfy the assumptions of Bernoulli trials. Below is one summary of the data obtained by M9.

Previous	Current Shot		
Shot	S	F	Total
S	38	15	53
F	16	30	46

The researchers describe these data in two ways. First, note that M9 made 72 percent of his shots after a hit, but only 35 percent after a miss; a difference of 37 percentage points. Second, the researchers compute the serial correlation and obtain 0.37.

The researchers analyze each player's data with three test statistics. The first two are a test of the serial correlation and the runs test. They summarize their findings as follows.

> With the exception of one player ($r = 0.37$) who produced a significant positive correlation (and we might expect one significant result out of 26 just by chance), both the serial correlations and the distributions of runs indicated that the outcomes of successive shots are statistically independent.

Their third test is a test of fit and the researchers refer to it as a test of stationarity. The test is nonstandard, but simple to describe. Suppose that the data are

1100100011110101 . . . .

Group the data into sets of four,

1100 1000 1111 0101 . . . ,

and count the number of successes in each set,

2, 1, 4, 2 . . . .

Use the 25 counts to test the null hypothesis that the data come from a binomial distribution with $n = 4$ and p estimated as the proportion of successes obtained in the data. The first difficulty with implementing this test is that typically one or more of the expected counts is quite small. The researchers overcame this problem by combining the O's and E's to yield three response categories: fewer than 2, 2, and more than 2, and then applied a χ^2 test with one degree of freedom. The test can be made one-sided by rejecting if and only if the χ^2 test would reject at 0.10 and $E > O$ for the middle category (corresponding to two successes). The rationale for this decision rule is that $E > O$ in the central category indicates heavier tails, which implies more streakiness. The theoretical basis for this test is shaky, but the simulation study reported in Section 3 suggests that its probability of type 1 error might be close to its nominal level. It is unclear whether the researchers apply the test of fit as a one- or two-sided test. In any event, the researchers apply their test of fit to their 26 sets of data and report,

> The results provided no evidence for departures from stationarity for any player but (M)9.

The researchers describe the shooting data in two other ways. First, they explore the possibility that the past influences the present only after two consecutive successes or failures. For example, this approach yields the following data for M9.

Previous	Current Shot		
Two Shots	S	F	Total
SS	30	8	38
FF	10	20	30

Note that M9 shot better after two hits than after two misses by 46 percentage points. Second, they condition on the three previous shots, and again M9 is most extreme yielding the data shown below.

Previous	Current Shot		
Three Shots	S	F	Total
SSS	25	5	30
FFF	7	13	20

The researchers do not perform any tests for these two presentations of the data. Note, in addition, that the researchers do not describe their data by computing the longest streak of successes, nor do they perform any tests based on the longest streak of successes.

Next, I will criticize the above analysis.

Performing a test of serial correlation and a runs test is largely redundant. For the 26 players studied, the correlation coefficient for a player's serial correlation and standardized number of runs is -0.993. (Two notes: Positive serial correlation corresponds to fewer runs than expected and a negative z; hence, the correlation is negative. While performing this computation I discovered two misprints in Table 4 of GVT. First, the entry for player M10 for P(hit$|$1 hit) should read 0.58(60). Second, the serial correlation for F12 is -0.070.) Thus, it is incorrect to believe that performing these two tests gives the data "two chances" to reveal that the Bernoulli trials model is inadequate. Throughout the remainder of this paper I will focus on the runs test and disregard the test of serial correlation. The simulation study reported in Section 3 reveals that the runs test has some ability to detect autocorrelation, but is poor at detecting nonstationarity.

In GVT, the test of fit is introduced under the heading, "Test of Stationarity." In TG1, it is described as "A more sensitive test of stationarity." The simulation study reported in Section 3 indicates that the test of fit is, in fact, abysmally poor at detecting any but the most extreme form of nonstationarity. In addition, for every alternative examined in Section 3, other simple tests are far superior to the test of fit. To put it bluntly, the test of fit should not be used.

Thus, my first conclusion is that the researchers used only one (distinct and possibly effective) test statistic, the runs test. As noted earlier the researchers find one significant result and note that, "We might expect one significant result out of 26 by chance." I will now explain why I consider this quote to be an incomplete description of their findings.

The significant result was obtained by M9 and his exact one-sided P-value for the runs test is 0.000044. Having noted that only one P-value was smaller than 0.05, would it not have been fair to mention that one out of 26 P-values was smaller than 1 in 20,000? Is this result not a bit surprising to one who believes in the omnipresence of Bernoulli trials?

Throughout the two papers the researchers refer to the alternative as being "the hot-hand." For the two test statistics considered, the runs test and the test of fit, the hot-hand alternative, whether it means autocorrelation or nonstationarity, is naturally a one-sided alternative and should therefore, in my opinion, have a one-sided test. The researchers acknowledge this "one-sidedness" repeatedly; for example, in TG1 they refer to a negative serial correlation as coming from data that, "Run counter to the streak-shooting hypothesis."

The one-sided versus two-sided debate has a long history and is not going to be settled here. I believe it is worth noting, however, that in addition to M9 two other players had one-sided P-values below 0.05 and two others had values slightly above 0.05. In particular, M3 shot better after a hit than after a miss by 18 percentage points, and gave an exact one-sided P-value of 0.0375. Similarly, M6 shot better after a hit than after a miss by 17 percentage points, and gave an exact P-value of 0.0403. Player F3 shot better after a hit than after a miss by 16 percentage points, and gave an exact P-value of 0.0680. Player M7 attempted only 75 shots and was better after a hit than after a miss by 15 percentage points; but with so little data the runs test is not statistically significant: the exact one-sided P-value is 0.0636. Disregarding M8 because he took only 50 shots, we find that:

- One of 25 P-values is extremely small, and
- Three of 25 P-values are smaller than 0.05 and two others are only slightly larger than 0.05.

This is not overwhelming evidence in support of the hot-hand theory, but it is equally wrong to interpret these data as indicating, as the researchers write in TG1, "People ... tend to detect patterns even where none exist." In addition, as shown in Section 3, the runs test is poor at detecting nonstationarity. Who can say how many of these 25 players exhibited some form of nonstationarity?

There is, of course, an advantage to using only one test statistic; namely, one test statistic makes it relatively easy to keep track of the probability of a type 1 error. If, however, the analyst guesses wrong about the form of the alternative, then performing the test is, at best, a waste of time, and, at worst, misleading.

How is one to decide on an alternative? The definitive answer can be obtained only by examining the performances of a great number of basketball players. Such an examination might show that departures from Bernoulli trials are:

- Almost always in the form of autocorrelation,
- Almost always in the form of nonstationarity,
- Frequently in the form of autocorrelation and frequently in the form of nonstationarity,
- So rare and so minor as to be unworthy of further attention, or
- Some other pattern.

Until such data are available, in addition to keeping an open mind one might choose to be guided by the opinions of experts (in basketball, not statistics!).

The researchers preface TG1 with a quote from professional basketball player Purvis Short:

> You're in a world all your own. Its hard to describe. But the basket seems to be so wide. No matter what you do, you know the ball is going to go in.

The researchers then write,

> This statement describes a phenomenon known to everyone who plays or watches the game of basketball, a phenomenon known as the "hot hand."

Clearly, Short is describing an occasional phenomenon and is not describing anything as omnipresent as lag one autocorrelation. Unfortunately, the researchers appear to be confused about the distinction between autocorrelation and nonstationarity. For example, immediately after the previously quoted statement, they write,

> The term refers to the putative tendency for success (and failure) in basketball to be self-promoting or self-sustaining.

In others words, they are saying that Short's description of nonstationarity is the hot-hand which in turn is autocorrelation! Later they write,

> Do players occasionally get a "hot hand"?

This question now suggests that the researchers acknowledge that the hot-hand might be nonstationarity rather than autocorrelation.

The concluding section of TG1 is titled, "The Hot Hand as Cognitive Illusion," and it contains further evidence of the researchers' confusion about alternatives. They write,

> Naturally, every now and then, a player may make, say, nine of ten shots, and one may wish to claim—after the fact—that he was hot. Such use, however, is misleading if the length and frequency of such streaks do not exceed chance expectation.

Nowhere in either paper do the researchers analyze their data searching for streaks of successes or streaks of k successes in $k+1$ shots. But in this passage they seem to be suggesting that their data contained no such unusual streaks. In a later paper, TG2, the researchers report,

> [In] Our previous analyses . . . We found . . . the frequency of streaks of various lengths was not significantly different from that expected by chance.

This statement puzzles me because I found no evidence in GVT or TG1 that the researchers examined the lengths of streaks of successes.

The following revealing statement also appears in TG2.

> Many observers of basketball believe that the probability of hitting a shot is higher following a hit than following a miss, and *this conviction is at the heart of the belief in the "hot hand"* (emphasis added).

The belief in autocorrelation is certainly not at the heart of my belief in the hot-hand. My belief is that on those somewhat infrequent occasions when the Bernoulli trials model is inadequate, nonstationarity is much more common a phenomenon than autocorrelation. If Mr. Purvis Short spoke in the language of statisticians, he would certainly say that he believes in nonstationarity rather than autocorrelation.

I do acknowledge that for persons who have studied neither probability nor statistics carefully, it is easy to be confused about differences between autocorrelation and nonstationarity. My generosity to such persons, however, is strained when their extremely incomplete data analysis is followed by a statement that anyone who disagrees with them is suffering from a "Cognitive Illusion."

Lest I be charged with ignoring the researchers "survey data," let me turn to that issue. The researchers asked a convenience sample of 100 "avid basketball fans" to consider a hypothetical player who shoots 50 percent from the field. Each fan was asked two questions about this player.

- 1. (2.) What is your estimate of his field goal percentage for those shots that he takes after having just made (missed) a shot?

The mean of the responses to questions 1 and 2 were 61 percent and 42 percent, respectively.

In contrast to Short's statement, these fans are describing autocorrelation. There are three points, however, that makes Short's description more compelling than the opinion of the fans. First, he is the expert and his opinion should be better informed than those of the fans. Second, the Short quote appears to be a "free response," whereas the fans' opinion is largely a result of the bias, perhaps unconscious, in the research method. For example, what would the results have been if the fans had been presented with the following two-part question?

> Consider a collection of hypothetical players whose "usual ability" is to have a probability of 50 percent of making an 18 foot jump shot.
>
> 1. What percentage of such players occasionally "get hot" and perform at a level above his or her usual ability?
> 2. Consider a player who is typical of those who occasionally "get hot." What is your estimate of his or her probability of making an 18 foot jump shot when "hot?"

If the researchers had substituted this two part question for their questions, they might well have reached the conclusion that, "A belief in nonstationarity is at the heart of the belief in the hot-hand."

Third, the fans' response is so ludicrous that it does not merit further consideration. The practical difference between 61 percent and 42 percent shooters is *huge*. It is absurd to believe that the *norm* is for a player to be constantly fluctuating between two states—a great shooter and a horrible shooter.

Finally, the researchers' interpretation of results is extremely biased. If the null hypothesis of Bernoulli trials fails to be rejected, they interpret this as lack of evidence for the hot-hand. Fair enough, power considerations aside. But data that reject a null hypothesis are disdained by the researchers also, in three ways:

1. Differences are not important unless they exceed in magnitude the 19 percentage points difference obtained by questioning the fans. (See the discussion of the prediction data below. In particular, the researchers state that the difference between 60 and 40 percent shooters is “small.”)

2. Differences are not important unless they match the omnipresence expressed by the fans. In a criticism of LSK, the researchers write in TG2,

   > As our survey shows, it is widely believed that the hot hand applies to most people. . . . Because LSK’s entire argument is based on the performance of a single player, we could rest our case right there.

   (Note: In addition to revealing a biased approach to data analysis, this quote is an extremely unfair evaluation of LSK. LSK did not conduct hypothesis tests; the goal of the research was descriptive; they wanted to find the streakiest among several players. Also note that their convenience sample of 100 basketball fans is interpreted as showing that “. . . it is widely believed.”)

3. After rejecting the null hypothesis of Bernoulli trials, the researchers decide that the alternative really is not the hot-hand, but a tendency to “try harder.” (See my discussion of the prediction data below.)

In short, the researchers view the fans’ opinion as the gold standard; I see it as a straw man that, despite whatever data might be obtained, allows them to continue to proclaim the folly of all who believe that perhaps on occasion basketball is more complex than Bernoulli trials.

In the remainder of this section I will examine the researchers’ prediction data.

Return to the Short quote. Notice that he is describing how he feels. He does not say, for example,

> When I analyze lengthy records of my shooting data I detect patterns. Therefore, I conclude that the null hypothesis of Bernoulli trials should be rejected.

The researchers realize the difference between data analysis and feelings. In GVT they write,

> There is another cluster of intuitions about “being hot” that involves predictability rather than sequential dependency. If, on certain occasions, a player can predict a “hit” before taking a shot, he or she may have a justified sense of being “hot” even when the pattern of hits and misses does not stray from chance expectation.

They investigate this notion with the prediction data.

Each player attempted 100 shots from a location at which he or she was believed to be about a 50 percent shooter. Before each shot a player would bet *high* or *low* and was advised to bet high if and only if he or she felt confident about the pending attempt. A success on a high bet would earn the shooter five cents, while a miss would cost four cents; for a low bet the values were two and one cents, respectively. The data analysis strategy and conclusions are presented in the following passage, taken from GVT.

> If players can predict their hits and misses, their bets should correlate with their performance. . . . These data reveal that the players were generally unsuccessful in predicting hits and misses. . . . Only 5 of the 26 individual correlations were statistically significant, of which four were quite low (0.20 to 0.22) and the 5th was negative (-0.51). The four small but significant positive positive correlations may reflect either a limited ability to predict the outcome of an upcoming shot, or a tendency to try harder following a high bet.

This is quite a passage! Unlike the shooting data for which one significant result is discounted as being due to chance, the researchers do not acknowledge that obtaining four (or five if you look at both tails) significant results is noteworthy. Instead they label the correlations “quite low” and then “small.” Note, however, that the table below gives a correlation of 0.20.

Size of	Current Shot		
Bet	S	F	Total
High	30	20	50
Low	20	30	50

As argued above, for this application, a difference of 20 percentage points is not small! Finally, having obtained results counter to what they hoped to find, the researchers suggest that the data are due to players trying harder. In other words, if the null hypothesis is not rejected, there is no hot hand; if the null hypothesis is rejected, there is no hot hand. Why bother collecting data?

3 A Simulation Study of Power

In this section I study the performances of seven test statistics: the runs test and test of fit used in GVT and TG1; two tests motivated by data summaries presented in GVT and TG1; and three tests motivated by the work of LSK. The following is an overview of the results.

1. None of the tests possesses much power unless the departure from Bernoulli trials is fairly substantial.
2. Three of the tests—the runs test and the two motivated by GVT and TG1—are good at detecting autocorrelation, but poor at detecting nonstationarity.
3. The three tests motivated by the work of LSK are good at detecting nonstationarity, but poor at detecting autocorrelation.
4. The test of fit is inferior to the other tests at detecting any departure from Bernoulli trials.

I will begin with a description of the tests and the critical regions used. Recall the following property of Bernoulli trials. Let $X_1, X_2, \dots X_n$, be Bernoulli trials and let $T = \sum X_i$ be the total number of successes. Given that $T = t$ all sequences of t 1's and $(n-t)$ 0's are equally likely. Note in particular, that, conditional on T, the value of p is irrelevant.

The runs test. Let R be the number of runs in a sequence of 100 trials. Given $T = t$ and the null hypothesis, the mean, μ_R, and standard deviation, σ_R, of R have simple known formulas. Let $Z = (R - \mu_R)/\sigma_R$ denote the standardized value of R. The decision rule is to reject H_0 if, and only if, $Z \geq 1.645$.

The test of fit. This test was described in Section 2. I will use the one-sided version with nominal $\alpha = 0.05$. More precisely, if χ^2 denotes the test statistic, reject H_0 if, and only if,

- $\chi^2 \geq (1.645)^2$, and
- $E > O$ for the number of four-tuples with exactly two successes.

The AC2 test. Summarize the data with the following table.

Previous	Outcome		
2 Shots	S	F	Total
SS	a	b	n_1
FF	c	d	n_2
Total	m_1	m_2	n

Compute $Z = \sqrt{n}(ad - bc)/\sqrt{n_1 n_2 m_1 m_2}$, and reject H_0 if, and only if, $Z \geq 1.645$.

The AC3 test. Summarize the data with the following table.

Previous	Outcome		
3 Shots	S	F	Total
SSS	a	b	n_1
FFF	c	d	n_2
Total	m_1	m_2	n

Compute $Z = \sqrt{n}(ad - bc)/\sqrt{n_1 n_2 m_1 m_2}$, and reject H_0 if, and only if, $Z \geq 1.645$.

Before introducing the final three test statistics, consider the following sequence of 19 dichotomous trials.

0011011111110101110.

Define S_0 to be the length of the longest run of successes. For the above sequence, $S_0 = 7$:

00110 **1111111** 0101110.

Define $S_1 + 1$ to be the length of the longest run that contains exactly one failure. For the above sequence, $S_1 = 9$:

00 **1101111111** 0101110.

Finally, define $S_2 + 2$ to be the length of the longest run that contains exactly two failures. For the above sequence, $S_2 = 11$:

00110 **1111111010111** 0.

The S_j tests, $j = 0, 1, 2$. Each test rejects H_0 if, and only if, S_j is sufficiently large. For each test, the critical region depends on the value of T, and is discussed in the Appendix.

I begin with a study of the sizes of these seven tests. I performed simulation studies for each of five values of $p_B = 0.3(0.1)0.7$ with each study consisting of 10,000 runs. For each run, the computer generated 100 Bernoulli trials with success probability equal to p_B and each of the seven tests were applied to the generated sequence. The results of the simulation studies are presented in Table 1.

The runs test and test of fit reject the true null hypothesis at a rate very close to the target value of 0.05. The remaining tests do not reject often enough. Thus, those who view my research as critical of the work presented in GVT and TG1 can note that when I examine power, these researchers' two tests have the advantage of having a larger probability of type 1 error.

The low probabilities of type 1 error for AC3 are noteworthy. For $p_B = 0.7$ it was not uncommon for a generated sequence to contain no triple consisting of 'FFF.' With no such triple the test statistic cannot be computed and, thus, the null hypothesis cannot be rejected. A similar problem occurs for $p_B = 0.3$ and the triple 'SSS.' This difficulty arises again below when I examine the power of AC3.

Note, finally, that the estimated probability of type 1 error increases with increasing p_B for tests S_j. The reason for this pattern is explained in the Appendix.

I performed two studies of power. The first of these considered nonstationary alternatives and its results appear in Table 2; the second considered autocorrelation alternatives and its results appear in Table 3. I will discuss these tables in detail below.

I will begin with a discussion of the simulation study reported in Table 2. The 38 alternatives studied were constructed as follows. The data consisted of a sequence of Bernoulli trials with probability of success equal to p_B. At a random trial number U the probability of success increased to a value $p_H > p_B$. The probability of success remained at the higher level for a duration of D trials and then returned to its original value for the remainder of the sequence, which consisted of a total of 100 trials. The simulation study crossed five values of p_B, 0.3 (0.1) 0.7, with four values of p_H, 0.7 (0.1) 1.0, and two values of D, 10 and 20. The two combinations with $p_B = p_H = 0.7$ are omitted from the study, leaving 38 alternatives. Finally, U was selected from the uniform distribution on the integers 1, 2, . . . , $(101 - D)$.

For each alternative a simulation study with 1,000 runs (sequences of 100 trials) was performed. Each of the seven tests was applied to each sequence and the proportion of times, out of 1,000, that the test rejected the null hypothesis is reported in Table 2. For example, consider the alternative with $p_B = 0.4, p_H = 0.8$, and $D = 20$. In the context of basketball I consider this to to be a very substantial departure from Bernoulli trials; the probability of success increases a huge amount and persists for one-fifth of the sequence. Table 2 reveals that the test of fit is the worst of the seven tests, rejecting H_0 only 139 times. All tests of autocorrelation perform much better than the test of fit, with the runs test the least effective and AC2 and AC3 essentially equally effective. Among the tests of stationarity, S_2 is the best and S_0 is substantially poorer than the other two.

A careful study of the 38 alternatives reveals the following features.

1. The estimated powers exhibit the monotonicity relationships that one would expect. With some exceptions for its values below 0.120, when the other two parameters are held fixed, the estimated power is increasing in D and p_H and decreasing in p_B.

2. The test of fit is a disaster! For 30 of the 32 alternatives with $p_B \leq 0.6$, the test of fit has the lowest estimated power of any test. The two exceptions are for alternatives for which no test has any noticeable power, $p_B = 0.6, p_H \leq 0.8$, and $D = 10$. For the six alternatives with $p_B = 0.7$ the test of fit outperforms a smattering of tests, but its estimated power never exceeds 0.113. The test of fit is poor even at detecting huge departures from Bernoulli trials. For example, for the alternative with $p_B = 0.5, p_H = 1.0$, and $D = 20$, each test of stationarity rejects more than 990 times, each test of autocorrelation rejects approximately 500 times, but the test of fit rejects only 178 times!

3. For 22 of the 38 alternatives studied, both the runs test and the test of fit rejected the null hypothesis fewer than 200 times. Thus, the tests used in GVT and TG1 have little chance of detecting these alternatives.

4. For 13 of the 38 alternatives studied, each of the seven tests rejected the null hypothesis fewer than 200 times. For each of the remaining 25 alternatives, S_1 and S_2 rejected more often than the tests of autocorrelation or fit; S_0 lost slightly to AC2 for three alternatives ($p_B = 0.3, p_H \leq 0.9$ and $D = 20$) and to AC3 for one of these three, but McNemar's test applied to each of these four combinations reveals that in no case is AC2 or AC3 statistically significantly more powerful than S_0. In view of this item, I will turn my attention to finding the best among the tests of stationarity.

5. Consider tests S_j and $D = 10$. If $p_H - p_B \leq 0.2$, all estimated powers are below 0.100. If this difference equals 0.3 all estimated powers are below 0.200. For the remaining alternatives, a quick rule is to use S_0 if you believe that p_H is one (or, perhaps, very close to one) and use S_1 otherwise.

6. Consider tests S_j and $D = 20$. If $p_H - p_B = 0.1$, all estimated powers are below 0.100. If this difference equals 0.2, the estimated power increases in p_B, but never exceeds 0.250. For the remaining alternatives, a quick rule is to use S_0 if you believe that p_H is one (or, perhaps, very close to one) and use S_2 otherwise.

I also performed a simulation study of power for 36 autocorrelation alternatives. The results appear in Table 3. The alternatives are obtained by crossing three values of L, 1, 2, and 3, with four values of δ, 0.05 (0.05) 0.20, and three values of $p_B = 0.50, 0.60, 0.70$. (See below for a discussion of $p_B = 0.30$ or 0.40.)

The data are 100 dichotomous trials. The first trial is a Bernoulli trial with probability of success p_B. For subsequent trials: after L or more consecutive successes, the probability of a success is $p_B + \delta$; after L or more consecutive failures, the probability of success is $p_B - \delta$; otherwise the probability of success is p_B. For example, for $L = 2$, $\delta = 0.10$ and $p_B = 0.50$, the trials are independent with probability of success equal to: 0.60 after two or more consecutive successes; 0.40 after two or more consecutive failures; and 0.50 otherwise. Notice that $L = 1$, $\delta = 0.10$, and $p_B = 0.50$ gives a very close approximation to the mean response of the fan survey reported in TG1. Note the following features revealed by Table 3.

1. The test of fit and the S_j tests are inferior to AC2, AC3, and the runs test.

2. For $L = 1$: The runs test is much better than the other tests.

3. For $L = 2$:

   - For $p_B \leq 0.6$: AC2 is the best test for $\delta \geq 0.10$.
   - For $p_B = 0.6$ and $\delta = 0.05$, the runs test and AC2 are equally effective.
   - For $p_B = 0.5$ and $\delta = 0.05$, the runs test is best. This is mildly surprising.
   - For $p_B = 0.7$, AC2 and the runs test are best.

4. For $L = 3$:

   - For $\delta = 0.05$, none of the tests is effective.
   - For $\delta \geq 0.10$ and $p_B \leq 0.6$, AC3 is the best test.
   - For $\delta \geq 0.10$ and $p_B = 0.7$, the runs test is best.

5. The largest estimated power is modest for alternatives that I would label realistic, namely $\delta \leq 0.05 \times L$.

Finally, note that by interchanging the roles of success and failure, the entries in Table 3 for $p_B = 0.6$ (0.7) are also valid for $p_B = 0.4$ (0.3) for the test of fit, the runs test and tests AC2 and AC3. This argument does not apply to the tests S_j.

4 A Case Study

Katie Voigt was a four-year starter on the UW-Madison women's varsity basketball team and graduated in 1998 with a degree in Secondary Education–Mathematics. Her usual position was shooting guard, also called off-guard and position two. I selected Katie for this study because she is an excellent shooter and I knew her from a class I had taught. I had no sense whether Katie was a particularly streaky shooter and did not tell her the purpose of my study. Katie agreed to record the results of her practice for twenty days during the (summer) off-season. On each day, after warming-up, Katie would attempt 100 shots from behind the three point arc and record her sequence of outcomes. In this section I will present my analysis of Katie's data. (If you would like to try your hand at analyzing Katie's data, contact me at wardrop@stat.wisc.edu, and I will send the data in reply.)

Katie shot better than I had expected, making 65.05 percent of her 2,000 attempts. I performed a χ^2 test of homogeneity on the 20×2 table obtained from the days (rows) and shot outcome (columns). The observed value of the test statistic is 31.359; the right-tail area of the χ^2 curve with 19 degrees of freedom is 0.0368. Thus, there is statistically significant evidence against the hypothesis of Bernoulli trials with constant (day-to-day) probability of success. The tests of Section 3 are designed to analyze the data *within* days; this first result suggests that collapsing data over days could be misleading, but an appropriate combination of data across days, given later, could be useful. Finally, I note that the standard deviation of Katie's 20 day-totals is 6.13, slightly more than 22 percent larger than the maximum population standard deviation for constant p binomials ($10\sqrt{pq}$).

In the remainder of this section I will focus on within-day analyses of Katie's data, using four of the tests described in Section 3 and three new tests described below. I will not use the test of fit; I believe that it has been established that this test is worthless. I will use neither AC2 nor AC3. My power study of these tests was motivated by the descriptive statistics in GVT and TG1. In terms of power, these tests perform best for alternatives that I do not consider very realistic. In addition, Katie shot very well and the results of Section 3 show that these tests perform worse when the baseline probability of success differs substantially from 0.5.

To each day's data I applied the three S_j tests and the runs test. I also applied three additional tests, denoted by $F_j, j = 0, 1, 2$. Each F_j is defined as S_j with the rolls of success and failure interchanged. Thus, for example, F_0 equals the longest streak of failures in the data set. Clearly, the sampling distributions for S_j given $T = t$ are also the sampling distributions for F_j given $T = 100 - t$. Thus, the results in Table 1 for S_j have immediate application to F_j; the estimated probability of type 1 error for S_j and p_B is also the estimated probability of type 1 error for F_j and $(1 - p_B)$.

Table 2 presents estimated powers for S_j for some alternatives in which a steady state is interrupted by a brief duration of *hot shooting.* These are also estimated powers for F_j for some alternatives in which a steady state is interrupted by a brief duration of *cold shooting.* For example, consider an alternative in which the probability of a success is 0.6, but at a random time it declines to 0.3 for a duration of 20 trials. The estimated power for F_j for such an alternative equals the entry in Table 2 for $p_B = 0.4$, $p_H = 0.7$, and $D = 20$.

I computed two one-sided P-values for each of the seven tests; one for the alternative of "streakiness" and one for the alternative of "antistreakiness." For S_j, and F_j the right tail is for streakiness and the left tail for antistreakiness. For the runs test, the reverse is true.

Table 4 presents all P-values that are 0.1000 or smaller, for either alternative. For S_j and F_j, the P-values are obtained conditional on $T = t$ from a simulation study with 10,000 runs. For the runs test, the P-values are exact, again conditional of $T = t$.

For days 3–8 there is substantial and strong evidence that Katie was streaky, but the nature of the streakiness was inconsistent. The shots on day 3 (day 8) exhibited strong (borderline) evidence of positive autocorrelation. On days 3 and 5 tests F_j provide evidence of cold streaks, while on days 4, 7, and 8 tests S_j provide evidence of hot streaks.

For days 10–14 there is a smattering of evidence of streakiness, mostly in the form of cold streaks, but the evidence is substantially weaker and less prevalent than on the earlier days.

Finally, in three of the last four days the data provide evidence of antistreakiness, with a particularly small P-value for the runs test on day 17.

I now turn to a description of the data behind Table 4. Beginning with the runs test, Katie shot 32 (day 3), 15 (day 8), and 14 (day 13) percentage points better after a hit than after a miss. By contrast, she shot 27 percentage points better after a miss than after a hit on day 17.

For tests S_j and F_j, Table 5 presents the value of the test statistic and its estimated null mean for each combination of day and test that has an entry in Table 4. For example, on day 3 Katie obtained $t = 66$ successes and 34 failures. Her longest streak of failures equaled nine, was much larger than its estimated null mean of 3.9, and yielded an estimated P-value of 0.0019.

Conditional on $T = t$ the null mean and variance of the runs test has a simple formula. Again conditionally, my simulation studies yield an estimate of the null mean and variance for S_j and F_j. Any of the seven tests can be used to test the null hypothesis that the data *every day* are Bernoulli trials with probability of success on day k equal to p_k. (The χ^2 test, discussed earlier, was a *between days* test and rejected the null hypothesis that every day had the same probability of success. The tests presented below are *within days* tests that search for evidence of streakiness without assuming homogeneity between days.) In particular, let Y_k be the value of the test statistic (the number of runs, or the value of S_j or F_j) on day k. Define W to be the sum of the Y_k's. A large value of W for S_j or F_j or a small value of W for the runs test would provide evidence against the null hypothesis of Bernoulli trials every day and in support of the alternative of streakiness on one or more days. The value of W is given in Table 6 for each of the seven test statistics. Each value of W was used two ways to obtain an approximate P-value. First, W was standardized and the standard normal curve was used in the ordinary way. Second, a simulation experiment yielded 10,000 values of W and the simulated distribution of W was used to obtain an approximate P-value.

The results are presented in Table 6. The P-values for the tests F_j are smaller than 0.01 and, for example, the value of W for F_1 is approximately 15 percent larger than its expected value. The test statistic W yields P-values below 0.05 for S_1 and S_2. The value of W for S_1 is approximately 10 percent larger than its expected value. Finally, the W's for the runs test and S_0 do not achieve statistical significance.

In summary, there are several reasons to believe that Katie's shooting is more complex than Bernoulli trials. The frequency and size of the hot-hand may or may not match the expectations of many persons, but it is too simplistic, I believe, to assume that Bernoulli trials are always appropriate.

The reader is encouraged to collect data and use the ideas of this paper to analyze them.

References

[1] Albright, S. C. (1993). A Statistical Analysis of Hitting Streaks in Baseball. *Journal of the American Statistical Association*, 88(424), 1175–1183.

[2] Albert, J. (1993). Comment. *Journal of the American Statistical Association*, 88(424), 1184–1188.

[3] Bennett, J. (1998). *Statistics in Sport*. London, Arnold Publishers.

[4] Forthofer, R. (1991). Streak Shooter–the Sequel. *Chance*, 4(2), 46–48.

[5] Gilovich, T., Vallone, R., and Tversky, A. (1985). The Hot Hand in Basketball: On the Misperception of Random Sequences. *Cognitive Psychology*, 17, 295–314.

[6] Hooke, R. (1989). Basketball, Baseball, and the Null Hypothesis. *Chance*, 2(4), 35–37.

[7] Larkey, P., Smith, R., and Kadane, J. (1989). It's Okay to Believe in the "Hot Hand." *Chance*, 2(4), 22–30.

[8] Stern, H. (1995). Who's Hot and Who's Not: Runs of success and failure in sports. In *1995 Proceedings of the Section on Statistics in Sports*. American Statistical Association, 26–35.

[9] Stern, H., and Morris, C. (1993). Comment. *Journal of the American Statistical Association*, 88(424), 1189–1194.

[10] Tversky, A., and Gilovich, T. (1989a). The Cold Facts about the 'Hot Hand' in Basketball. *Chance*, 2(1), 16–21.

[11] Tversky, A., and Gilovich, T. (1989b). The 'Hot Hand': Statistical Reality or Cognitive Illusion? *Chance*, 2(4), 31–34.

[12] Wardrop, R. (1995). Simpson's Paradox and the Hot Hand in Basketball. *The American Statistician*, 49(1), 24–28.

5 Appendix

In this section I will describe my estimates of the null sampling distribution of $S_j, j = 0, 1, 2$.

Recall that the (null) data consists of 100 Bernoulli trials. In practice, the probability of success p will be unknown, and this problem can be handled by conditioning on the number of successes, T, in the sequence. Conditional on $T = t$, the $C(100, t) = 100!/(t!(100 - t)!)$ possible arrangements of t successes and $(100 - t)$ failures are equally likely. Thus, while the exact distribution of S_j given $T = t$ can be difficult to obtain, it can be simulated quite easily.

For $t = 3, 4, \ldots, 97$ I performed a simulation study with 10,000 runs. (If t is smaller than, say, five or larger than 95 then, intuitively, there is not much hope of finding streakiness.) For each run, the computer selected t numbers at random without replacement from 1, 2, ..., 100; these numbers gave the positions of the successes in the sequence. Once these positions were known, the computer calculated the value of S_0. This entire process was repeated for S_1 and then again for S_2.

The simulation was a bit more sophisticated than suggested above. For each run the computer first picked $t = 3$ numbers at random, for example, 17, 34, and 81. Then it added one number to this list to get the positions for $t = 4$, for example, 17, 34, 56, and 81. And so on. This method forces a stochastic ordering in the estimated sampling distributions, namely, the simulated distribution of S_j, conditional on $T = t + 1$ is stochastically larger than its distribution conditional on $T = t$.

A summary of the simulation studies is available via email at the address reported earlier. I noted in Section 3 that the estimated probability of type 1 error for each of the tests S_j, increased as p_B increased. As p_B increases, the values of t obtained will tend to increase. As the value of t increases, of the estimated tail probabilities that are 0.0500 or smaller (and, hence, will lead to rejecting the null hypothesis) the largest tends to be closer to 0.0500. Thus, the null hypothesis will tend to be rejected more often for larger t's or larger p_B's. (This phenomenon is caused by the distribution of S_j becoming less granular as t increases.)

Table 1: Estimated probability of type 1 error for seven tests based on simulation studies with 10,000 runs. The data are 100 Bernoulli trials with probability of success equal to p_B.

			p_B		
Test	0.3	0.4	0.5	0.6	0.7
Runs	0.0532	0.0503	0.0484	0.0520	0.0508
AC2	0.0387	0.0360	0.0354	0.0391	0.0399
AC3	0.0245	0.0284	0.0300	0.0287	0.0283
S_0	0.0250	0.0329	0.0373	0.0395	0.0415
S_1	0.0270	0.0321	0.0356	0.0407	0.0419
S_2	0.0279	0.0315	0.0337	0.0359	0.0397
Fit	0.0570	0.0441	0.0477	0.0474	0.0560

Table 2: The estimated power for seven tests based on simulation studies with 1,000 runs. The data are 100 Bernoulli trials with probability of success equal to p_B. At a random time the probability of success jumps to p_H, remains at that value for duration D and then returns to p_B. The largest estimated power for each alternative is in bold face (provided it exceeds 0.100) as is any estimated power that is not statistically significantly different from the largest.

	$p_B = 0.3$							
D:	10				20			
p_H:	0.7	0.8	0.9	1.0	0.7	0.8	0.9	1.0
Runs	0.162	0.235	0.346	0.527	0.288	0.494	0.716	0.921
AC2	0.187	0.308	0.490	0.805	0.367	0.622	0.878	0.992
AC3	0.178	0.321	0.521	0.894	0.336	0.620	0.857	0.986
S_0	0.205	0.399	0.640	**1.000**	0.339	0.607	0.869	**1.000**
S_1	**0.265**	**0.503**	**0.748**	0.992	0.457	0.746	0.950	**1.000**
S_2	**0.279**	**0.481**	0.704	0.955	**0.521**	**0.805**	**0.979**	**1.000**
Fit	0.116	0.144	0.146	0.179	0.182	0.240	0.362	0.470

	$p_B = 0.4$							
D:	10				20			
p_H:	0.7	0.8	0.9	1.0	0.7	0.8	0.9	1.0
Runs	0.101	0.136	0.197	0.305	0.148	0.251	0.455	0.724
AC2	0.103	0.141	0.261	0.434	0.160	0.328	0.569	0.871
AC3	0.093	0.154	0.333	0.525	0.152	0.320	0.594	0.866
S_0	**0.138**	0.271	**0.519**	**0.967**	0.191	0.415	0.748	**1.000**
S_1	**0.155**	**0.299**	**0.543**	0.837	**0.255**	0.565	0.868	**1.000**
S_2	**0.141**	0.268	0.451	0.681	**0.262**	**0.622**	**0.912**	**1.000**
Fit	0.074	0.071	0.120	0.111	0.097	0.139	0.203	0.251

	$p_B = 0.5$							
D:	10				20			
p_H:	0.7	0.8	0.9	1.0	0.7	0.8	0.9	1.0
Runs	0.070	0.075	0.124	0.185	0.071	0.157	0.266	0.494
AC2	0.056	0.069	0.128	0.201	0.065	0.158	0.269	0.515
AC3	0.038	0.071	0.124	0.211	0.069	0.136	0.288	0.498
S_0	0.070	**0.133**	**0.330**	**0.739**	0.091	0.259	0.603	**1.000**
S_1	0.061	**0.143**	0.274	0.510	**0.124**	0.338	0.734	**0.999**
S_2	0.059	**0.126**	0.228	0.390	**0.133**	**0.393**	**0.801**	0.993
Fit	0.049	0.058	0.089	0.096	0.060	0.108	0.140	0.178

Table 2: (Continued) The estimated power for seven tests based on simulation studies with 1,000 runs. The data are 100 Bernoulli trials with probability of success equal to p_B. At a random time the probability of success jumps to p_H, remains at that value for duration D and then returns to p_B. The largest estimated power for each alternative is in bold face (provided it exceeds 0.100) as is any estimated power that is not statistically significantly different from the largest.

	$p_B = 0.6$							
D:	10				20			
p_H:	0.7	0.8	0.9	1.0	0.7	0.8	0.9	1.0
Runs	0.043	0.058	0.094	0.136	0.054	0.084	0.141	0.320
AC2	0.044	0.047	0.085	0.101	0.033	0.068	0.132	0.274
AC3	0.034	0.042	0.071	0.099	0.027	0.053	0.123	0.215
S_0	0.040	0.057	**0.153**	**0.336**	0.060	0.132	0.381	**0.996**
S_1	0.046	0.067	**0.148**	0.269	0.070	**0.173**	**0.469**	0.949
S_2	0.048	0.062	**0.129**	0.202	0.062	**0.168**	**0.467**	0.862
Fit	0.045	0.052	0.074	0.084	0.070	0.070	0.097	0.147

	$p_B = 0.7$							
D:	10				20			
p_H:	0.7	0.8	0.9	1.0	0.7	0.8	0.9	1.0
Runs	—	0.066	0.077	0.109	—	0.057	0.076	0.169
AC2	—	0.049	0.064	0.056	—	0.049	0.072	0.123
AC3	—	0.035	0.032	0.045	—	0.028	0.046	0.082
S_0	—	0.057	0.078	**0.176**	—	0.056	**0.248**	**0.848**
S_1	—	0.055	0.069	0.147	—	0.070	**0.245**	0.664
S_2	—	0.056	0.067	0.126	—	0.064	**0.234**	0.530
Fit	—	0.056	0.065	0.083	—	0.062	0.072	0.113

Table 3: The estimated power for seven tests based on simulation studies with 1,000 runs. The data are 100 dichotomous trials. The first trial is a Bernoulli trial with probability of success p_B. For subsequent trials: after L or more consecutive successes, the probability of a success is $p_B + \delta$; after L or more consecutive failures, the probability of success is $p_B - \delta$; otherwise the probability of success is p_B. The largest estimated power for each alternative is in bold face (provided it exceeds 0.100) as is any estimated power that is not statistically significantly different from the largest.

$p_B = 0.5$

	$L = 1$				$L = 2$				$L = 3$			
δ:	0.05	0.10	0.15	0.20	0.05	0.10	0.15	0.20	0.05	0.10	0.15	0.20
Runs	**0.237**	**0.597**	**0.903**	**0.986**	**0.173**	0.293	0.476	0.743	0.081	0.176	0.253	0.363
AC2	0.134	0.380	0.706	0.898	0.137	**0.374**	**0.623**	**0.877**	0.089	0.189	0.295	0.437
AC3	0.090	0.244	0.473	0.748	0.092	0.239	0.434	0.714	0.087	**0.234**	**0.381**	**0.579**
S_0	0.080	0.161	0.287	0.460	0.080	0.129	0.232	0.419	0.068	0.141	0.206	0.335
S_1	0.075	0.146	0.245	0.420	0.080	0.145	0.244	0.422	0.069	0.145	0.213	0.308
S_2	0.079	0.136	0.236	0.353	0.072	0.134	0.227	0.394	0.067	0.127	0.183	0.286
Fit	0.110	0.196	0.283	0.444	0.105	0.178	0.255	0.389	0.061	0.063	0.092	0.136

$p_B = 0.6$

	$L = 1$				$L = 2$				$L = 3$			
δ:	0.05	0.10	0.15	0.20	0.05	0.10	0.15	0.20	0.05	0.10	0.15	0.20
Runs	**0.266**	**0.621**	**0.894**	**0.986**	**0.135**	0.294	0.511	0.710	0.087	0.150	0.257	0.420
AC2	0.156	0.384	0.687	0.902	**0.137**	**0.348**	**0.588**	**0.806**	0.081	0.162	0.270	0.421
AC3	0.081	0.248	0.459	0.750	0.085	0.223	0.401	0.627	0.075	**0.188**	**0.343**	**0.514**
S_0	0.071	0.139	0.276	0.410	0.067	0.136	0.242	0.398	0.071	0.125	0.220	0.375
S_1	0.081	0.147	0.249	0.374	0.074	0.137	0.245	0.386	0.069	0.110	0.209	0.302
S_2	0.074	0.141	0.238	0.329	0.086	0.134	0.210	0.350	0.056	0.104	0.195	0.288
Fit	0.109	0.182	0.302	0.498	0.082	0.155	0.272	0.407	0.059	0.080	0.106	0.150

$p_B = 0.7$

	$L = 1$				$L = 2$				$L = 3$			
δ:	0.05	0.10	0.15	0.20	0.05	0.10	0.15	0.20	0.05	0.10	0.15	0.20
Runs	**0.262**	**0.605**	**0.851**	**0.935**	**0.127**	**0.291**	**0.494**	**0.661**	**0.095**	**0.185**	**0.313**	**0.467**
AC2	0.135	0.341	0.589	0.730	**0.120**	**0.312**	**0.501**	**0.637**	**0.080**	0.156	0.280	0.372
AC3	0.096	0.202	0.362	0.509	0.069	0.160	0.310	0.418	0.075	0.139	0.232	0.345
S_0	0.076	0.147	0.228	0.334	0.082	0.142	0.201	0.311	**0.102**	0.140	0.271	0.389
S_1	0.091	0.151	0.218	0.304	**0.101**	0.118	0.209	0.322	0.081	0.138	0.239	0.336
S_2	0.073	0.133	0.195	0.299	0.076	0.127	0.222	0.325	0.070	0.126	0.222	0.309
Fit	0.120	0.201	0.335	0.440	0.079	0.177	0.311	0.382	0.076	0.094	0.154	0.201

Table 4: P-values for seven tests for Katie's daily data. Only one-sided P-values below 0.1000 are printed. **Bold face** P-values are for the alternative of antistreakiness; all others are for the alternative of streakiness.

				Test			
Day	Runs	S_0	S_1	S_2	F_0	F_1	F_2
3	0.0007				0.0019	0.0109	0.0399
4			0.0362	0.0924			
5					0.0324	0.0541	0.0221
7		0.0074	0.0543	0.0672			
8	0.0764	0.0242	0.0238	0.0205			
10				0.0924			
11						0.0363	
12					0.0495	0.0430	
13	0.0947						
14						0.0718	0.0717
17	**0.0041**				**0.0437**		
19			**0.0946**				
20		**0.0818**					

Table 5: Values of the test statistic (and its estimated null mean) for the entries in Table 4.

				Test			
Day	t	S_0	S_1	S_2	F_0	F_1	F_2
3	66				9 (3.9)	9 (5.0)	9 (5.9)
4	68		20 (13.1)	21 (16.0)			
5	64				7 (4.1)	8 (5.3)	10 (6.2)
7	68	19 (9.8)	19 (13.1)	22 (16.0)			
8	59	13 (7.5)	16 (10.0)	19 (12.1)			
10	68			21 (16.0)			
11	61					9 (5.8)	
12	56				8 (5.1)	10 (6.6)	
14	68					7 (4.7)	8 (5.5)
17	76				2 (3.7)		
19	67		9 (12.7)				
20	71	7 (10.8)					

Table 6: Results of combining data across days. The value of the test statistic (number of runs or value of S_j or F_j) on day k is denoted by Y_k. The null mean and variance of Y_k are denoted by μ_k and σ_k^2. P-values are presented for the one-sided alternative of streakiness. Each P-value is computed twice; by referring Z to the standard normal curve, and by a 10,000 run simulation experiment.

	Runs	S_0	S_1	S_2	F_0	F_1	F_2
$W = \Sigma Y_k$	889	193	270	322	94	119	138
$\mu_W = \Sigma \mu_k$	915.1	183.8	245.8	298.4	81.0	103.7	121.8
$\sigma_W = \sqrt{\Sigma \sigma_k^2}$	19.96	11.11	12.66	13.99	4.95	5.57	6.13
$Z = (W - \mu_W)/\sigma_W$	-1.31	0.82	1.92	1.68	2.62	2.74	2.64
P-value (snc)	0.0951	0.2061	0.0274	0.0465	0.0044	0.0031	0.0041
P-value (sim)	0.0965	0.2199	0.0387	0.0490	0.0097	0.0065	0.0085