Do Baseball Players Have Hot Streaks?

David A. Levine

March 21, 2007

You can see a lot simply by observing.

Yogi Berra

…I know it when I see it…

Justice Potter Stewart

…we all know from personal experience that the world is flat. Indeed, if it were round gravity would cause the oceans on the top to flow around and drip off the bottom, which of course is ridiculous.

J. Crocker Fisher

"It'spring!" he shouts."It's spring!"

Babar, the King of Elephants

Acknowledgments

Much of the information used here was obtained free of charge from and is copyrighted by Retrosheet. Interested parties may contact Retrosheet at "www.retrosheet.org".

I am grateful for the assistance and counsel of Rick Kaye and Jonathan Reiss. Rick is my "supplier" – he obtained most of the Retrosheet data and organized it in such a way that saved me countless hours of work. Both Rick and Jonathan helped steer me to the right kinds of statistical tests, provided me with supplemental information (some of which is referred to in the text) and offered many helpful comments on the original text of this paper. I would also like to thank Danny Mintz and Zalmen Rosenfeld for many provocative conversations and e-mail exchanges. I owe all of these friends a debt of gratitude and if there are any screw-ups in here, it's…you guessed it…my fault (unless it's really theirs).

Introduction

I've been hearing and reading about streakiness in the context of sports – baseball and basketball especially – since I was a kid. Depending on the sport, the broadcaster will say that Smith "can't miss," or "is on fire" or "is seeing the ball real well" or "is in the zone." Jones, by way of contrast, is "slumping" or is "having trouble focusing" or "can't find a pitch to hit."

This way of thinking of things became more salient to me when I entered the investment research business in 1972. As it turns out, investors think that stocks and bonds are streaky too. I've been (sporadically) writing about market and economic streakiness for about three decades but I recently realized that it was high time that I write about something that's actually important – baseball! [1]

What Do Sports Fans Mean by "Streakiness?"

The American Heritage Dictionary defines "streaky" as "1. Marked with, characterized by, or occurring in streaks. 2. Variable or uneven in character or quality." It goes on to offer the following "informal" definition of "streak" as "A brief run, or stretch, as of luck." The phrase "as of luck" suggests it might sometimes be luck and sometimes not be luck. But, in general, when sports fans say a batter is "hot" or "in a slump" they are most emphatically *not* talking about a fluctuation in performance that is *merely* the result of chance. Rather, they are talking about the establishment of a positive (or negative) *psychological feedback loop* that contributes to a change in the player's performance. In this view, success breeds success and failure breeds failure.

To be sure, the worst major league hitters occasionally get four hits in a row and the best occasionally go 0/12. Thus, there is no doubt that streaks happen. Indeed, any run of two hits in a row or two outs in a row can be considered a streak. But that's not what sports fans mean when they say players are "streaky" or that "hot streaks" or "slumps" exist: What they mean (even if they do not spell it out this way) is that streaks are more frequent and longer than what one would expect from a random process. If this is true, then ballplayers are, in fact, streaky. But if streaks are *not* more common/longer than one would expect from a random process, then "hot streaks" and "slumps" are myths.

---

[1] Using the definition of streakiness laid out in the next section, I can report that (a) the economy is somewhat streaky, (b) the rate of inflation is very streaky, (c) short-term interest rates are extremely streaky (because the Fed manages them that way), (d) intermediate and long-term interest rates are ever-so-slightly streaky, (e) individual stocks have a tendency to be streaky, (f) broad stock market indices used to be slightly streaky over the short run but haven't been since the late 1970s, (g) broad stock market indices are the *opposite of streaky* (i.e., they are "mean reverting") over long-term and (even) medium-term time horizons, and (h) currencies are streaky in the short term and mean-reverting in the long run (adjusted for inflation). A correct understanding of (c), (e) and (h), can help you make winning investment bets if incorporated into the right kind of model and recognition of (g) can help you make money if it convinces you to increase your exposure to the stock market (because the large fluctuations in the stock market over the short run really do have some tendency to "even out" over the long run). If any of you would like to discuss these economic/financial subjects with me – I'm available. As you might infer from this (oh, 49-page) paper, I have some extra time on my hands.

Would You Know Streakiness if You Saw It?

Let's start with something whose statistical properties are very familiar – namely, coin flips. Obviously, when you toss a coin repeatedly you will end up averaging close to 50% heads and 50% tails. The more flips you do, the closer your percentages are likely to be to the theoretical 50/50 average. But despite that average, here's a pattern you will probably never see in your lifetime.[2]

H T H T H T H T H T H T H T H T H T H T H T H T H T H T etc.

The reason, of course, that you will not see anything like the above, is that over that many flips you will almost inevitably encounter stretches of two in a row, three in a row, etc. Indeed, streaks as long as ten heads in a row will happen about once per thousand flips and even 20 in a row will happen about one time per million flips.[3] With this in mind, can the reader guess which of the following patterns of 32 coin flips is streaky and which is not? (Note: Each row has exactly 16 heads and 16 tails.)

HHH T HH T HH TT H TT HH T HHH T H TTTT H TTTT H

T H T HHHHHH T HH TT H TT H T H T HH T H T H TTTTT

As you, dear reader, mull over the question of whether you can recognize streakiness when you see it, let us think about why the average sports fan believes streaks exist and are not a figment of his (or the collective) imagination. Hardly anyone studies the question, even informally.[4] Nevertheless, the vast majority of sports fans believe that streakiness exists *and that they can recognize* when a player has gotten "hot" or fallen into a "slump." How do they this?

That sportscasters constantly spout statistics in support of streakiness ("Gonzales is 'red hot' – he's got 8 hits in his last 11 at-bats.") certainly encourages the idea. But, mostly fans believe that streakiness exists *simply from watching*[5]. As a result of a poll I took on this subject – which I'll be reporting on a bit later – I had any number of conversations and e-mail exchanges with friends convinced that streakiness exists. There are two main reasons for this belief: (1) observation and (2) personal experience. I will address

---

[2] The chances of 14 heads alternating with 14 tails and starting with a head are about one in 268 million. If you spend your entire life flipping coins, it could happen. But at 10 flips per minute, 12 hours per day, you would expect to encounter this pattern only once every 102 years or so.

[3] Most of you are familiar with the math here: The chance of a single head is ½: The chance of ten in a row is ½ raised to the tenth power – i.e., once chance out of every 1,024 attempts. The chance of 20 in a row is ½ raised to the twentieth power or one chance in 1,048,576 tries.

[4] One excellent exception – "The Hot Hand in Basketball: On the Misperception of Random Sequences" by Thomas Gilovich, Robert Vallone and Amos Tversky – appears in "Cognitive Psychology, 1985, 17: 295-314." I have a copy, courtesy of Jonathan Reiss, and will e-mail it to the reader upon request.

[5] This conjures up one of the most famous lines in the history of Supreme Court decisions: Justice Potter Stewart's opinion in *Jacobellis v.Ohio* (1964) said the Constitution protected all obscenity except "hard-core pornography." He wrote: "I shall not today attempt further to define the kinds of material I understand to be embraced within that shorthand description; and perhaps I could never succeed in intelligibly doing so. But *I know it when I see it,* and the motion picture involved in this case is not that." [Emphasis added.]

personal experience later but, for now, let me just say that sports fans who know streakiness exists by virtue of observation are effectively saying that they can recognize a streak when they see it. Hence my little experiment above: Are you one of those who can *see* the difference?

I hope the reader will not consider it an ethical breach when I confess that neither of those sequences was streaky. Here they are again:

HHH T HH T HH TT H TT HH T HHH T H TTTT H TTTT H

T H T HHHHHH T HH TT H TT H T H T HH T H T H TTTTT

Indeed, one of them (from a statistical standpoint) is very close to a typical random sequence and one of them is actually somewhat "UNstreaky." A sequence is "unstreaky" when it is characterized by fewer/shorter streaks than might be expected from a random sequence. Can you guess which one is *less* streaky than typical random streakiness?

One measure of streakiness is the correlation between successive coin flips – the so-called "serial correlation." Given the modest sample size here, this correlation would have to be +0.35 to be considered genuinely streaky from a statistical standpoint.[6] (A correlation between +0.10 and +0.30 might be considered mild evidence of streakiness.) But the first sequence shown above has a serial correlation of -0.03 (i.e., very close to zero and, therefore, about "as random as you can get") while the second is the one on the UNstreaky side (albeit *not* statistically significant), with a correlation of -0.16.

Some might think that the second series is streakier than the first because it has the two longest streaks in either sequence – i.e., one run of six heads and another of five tails. But the second sequence also has 13 flips (out of the total of 32) that are "singles" (a head surrounded by tails or a tail surrounded by heads[7]), only four pairs, and no runs of three or four. In a "typical" sequence one would expect only 8-9 "singles." As it turns out, the plethora of singles more than outweighs the two long-ish streaks.

Sorry to have pulled a fast one on you but it is the best I could do (without gathering all of you in one room) to replicate the spirit of a story told to me by Rick Kaye about a professor of statistics who would divide his class into two parts. The professor gave each a task and left the classroom while they performed it. Group 1 was supposed to flip a coin 100 times and write the resulting sequence on the blackboard. Group 2 was supposed to consult with each other and *invent* a hypothetical sequence of 100 flips that they thought would *look* random. On re-entering the classroom, the professor would glance at the blackboard and instantly distinguish the artificially-constructed series from the actual coin-flip sequence. How did he do it? The artificial sequence would always have much too much alternation between heads and tails, too many short streaks (two in a row, three in a row) and not enough streaks of five or six or longer.

[6] At the .05 confidence level.

[7] Obviously the first and last flips cannot be "surrounded." There are 12 heads or tails surrounded by their opposites. The initial tail is adjacent to a head.

Evidently, most people think randomness entails *a tendency towards alternation* and so when they see something that's genuinely random, they are inclined to think it is streaky. If this is true with regard to coin flips, it seems reasonable to assume that this might also be true with regard to other phenomena – like baseball batting sequences – i.e., fans will be inclined to think they are seeing streaki*ness*, when all they are seeing are the streaks that are a typical part of random sequences.

The propensity to see streakiness that does not exist means sports fans' impressions cannot be trusted. Perhaps the streakiness they think they see is real; perhaps it is not. The point is that we need to use objective measurement techniques – statistical tests – to determine whether baseball players are genuinely streaky.

Streaks, Intuition, Faith

Most people understand "coin-flip randomness." Perhaps, they cannot *recognize* a typical random sequence when they see it but they do not need to be convinced that the process of flipping coins is a random one – i.e., entirely a question of luck.[8] On the other hand, coin flips don't matter very much to people. But, when it comes to subjects that people care about, they *crave* causality.

Sometimes, that craving is self-contradictory! A batter who is 0/14 comes to the plate. He strikes out and the fan says: "Of course he struck out. It was likely to happen. He's in a terrible, terrible slump." But, if the same player in the exact same situation gets a hit, the very same fan might exclaim "of course… he was 'due' to finally get at hit." One cannot (or, at least, should not) believe that both of these statements make sense, but many baseball fans (and announcers) seem capable of making either statement, depending on the outcome of the particular at-bat.

Sometimes, the craving that every result needs to have an identifiable, knowable cause is at the highest possible intellectual level: (The discussion is way above my pay-grade, but) Einstein, in rejecting some of the implications of quantum theory related to the random properties of matter and energy, is famously quoted to have said "God does not play dice with the universe."[9] Sometimes, the craving that things have to happen for a reason is about the most important possible subject – life and death. It is discomfiting to recognize this but many medical outcomes have a random element to them. While working on this rumination on baseball streaks an article appeared in the New York Times regarding the role of luck in medicine.[10] An excerpt:

---

[8] On the other hand, you'll find plenty of people in Casinos who, while they are convinced that coin tossing is random, nevertheless believe the dice on craps tables and the balls on roulette wheels are not.

[9] This wording (or a variation) is commonly used to *paraphrase* what Einstein actually wrote (in a letter to Max Born, a Nobel Prize winner and – I'm not kidding – the grandfather of Olivia Newton-John): "The [quantum] theory accomplishes a lot, but it does not bring us closer to the secrets of the Old One. In any case, I am convinced that He does not play dice." (Yes, Einstein capitalized "He.")

[10] New York Times, September 19, 2006, page F5. "In Science-Based Medicine, Where Does Luck Fit In?" by Barron H. Lerner, M.D.

> "Luck seems to have become particularly anathema in an era of evidence-based medicine, in which physicians and patients are encouraged to learn the latest relevant data to guide decisions. Dr. Peter A. Ubel, a University of Michigan internist and author of 'You're Stronger Than You Think,' believes that his patients prefer biological explanations of why they are sick, rather than hearing that they have bad genes or bad luck. But given the biological variability within given diseases, like cancer, and the fact that variable genetic makeup leads different individuals to respond differently to diseases and therapies, even better scientific knowledge will not eliminate the role played by luck. Chance, the British physician R. J. Epstein wrote in the Quarterly Journal of Medicine, ensures different outcomes within given sick populations. A few examples? Roughly 1 percent of North American whites are highly resistant to H.I.V. infection because they lack a certain cell surface protein. Lucky. Roughly 5 percent of people infected with the hepatitis B virus develop chronic active hepatitis, an often serious liver disease. Unlucky. This phenomenon can be seen in individual cases of disease as well. When the cyclist Lance Armstrong was given a diagnosis of testicular cancer in 1996, it had already metastasized throughout his body, including his brain. His doctors gave him less than a 50 percent chance of survival. Mr. Armstrong's subsequent cure can surely be attributed to the chemotherapy he received, but the fact is that other men with similar cases of testicular cancer died despite the same regimen. He has noted his good fortune, saying that his survival was mostly 'a matter of blind luck.' But others shy away from identifying either good or bad luck."

Conceivably in medical matters, we will ultimately discover why some people with certain conditions benefit from certain therapies and others do not. Quite possibly (indeed, probably) the reasons are genetic or something else we've yet to learn about and are not random. But, for now, at least, the efficacy of certain medicines is beyond our understanding and, therefore, *effectively* random.

Completely Random, Effectively Random, Not Random and Part Way Between

> *"'You see,' Max explained as he pumped, 'there's different kinds of dead: there's sort of dead, mostly dead, and all dead.'"*
>
> Miracle Max in The Princess Bride

I should first sidestep the philosophical question of whether people in the real world ever encounter anything that is completely random or completely determined.[11] While I assume that from a logical standpoint the answer is "no," from a practical standpoint the answer is "yes." Let me explain.

[11] The clumsy language ("[do] people in the real world ever encounter anything…") is intentional – an attempt to avoid the question of issues like the speed of light in a vacuum – something that we know with certainty. However, since there is no such thing as a perfect vacuum and non-vacuums slow light down, even the speed of light is not known with absolute certainty *in the real world*.

One might think coin flips are completely random, but it must be true that actual coins cannot be *perfectly* balanced and, thus, are at least a tiny amount more likely to land on one side than the other? The effect might be very small but it is surely not zero. And, as it turns out, other factors are even more important.[12] Still, coin flips are close enough to the ideal that, as a practical matter, we would be foolish not to treat them as random.

Similarly, near the opposite end of the random-vs.-determined spectrum, there are many physical phenomena that one might think always occur in a known way. For example, water "always" boils at 100° Celsius and objects "always" fall to earth at 32 feet per second per second. Except…the boiling point of water varies with altitude, barometric pressure and the purity of the water, and the speed with which objects actually fall to earth varies by latitude and altitude (both of which affect one's distance from the center of the earth) and by other factors.[13]

They may be at opposite ends of the random-vs.-determined dimension, but coin flips and falling objects share one very important attribute: We understand them extremely well and can quite "fully" (or so we would claim) explain how and why they vary. But how about things that fall in between these extremes?

Let's Talk About the Weather[14]

The weather is something that is surely not random and we know a great deal about the forces that influence it. Nevertheless, there are huge gaps in that knowledge, which is why even short-term forecasts so often go awry. When we think about more distant forecasts (e.g., the high temperature in New York City[15] this coming summer – say, July 15), we are even less confident of our ability to forecast, because even though we know the factors that will govern it and we know something about how those forces evolve from one day to the next, there are simply too many days between now and then. Presumably there is a solution to this forecasting problem, but it is beyond our ken.

Still, it's not as if there is nothing we can say about the high temperature next July 15. For example, I know that the past 20 July 15^{ths} had an average high temperature of 86.2° and that that figure would be a much better guess than, say, a random number plucked from the air or, even, the 62.8° average high temperature posted over *all* days (including non-summer days) in New York City over that 20-year period. However, 20 is a very

---

[12] Apparently coins tend to land with the side facing up that was facing up when the coin was flipped. If you find this hard to believe, see http://www-stat.stanford.edu/~cgates/PERSI/papers/headswithJ.pdf. What makes coins "effectively random" is that people don't know this. If they did then flippers would always position the tail facing up, since more people call heads than tails (according to conventional wisdom).

[13] The farther you are from the center of the earth, the smaller the gravitational pull. Since the rotation of the earth causes it to bulge at the equator, the closer you are to the equator (or the higher the altitude of the land that you happen to be on), the slower a given object will fall to earth. Also, FYI, under "standard" conditions, water boils at a bit under 100° and objects fall to earth at slightly faster than 32 feet/sec^2.

[14] It is a (*very*) sad reflection on how long it has taken me to do this study and write this paper, but all of data in this section refer to the 20 years ended December 31, 2005. The reason: my statistical analysis of the weather was done in the summer of 2006 and I wanted to look at the previous 20 *calendar* years.

[15] All references to New York City weather represent readings taken at LaGuardia Airport.

small sample[16] and one can improve the guess (slightly) by using the average of all (620) July days since July 1, 1986 – namely, 85.0°.

This last answer can be improved upon as well. If we examine 5-day, 10-day, etc moving averages of the past 20 years, it becomes clear that one should expect July 15 to be a *little* warmer than the average day in July. (In particular, the early days of July are cooler than the rest of the month.) This factor might boost one's estimate to 85.5° or so. One might also want to add an adjustment for the warming of the earth's atmosphere. In New York City, as it happens, the past 10 Julys have been cooler than the previous 10 (84.4° vs. 85.6°) and the past 10 *years* have also been cooler (62.7° vs. 62.9°), but most other places have warmed up. So, a best guess might be something like 85.7° or so.

Beyond knowing that 85-86° represents a pretty good guess of the high temperature next July 15, we also know that the standard deviation of all 620 July days over the past 20 years has been 6.6°, and since we know how "normal distributions" vary,[17] we can say…

- there is a two-thirds chance that the high temperature that day will be within one standard deviation of the our mean expectation (i.e., between 79° and 92°) and

- there is only a 5% chance – just one in 20 – of being more than two standard deviations from the mean (i.e., below 72½° or above 98½°)[18]

What about super-extreme readings? A three-standard-deviation event (the equivalent of 105° or 66° on July 15) should happen about once every 370 (July) days. With 31 days in the month one might expect such an extreme reading at some point during the month about once every 12 years. However, we have not experienced any readings this hot or cold over the past 20 years.[19] Was this "unlucky"? Perhaps, but as the reader shall see, weather IS streaky. Heat waves and cold fronts are real phenomena and the next time such an extreme condition hits New York, it would not be a great surprise if we had two or three days in a row of such extremes (followed by 30 or 40 years of no such extremes).

Weather as a Model for Streakiness

We know that July is New York's warmest month; we know that Derek Jeter is a far-above-average batter. We know that July's historical average high of about 85° and Jeter's lifetime batting average of .317 are pretty decent forecasts for those values next

[16] For example, the high temperature over the past 20 July 12^{ths} has averaged 82.1°. There is no chance the "true" average temperature rises 4.1° between July 12 and July 15 – i.e., from 82.1° to 86.2°.
[17] I'm assuming the reader is familiar with the concept of the normal distribution. Those who are not and wish to read an explanation can try this link: http://en.wikipedia.org/wiki/Normal_distribution It is, however, pretty difficult slogging and no substitute for once having learned about it in school.
[18] Expressions like "two-thirds…within one standard deviation" and "95% chance…within two standard deviations" represent rounded values. The precise estimates are 68.3% within one standard deviation, 95% within 1.96 standard deviations and, for later in this paper, 99% within 2.58 standard deviations.
[19] 105° or 66° are the applicable three-standard-deviation high temperatures (within a few tenths of a degree) for the final two-thirds of July. However, the beginning of the month is about 1½ degrees cooler and so for early July 104° and 65° are the correct (rounded) three-standard-deviation temperatures.

July. But, what if the temperature next July 12-14 is above 90° every day? What if Jeter goes 7 for 12 for a three-game stretch on the same dates. Is the temperature on July 15 likely to be above average? Is Jeter more likely to hit better-than-usual on July 15?

I'll get to baseball a little later, but let me just establish the fact that there is no doubt whatsoever that weather is streaky. A handful of statistics:

- The correlation between each day's high temperature and the next day's high temperature in New York City has been 0.91 over the past 20 years. To be sure, this connection is "inflated" by seasonal change – i.e., the fact that all January days are colder than all July days boosts this correlation. However, when we restrict ourselves to 30-day stretches[20] (where seasonal change is always minimal) the correlations remain highly statistically significant (0.52).

- Looking solely at July data for the past 20 years the correlation between the high temperatures on successive days is 0.58

- The slope of the difference in high temperature from one July day to the next is 0.56. This means that if one day is 10° above (or below) average, the guess that minimizes your error[21] for the next day is 5.6° above (or below) normal.

- Hotter-than-normal days are followed by hotter-than-normal days 71% of the time; cooler-than-normal has followed cooler-than-normal 70% of the time.

- Looking only at pairs of days where both are either above-average or below-average, we find essentially *no change* (!) in temperature from one day to the next. That is, when a hotter-than-normal day is followed by another hotter-than-normal day, that second day averages only 0.2° cooler than the first of the two days (the average moderates from +5.9° to +5.7°). With regard to below-average days there is no change at all (it stays at -5.8°).

As we can see, weather is extremely streaky. Unless you are tipped off by a meteorologist, you should not expect a really hot day to give way to a succession of days, each of which is a bit less hot than the day before – eventually returning to the average. Rather, the most typical pattern is a bunch of hot days in a row followed by a sharp drop in temperature (to below average) when a new weather system moves through.

I know, I know. You knew all this already. But are ballplayers like this?

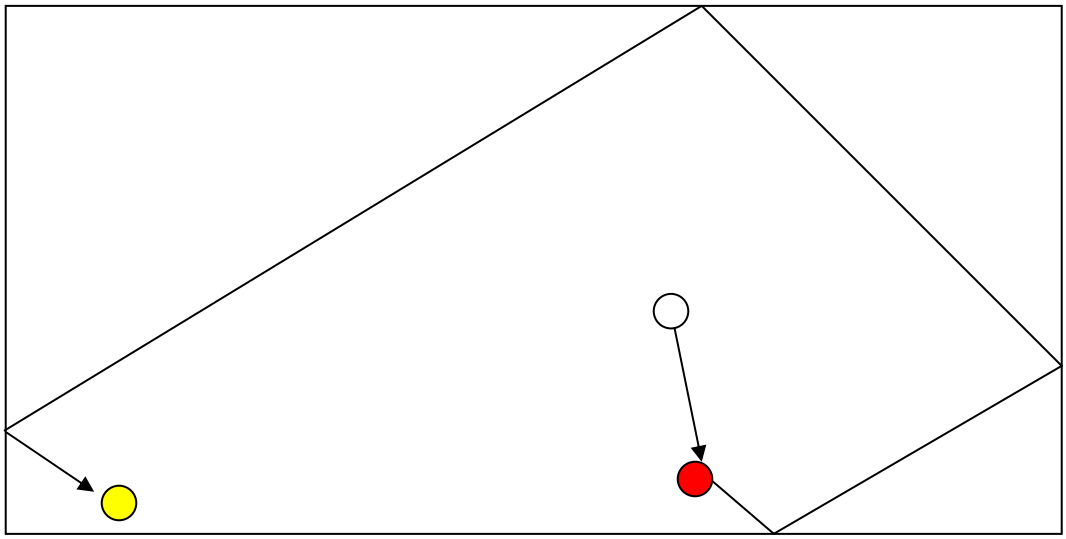
A Billiard Interlude … (I swear I *will* get to baseball)

[I wish to tell a story about myself because I believe the feelings described below are similar to those that everyone who plays a sport or game (even half-seriously)

[20] Looking at Jan. 1-Jan. 31, Jan. 2-Feb. 1, Jan. 3-Feb. 2, and so on, all around the year.
[21] Actually, it minimizes the *square* of your errors.

experiences from time to time. Indeed, I think most people are convinced ballplayers are streaky because they believe they, *themselves*, are streaky *from personal experience*.]

I am an aficionado of three-cushion billiards – a beautiful, complex game in which the object is to use one's cue ball to hit the other two balls on the table. To score one must hit at least three cushions before striking the second of the two object balls. (The first object ball can be hit before or after one or more cushions has been hit.) Here is an example of a shot where four cushions are struck before the final ball is contacted. (The white ball is the cue ball and the colored balls are the object balls.)



I play about 10 hours a week and have been keeping careful records of my performance since 1993. Over the past decade I've made approximately 41% of my shots. In a normal session I play for four-five hours and take anywhere from 200 and 300 shots. On a typical day I might make, say, between 95 and 110 of 250 shots.

However, most of the time I do not perform very close to my average performance. Sessions are littered with episodes – covering 20 shots, 30 shots, or sometimes more – when I get "hot" (making more than half of my shots) or cold (say, one-third or less). On a great day, I might have three or four hot stretches and no cold periods – on a terrible day, the opposite.

Part of this variation clearly has to do with luck. Shots in three-cushion billiards are so complicated (involving three balls, multiple cushions and a cue ball that often has to travel more than 20 feet) that it is very difficult to "play position" – and when you have the opportunity to do so, you are doing it only in a most approximate way. Change the position of just one of the three balls by a few inches – something that even the best players in the world cannot control – and a difficult shot becomes easy (or vice versa). Moreover, there are lucky shots where you intend to make a shot one way but hit the ball so badly that you make it an entirely different way. In the most common variant here, you "kiss" the first object ball, get knocked around the table, yet end up making the shot

anyway. (A "kiss" happens when you hit the first object ball more than once…generally, after either the object ball or your cue ball or both have hit one or more cushions.)

But luck "evens out" – everybody knows that! – and, so, when I have one of those days where I am scoring like mad (or "can't score at all") the feeling that I am "in the zone" (or "suck") *and that those feelings are influencing my performance*, becomes overwhelming. It is almost impossible to escape the feeling that psychology matters.

The problem with this conclusion, however, is that a study of the statistical properties of my performance (since the summer of 1998) shows that it is essentially random. A player making 41% of his shots is supposed to make (to cite a few examples)…

85 or more of	250 shots	99.0% of the time
90	250	95.4
103	250	49.8
115	250	6.2
120	250	1.5

…and if you look at the distribution of my results over the years, you will find results that are quite close to these predictions. Admittedly, there has been a *slight* tendency to produce a few extra extreme results. For example, in my past 799 sessions instead of performing outside the 95% confidence interval 5% of the time, I have been outside it 6.1% of the time.[22] However, the extra 1.1% "extreme" results are not sufficient to achieve even minimum evidence of streakiness.[23]

Conclusion:

- My performance at the billiard table is random around my mean performance;
- External factors that may influence my performance on any given day (how many easy or difficult positions I have, how many lucky shots I happen to make, even things like how I feel physically or how many distractions there are in the billiard room) are randomly distributed and uncorrelated with each other and, thus…
- *When I feel like I am in "the zone" I feel that way because I've been scoring a lot. I do NOT score a lot because I AM in the zone.*

The upshot is that while I am *emotionally* convinced that I am streaky, I know intellectually (for a fact!) that I'm not.

Of course, this is just me. While my performance in billiards appears to be random around my mean performance, I don't have the data to test this for other billiard players. Fortunately, data is readily available for something that interests many more people – namely, baseball.

[22] Instead of 20 excellent sessions and 20 very poor sessions, I've had 27 and 22 respectively.
[23] If extreme results occurred 6.6% of the time it would be somewhat convincing; at 7% quite convincing.

Measuring Streakiness – How Many At-bats?

How long does it take for a player to get hot or fall into a slump? Are two at-bats enough? Are five? 10? 25? 50? I have tried to satisfy the statistical Gods as well as the people I've been corresponding with by examining various numbers of at-bats.

From a statistical standpoint, any number of at-bats can be tested for streakiness. But, baseball fans seem to believe that stretches of at-bats that are very short or very long are either not sufficient to put a player into a "groove" or "slump" or too long to be relevant to the concept. From my e-mail correspondence I gather that most fans believe that stretches of at-bats between 8-10-12 on the short end and/or 25-35-50 on the long end are appropriate for examining the hypothesis of whether players are streaky.

Below I report mostly on streakiness during stretches of 10 and 25 at-bats. I've steered clear of very long stretches because they produce samples too small to be statistically significant. Even players who miss very few games rarely bat much more than 650 times per season.[24] If we used 50 at-bats, our sample would consist of only 13 (non-overlapping) stretches in a full season; by using 25 at-bats, we double the number of stretches. (From a statistical standpoint 26 is not very large either, but it's better than 13.) Meanwhile, I assume most who believe that 35 or 50 at-bats would be *best* for measuring streaks will recognize that 25 at-bats would, at the very least, represent an acceptable length for testing the phenomenon.

Indeed, on the same logic I would expect that if streakiness exists because of positive or negative psychological feedback loops, then it should exist for *any* stretch of at-bats, regardless of how short – even as few as two! For that and two other reasons, I also examined streakiness over stretches of pairs of at-bats. The other reasons: (1) pairs of at-bats are least affected by opposing-pitcher variability – a factor that, as will be explained below, *reduces* streakiness; and (2) pairs produce very large sample sizes for individual batters.

Measuring Streakiness – What Criteria To Use for Determining Whether It Exists?

In July 2006, I polled a number of friends and relatives on whether they believe streakiness exists. Among other things I asked the following:

> Do you think that batters are in fact streaky (in the sense that they have more stretches of multiple at-bats where they do well or poorly than chance would suggest)?
>
> a) All players (or the vast majority) have a tendency to be pretty streaky.
> b) A significant proportion of batters are streaky.

[24] In the season I studied in detail (2004) only seven batters exceeded 650 at-bats and only four exceeded 657. The leader, Ichiro Suzuki, had 704 at-bats. Note that part of the reason the numbers aren't greater is that players are not charged with an official at-bat when they walk, sacrifice or are hit by a pitch. "Total Plate appearances are usually 50-100 higher for most full-time players. In 2004, Ichiro also led in this category with 762 plate appearances while 17 other players ranged from 700-748 and 40 additional players ranged from 651-699.

c) The batting process is pretty random – i.e., there will be no more "hot streaks" than mere chance would suggest.
d) Other?

Of the 28 respondents who answered that question, 21 answered a, b, or d with some additional comments that made it clear they believe that players (at least many of them) are streaky. For additional clarification on what people had in mind when they think about streakiness, I asked the following two-part question:

> Whether or not you believe that streakiness exists, what do you believe *would constitute* streakiness? (I'll flesh this out via an example of a batter who hits .300 for the full season but who from time to time shows evidence of being "hot" by getting 5 hits in his last 10 at-bats. The question arises, what should we make of this?)
>
> a) A .300 batter who has 5 hits in his last 10 at-bats can be expected to get a hit how often in his 11th at-bat? Answer this question by telling me his likely batting average in that 11th at-bat.
>
> [Note: an answer of .425 or .375 or .325 or .305 might each reflect a belief that "hot" and "cold" are valid concepts – but .425 is obviously a much (*much*) hotter response than .305. On the other hand, perhaps you believe that as a result of the "law of averages" he'll bat only .250 in those 11th at-bats. If you believe that, you believe in the OPPOSITE of streakiness.]
>
> b) What is the MINIMUM batting average that you would want to see to be convinced that streakiness exists?

There were 25 respondents who gave estimates of expected batting average for players who have just gotten five hits in their previous 10 at-bats. Of the 18 who said they believe players are streaky, here is a summary of what they said for those 11th at-bats:

Mean	.335
Median	.329
High	.400
Low	.300 (two think streakiness exists but 5/10 won't make a player "hot")

Not surprisingly, five of the seven responders who don't believe players are streaky said their expectation for the 11th at-bat was exactly .300. One of the responders believes in the "opposite" of streaks – i.e., a player who has gone 5/10 becomes *less* likely to get a hit in his 11th at-bat and that respondent expects the 5/10 batter to average .250 in that 11th at-bat.

To the question of what batting average one would require to become convinced that streakiness is a genuine phenomenon, I will again separate the answers of the 18 who believe streakiness exists from the six who do not. (One of the non-believers who answered the previous question chose not to answer this one.)

	Believers	Non-Believers
Mean	.339	.397
Median	.325	.335
High	.500	.500
Low	.302	.315

As you can see, the non-believers not only think streakiness does not exist, but they demand very strong evidence before they will be willing to change their mind.

What might a statistician say? From the standpoint of statistical significance, a great deal depends on sample size. A .300 batter who bats 550 times during the season will have stretches of 11 at-bats 540 times (i.e., at-bats #1-11 is the first such stretch; the second is at-bats #2-12, and so on, with the final stretch being at-bats #540-550). In a random distribution, a .300 hitter would be expected to go 5/10 just over 10% of the time – i.e., approximately 55 times during a season of 550 at-bats. Let's say that a particular batter, in fact, accomplishes this feat exactly 55 times. What does his batting average have to be in those 55 eleventh at-bats for a statistician to report that the batter is more than two standard deviations above "normal"? The answer is that he would have to go 24 for 55 which is a batting average of .436. So, for a single batter in a single season to show evidence of streakiness, the batting average we need to see is, indeed, very high.

Notice that a batter who bats .382 (21/55) would be well above what fans who believe streaks exist generally require to confirm their beliefs. (.382 is, in fact, high enough that only four batters have managed such a high average over the past 65 seasons.[25]) But 55 at-bats is only about one-tenth of a typical season and a statistician would tell you that a .300 batter who goes 21/55 is "only" 1.32 standard deviations above his normal performance – something that is quite common. (A random batting process will generate performance that "good" or better about 12% of the time for clusters of 55 at-bats.[26])

Baseball is so chock full of "situational statistics" I decided, as an experiment, to examine the ESPN website and check out the five players who happened to bat exactly .300 in 2006 and look for circumstances under which they batted (a) at least .360 and (b) exceeded expectations by 1.32 standard deviations or more. A *partial* list of these "above-normal" accomplishments is presented below. (I ignore stats that might be *expected* to produce superior performance – such as righties batting against lefties or batters hitting with a count of two balls and no strikes – and I limit myself to a maximum of three "notable accomplishments" per batter.)

[25] Since Ted Williams hit .406 in 1941, the only players with enough plate appearances to be eligible for the batting championship and who hit higher than .382 were (1)Ted Williams .388 in 1957 (at the age of 38!); Rod Carew .388 in 1977; George Brett .390 in 1980 and Tony Gwynn .394 in 1994. However, both Brett and Gwynn played "partial" seasons – Brett due to injury and Gwynn due to a player's strike.

[26] The greater the number of at-bats, the less common +1.32 standard deviations or better will be, but even at 550 at-bats – more or less a full season, it will happen a bit more than 10% of the time. For smaller clusters, the frequency rises (e.g., for 10 at-bats, 1.38 or more standard deviations will occur 15% of the time.)

Should we really be impressed (as you can see in the last row of Table 1) by the fact that Carlos Lee batted .778 with runners on 2nd and 3rd? Of course not; he was in that situation only nine times. There are so many different "splits" measured in baseball – each player batted against 18-20 other teams, played in almost as many ballparks, played in six separate calendar months, batted in situations with the bases empty or with a runner on 1B or a runner on 2B or runners on 1B and 2B etc. *ad nauseum*. If you look at enough data, some are *bound to look way above normal, even if the entire process is random.*

Table 1 – Selected Extreme Performances by .300 Batters in 2006

	At-bats	Hits	Average	# of Standard Deviations
Jamey Carroll				
month of June	108	39	.361	1.39
vs. Washington	31	14	.452	1.84
vs. Pittsburgh	20	10	.500	1.95
Jose Reyes				
month of June	110	41	.373	1.66
vs.Toronto	14	9	.643	2.80
2 out, men in scoring position	70	29	.414	2.09
Ivan Rodriguez				
vs. Minnesota	58	23	.397	1.60
vs. Seattle	25	12	.480	1.96
man on 1B only	119	46	.387	2.06
Rafael Furcal				
month of September	111	41	.369	1.59
vs. New York Mets	29	12	.414	1.34
men on 1B and 3B	12	7	.583	2.14
Carlos Lee				
vs. Chicago Cubs	42	17	.405	1.48
at Minute Maid Park	11	6	.545	1.78
men on 2B and 3B	9	7	.778	3.13

With hundreds of players and scores of ways to break down the statistics, it is inevitable that we will see many pieces of evidence potentially consistent with the hypothesis that players are streaky. This explains why baseball conventional wisdom is steeped in the notion of streaks, "clutch" hitting and the like. But the only way to fairly assess whether these ideas are valid, is to look at all the data for many players and conclude that players are streaky only if the totality of the data is consistent with streakiness.

When we move from individuals to the group an interesting thing happens – namely, the batting averages required to provide evidence of "hotness" drops considerably. The reader will recall that a .300 batter needed to bat .436 over 55 at-bats to produce statistically significant evidence of being hot. Let us contrast that with, say, what might be required from a group of 200 batters who bat .300. This group might have 11,000 at-

bats in a single season (i.e., 200 x 55) where they come to the plate having gotten five hits in their previous 10 at-bats. What do they have to hit to demonstrate evidence of being "hot"? Instead of .436, they need only bat a collective .309!

Of course, when we examine the data we will have collective data AND individual data. Because of small sample sizes the individuals are likely to be all over the lot. But if many or all players are in fact streaky, (a) the group data will almost surely show it (because the sample size is so large) and (b) more individuals are likely to appear to be streaky than would be expected by a random distribution.[27]

One final consideration: Let us say that all batters, after going five for 10, collectively batted .308 over the course of 11,000 at-bats in a single season, not quite reaching the .309 required to attain the 95% confidence level. We could always look at more seasons. Let's say that we examined 10 seasons and discover that over 10 years they averaged .306 or .305. Numbers like that over that many seasons (with a huge sample of 110,000) would handily provide the statistical support for the notion of streakiness. Indeed, even .303 would do for a sample that large. But would baseball fans care? I doubt it, because an extra .003 is not what fans have in mind when they believe a player is "hot."

The Role of "Fundamental" Contributors to Streakiness

There are a number of factors, other than psychology, that either contribute to or reduce streakiness. Paradoxically, the very *same* factor can sometimes add to streakiness and sometime subtract – with the effect varying according to the number of at-bats being examined. I can identify at least seven factors UNrelated to the psychology of performance that might affect the number of "hot" and "cold" streaks that we see.

1. Home-Away effects – players have historically hit better at home than on the road for at least two reasons: They are more familiar with the ballpark and they are less likely to be fatigued by recent travel.

2. Ballpark effects – some ballparks are "hitters' parks" because they have shorter fences and/or smaller foul territories and/or more closely cropped grass and/or more favorable weather conditions than "pitchers' parks."[28]

3. Some pitchers are better than others – batters will, obviously, tend to perform less well against above-average pitchers than against below-average pitchers;

4. Pitcher effectiveness declines the longer the pitcher stays in the game, partly because pitchers tire as they throw more pitches and partly because batters

[27] In a random distribution 2½% of all batters would appear to be streaky at the 95% confidence level and 2½% would appear to be the *opposite* of streaky. But if genuine streakiness exists, perhaps 10% (or even 30%) would appear to be streaky while only 1% (or ¼%) would appear to be the opposite.

[28] Many pitches are popped up into foul territory. If the stands are close to the field fewer of them will be caught and when they are not caught, the batter gets another chance to get a hit. Separately, when the grass on the field is short, ground balls have a better chance of getting through the infield.

become more familiar with a pitcher's motion with each successive at-bat.[29]

5. Defensive skills vary – some fielders cover a lot more ground than others, thereby preventing hits that other fielders would not be able to prevent.

6. Injury – if you play when you're hurt your performance is negatively affected.

7. Stress unrelated to baseball – e.g., illness or death of a loved one, marital difficulty, etc. can negatively affect player performance.

A .300 hitter leading off a game on the road, in a pitcher's park, against a great pitcher backed by a solid defense surely has less than his usual 30% chance of getting a hit. Indeed, under these conditions, the odds are probably below 25%. With a sore wrist and a gravely ill parent, his chances might be less than 20%.

To gain an understanding of the size of these effects, let's review some simplified hypotheticals. Let's say that a .300 hitter who actually bats .300 for the season, hits only .250 during the first half of the season owing to a combination of factors (strong opposing pitching, nagging injuries, etc.) and bats .350 during the second half of the season. Will he perform more streakily than a batter who hits .300 all season long? Very probably.

Consider the shortest possible streak – two straight hits. A player who is "always" (in an underlying sense) a .300 hitter will be expected to accomplish this 9% of the time:

.300 x .300 = .090

But a batter who hits .250 from April-June and .350 from July-September will be expected to get two hits in a row 6.25% of the time during the first half season and 12.25% of the time during the second half..

.250 x .250 = .0625 and .350 x .350 = .1225

If you spend the first half of the season getting two hits in succession just 6.25% of the time but the second half getting two in a row 12.25% of the time you will get two straight hits during 9.25% of your at-bats during the full season. [30]

(.0625 + .1225)/2 = .0925

---

[29] Every pitcher's motion is unique and it is rare for a player to face the same pitcher in more than four games per season. It is not even all that common for a player to face the same pitcher more than twice.

[30] OK, OK, pipe down. It's not exactly 9.25% of the time. Take, as an example, a player with 560 at-bats who bats .250 for his first 280 at-bats and .350 for his last 280 at-bats. He will be expected (from a random-statistical standpoint) to perform as follows: As a .250 batter over his first 280 at-bats or 279 *pairs* of at-bats, we expect 17.4375 two-for-twos (.250 x .250 x 279); as a .350 batter over his final 280 at-bats or 279 pairs, we expect 34.1775 two-for-twos. For the one pair of at-bats (at mid-season) when he is transformed from a .250 batter to a .350 batter his chances of going two-for-two are 8.75% (.025 x .035 = .0875 and, thus, the total number of expected two-for-twos is 17.4375 + 34.1775 + .0875 which equals 51.7025 which is 9.249% of the 559 pairs of at-bats during the full season.

This is almost 3% streakier than someone who is "always" a .300 performer:

9.25%/9.00% = 1.028

Now let us consider a player who keeps alternating *from one at-bat to the next* – i.e., he performs like a .250 batter during odd at-bats and like a .350 batter during even at-bats. Such a player will get two hits in a row only 8.75% of the time…

.025 x .035 = .0875 and (of course) .035 x .025 = .0875

…which is almost 3% *less* streaky than the steady .300 batter would be:

8.75%/9.00% = 0.972

We can see from the above that the more variable underlying conditions are, the greater the potential for *either* fundamentally-produced streakiness OR fundamentally produced *anti*-streakiness! The results we end up with will depend on the periodicity of the factors listed above and the length of the stretches of at-bats that we examine.

For example, if a particular set of conditions that sometimes helps batters and sometimes hurts them typically alternated every five at-bats, it would tend to increase the amount of streakiness we would observe if we focused on all pairs of at-bats over the course of a season but diminish the amount of streakiness we observe if we focused on all stretches of 25 at-bats over the course of the season. With this in mind, how long do the seven factors mentioned above persist and what is their role in streakiness?

1. Home-Away effects – the average homestand and roadtrip in the major leagues in 2004 averaged 6.6 games or about 22 official at-bats. The vast majority lasted between three and 11 games (i.e., eight at-bats to 40 at-bats) and only 4.2% were longer than 11 games or shorter than three.[31] As a result, any benefits of playing at home are likely to boost the amount of streakiness we see over stretches of two at-bats and 10 at-bats, but not have much influence over stretches of 25 at-bats.

[31] The longest homestands included five that were 14 games, 11 that were 13 games and seven that were 12 games. At the short end, there were 11 two-game series and a pair of one-game series. Both of the latter and most of the two-game series were the result of rain-related postponements. A few of the long series were lengthened by one game because of rain-related doubleheaders related to previous postponements. With regard to the estimates of the number of at-bats over 6.6 games, I have used the average number of at-bats by the players I studied (547) and divided by the average number of homestands and roadtrips for major league teams (24.6). The reason the number is a bit lower than one might guess (working out to only 3.4 at-bats per game is that even the "starting players" that I used in my study do not play every game. For the estimated number of at-bats in 3-game and 11-game stretches, I used four at-bats per game *played* and assumed that most regular players play at least two games of any 3-game stretch and few play all 11 in a typical 11-game stretch. Finally, in the spirit of full disclosure, I should add that my analysis of the length of homestands and roadtrips was done by looking at game-by-game logs of the 2004 season and I'm not 100% certain my statistics are completely accurate. However, I am certain that any errors I may have made here cannot be meaningful relative to the conclusions that I am drawing.

2. Ballpark effects – these are similar but not identical to home-away effects because while the average homestand is 6.6 games and most range from 3 to 11 games, the vast majority of visits to other ballparks are exactly three games and almost none are shorter than two games or longer than four. Thus, the average ballpark "stay" is probably a bit under 4½ games. As a result, ballpark effects probably reduce streakiness (slightly) over stretches of 25 at-bats, while continuing to boost it over stretches of two and 10 at-bats.

3. Some pitchers are better than others – this is surely the most important factor *reducing* the amount of streakiness we see. The quality of opposing pitching often changes materially from one game to the next and, even within a single game, it is common for a batter to face three different pitchers of quite disparate ability. Most of the variation has to do with pitcher skill, but part has to do with which side pitchers throw from. That is, right-handed batters tend to perform better against lefties than righties, and left-handed batters do better against righties than lefties. The upshot is that opposing pitcher variability reduces streakiness for all stretches of at-bats that are longer than three and, possibly, two.

4. Pitcher effectiveness declines the longer he stays in the game – thus even in stretches as short as two at-bats, the variability of single-pitcher effectiveness has a slightly moderating impact on streakiness.

5. Defensive skills vary – this will tend to generally reduce streakiness because individual fielders vary in their effectiveness.

6. Injury – the longer that injuries last, the more they promote streakiness. If the reader thinks this can only promote "bad streaks" (i.e., slumps) consider the following hypothetical: A .300 batter may average .300 because during the 80% of the time he is in good health he bats .310 and the 20% of the time he is bothered by minor injuries he bats .260. He is not only more prone to going 2/25 (or 0/2) during a .260 phase than a .310 phase, but he is more likely to go 12/25 (or 2/2) during a .310 phase than during a .260 phase *and* he is more prone during his .310 phase to go 12/25 or 2/2 than he would be if he was also a steady .300 batter.

7. Stress unrelated to baseball – just like injury, the longer situations like these persist, the more they boost streakiness.

How important are the seven items listed above? For two of the items – injury and stress – we can only guess. Major injuries and life events (deaths, births) are routinely reported but don't matter very much because players don't play when this happens. The lesser injuries and stresses that players continue to play with are rarely reported and even when we learn about them we don't know how to estimate their effects.

Defensive skills can be measured, but we do not have the data needed to know whether a batter got a hit or was denied a hit because of a particular fielder's ability. To do so, we

would need to know *exactly* where every ball was hit and the fraction of fielders at the pertinent defensive position who would have made the play. Moreover, from other literature and our own findings in this study (about the impact of batting average variability on streaks), the impact of this factor would surely be immaterial.

With regard to the first four fundamental factors listed above, we are able to "ballpark" (heh, heh) the effects though a combination of actual baseball statistics from the 2004 season and hypotheticals. Since the effects are quite (surprisingly?) small and discussion quite (unsurprisingly?) long, I've moved the latter to Appendix A – beginning on page 41. For now, I will simply summarize the results:

(1) Home-Away Effects – the average home batting average exceeds the average away batting average by only .006 (i.e., .269 vs. .263). This increases streakiness by a trivial amount. (Note: .266 squared is .070756 while the mean of .269 squared and .263 squared is .070765. Applied to 167,353 at-bats over the 2004 season, this difference is worth about 1½ extra streaks of two-straight hits. However small that impact is, the effects are diluted further whenever we examine stretches of at-bats that are greater than two.

(2) Ballpark Effects – vary significantly from ballpark to ballpark and in the aggregate are about 20 times as great as home-away effects, but (a) they are substantial at only a few ballparks and (b) in the aggregate would add only modestly to streakiness (perhaps amounting to one-seventh or so of what is needed to be statistically significant).

(3) Pitcher Variability – the dispersal in batting averages allowed by pitchers is much wider than "ballpark effects" a factor that *would* lead to a great deal of streakiness if batters faced great or horrible pitchers for many at-bats in a row. But they do not. The maximum effect promoting streakiness is surely for the shortest possible stretch of at-bats (two) where the number of two-for-twos is probably boosted about 1% by the variability in pitcher skill. For all longer stretches of at-bats, streakiness is surely reduced by pitcher variability, but the effect is very small.

(4) Pitcher Fatigue – face the same pitcher more than once and his effectiveness declines. This reduces the number of two-for-twos.[32] However, this only offsets about one-tenth of the short-term streak-boosting impact of pitcher variability.

The upshot is that these four effects taken together are expected to boost the number of very short-term streaks (two-for-twos and their complement, oh-for-twos) and have very little impact over longer stretches like 10 or 25 at-bats. None of these impacts will be even close to being statistically significant, let alone produce deviations in batting performance that would comport with fans' views of what streakiness is.

[32] If you're wondering "doesn't the fact that a batter is more likely to get a hit in his second at-bat increase the number of two-for-twos?" The answer is "yes" BUT, the fact that the batter was more likely to make an out in his first at-bat (when the pitcher was more effective) reduces the number of two-for-twos. The latter effect is more important than the former.

Does Psychology Trump the Fundamentals or Vice Versa? – Or, Can a Belief in Streaks Undermine Their Existence?

A hot or cold (mini) streak might arise initially as a product of any of the fundamentals mentioned in the last section and then blossom into a full-fledged hot or cold streak when positive or negative psychological feedback loops amplify the effects of these underlying fundamental factors. For example, a player bothered by a sore shoulder might hit worse than the physical injury *alone* would suggest if he loses confidence as well. Before you can say "Jack Robinson," he's pressing and he's in a slump. Indeed, about 20% of the people who responded to my poll made comments along these lines and I would not be surprised if many respondents believe this to be true but just didn't bother to say so. Indeed, the essence of the claim that athletes are streaky is the notion that positive or negative psychological feedback loops (a) happen and (b) have a significant influence on performance.

While it is undeniable that players sometimes feel better or worse about themselves and it almost *has to be true* that that actually has some influence on performance, it is not necessarily true that the effect is large. In fact, it is not even necessarily true that the effect is in the direction most people believe.

For example, what if success at the plate (say, six for eight over a two-game stretch) leads to *over*-confidence and a cavalier attitude at the plate over the subsequent few games? Similarly, going 0/8 over two-games may prompt the majority of players to take extra batting practice and concentrate harder at the plate. Perhaps this outweighs any "negative thoughts." If these reactions are typical, then it's possible that success breeds failure and vice versa – which is the *opposite* of the players-are-streaky hypothesis. If this is true then ballplayers will have *fewer* protracted streaks than would be expected by chance.

Measuring Streakiness or How Should We Estimate "Expected" Batting Average?

In the poll that I e-mailed last summer, I asked the respondents to consider the case of a .300 hitter who has a stretch where he gets five hits in 10 at-bats and asked them to estimate the likelihood of his getting a hit in the 11^{th} at-bat. Any answer above 30% would presumably reflect a belief that players have a tendency to "get hot." Any answer below 30% would suggest a belief in mean reversion ("the law of averages" in common parlance) and an answer of 30% would suggest the respondent believed that the batting process was random. In posing the question, however, I was careful *not* to define exactly what I meant by a ".300 hitter." The reason: It's almost impossible to know what a batter's true capabilities are at any given moment.

When a batter comes to the plate, what are the chances he will get a hit? On the face of it, it does not seem unreasonable to suppose that a player's lifetime batting average might be a pretty good predictor. At the very least, you are working with a large sample. However, players get better and worse over the course of their careers and so their lifetime batting average is usually biased – one way or another – during any particular plate appearance

during any particular season.[33] For this reason, many observers (and statisticians) would be inclined to use results from the current season only. But "sample-size" (of number of at-bats) is not very large for a single season, particularly early in the season.

The earlier in the season and/or the "hotter" or "colder" the player has been lately, the *less* representative his current-batting average will be. How great a distortion results from "recent performance?" How does this phenomenon evolve over the course of a season? And, how does this bear on the way we should measure streakiness?

Consider the most extreme case – i.e., a player who gets a hit in his first at-bat of the season and who, thus, sports a 1.000 batting average the next time he comes to the plate. I estimate that in a typical major league season close to 200 players will have this experience.[34] Obviously, the collective batting average of this group will be much, much lower than 1.000 during their second at-bats. (Indeed, they will very likely bat less than .300.) Of course, we are (mostly) concerned not with individual at-bats, but groups of at-bats. So let's return to the question of the .300 batter who goes five for 10.

There are about 420 position players on major league rosters at any given time. If they were all .300 hitters then *simply by chance* we would expect 10.1% of them – i.e., 42 or 43 players – to get 5 hits in their first 10 at-bats of the season. In an effort to stick to "round" numbers (which will make this part of the discussion a bit easier to follow), let's say that exactly 40 of them began the season in this fashion. Let us stipulate for the sake of argument, that the results so far have been mere luck. The question now, however, is "have they gotten 'hot' by virtue of starting the season 5/10 – i.e., has their early success put them in a positive frame of mind thus making them "more likely than usual" to get a hit in their 11th at-bat? And, how can we measure success?" If in their 11th at-bats, our group of 40 gets 18 hits, they will have batted a collective .450 (i.e., 18/40). There are two things we can say about that: (1) It is a fantastic batting average but (2) It is *below* the .500 batting average they each had as they came to the plate for their 11th at-bats. Clearly – since no player in modern baseball history has ever hit more than .424 for a

[33] Batting averages generally improve from a player's rookie season into his late 20s and commonly fall off after his mid 30s. Thus a 28-year old who enters a season with a lifetime batting average of .300 will more likely than not hit more than .300 while a 39-year old entering a season with a lifetime average of .300 should be expected to hit less than .300.

[34] The 30 major league teams have 270 players in their collective batting orders during the first game of the season. About 75 of them can typically be expected to get a hit in their first at-bat. Over the course of the season, however, many more players come to the plate. Reserves play, players get called up from the minor leagues and most pitchers get to bat. During the 2006 season, 976 players got at least one official at-bat. Of these, 792 got at least one hit, and 785 of them got two or more official at-bats – i.e., they had a chance to follow their first hit with a second. Of the 785 with at least one hit and two at-bats, 621 were position players and 164 were pitchers. I do not have the sequence of at-bats, but judging from the typical batting averages for position players and pitchers, I'll guess that about 170 position players and 25 pitchers were batting 1.000 when they went to the plate for their second official at-bat of the season. (For those of you who are really "into" this – and, if you're reading this deep into footnotes, I can tell that you are – I should inform you that of the 75 players with only one official at-bat, seven of them got hits. This group included 72 pitchers, one catcher and two pinch-hitters. All seven of the players who were 1/1 for the season were pitchers. Of the 116 players with two or more official at-bats but zero hits, 101 were pitchers most of whom had fewer than 10 at-bats. And, every non-pitcher with zero hits had fewer than 10 at-bats.

season – a batting average for any set of bats that is .450 is higher than one should expect. Our measurement system must recognize this and compensate for this phenomenon.

How big is the effect that we are talking about? Tables 2A and 2B show the effects of various "hot" and "cold" stretches at various stages of the season. In all cases, the hypothetical player was batting .300 before having either the good stretch or the poor one. For example, in the left-hand column of Table 2A, we examine a player who was batting .300 after starting the season with 12 hits in 40 at-bats. In the bottom row of the table we calculate that when this players proceeds to get 18 hits in his next 25 at-bats, he boosts his season-to-date average to .462 (because he would be 30 for 65). Even late in the season (after 480 at-bats), a very modest "hot streak" (5 for 10) will boost the batting average from .300 to .304. (See the upper right-hand corner of Table 2A.) In one of the middle cases, a player who is batting .300 after half a season (260 at-bats) and goes 13 for 25 will raise his average to .319.

Table 2A – The Impact of Hot Streaks on Batting Average

	After Hitting .300 for These Many At-bats				
At-bats	40	150	260	370	480
Hits	12	45	78	111	144
And Then Going	**The Player's Batting Average Rises To**				
5 for 10	.340	.313	.307	.305	.304
6 for 10	.360	.319	.311	.308	.306
7 for 10	.380	.325	.315	.311	.308
8 for 10	.400	.331	.319	.313	.310
9 for 10	.420	.338	.322	.316	.312
11 for 25	.354	.320	.312	.309	.307
12 for 25	.369	.326	.316	.311	.309
13 for 25	.385	.331	.319	.314	.311
14 for 25	.400	.337	.323	.316	.313
15 for 25	.415	.343	.326	.319	.315
16 for 25	.431	.349	.330	.322	.317
17 for 25	.446	.354	.333	.324	.319
18 for 25	.462	.360	.337	.327	.321

Reflecting on Table 2A one can see that whenever a player of a given skill has a stretch of at-bats where he hits above his "underlying" average, his season-to-date average will rise above his "underlying average." We can see the mirror-image of this phenomenon in Table 2B (next page): any player who has a stretch of at-bats that is below his underlying batting average will have his season-to-date average depressed. The longer the stretch and the earlier in the season this occurs, the greater the effect.

Table 2B – The Impact of Cold Streaks on Batting Average

	After Hitting .300 for These Many At-bats				
At-bats	40	150	260	370	480
Hits	12	45	78	111	144
And Then Going	The Player's Batting Average Declines To				
0 for 10	.240	.281	.289	.292	.294
0 for 25	.185	.257	.274	.281	.285
1 for 25	.200	.263	.277	.284	.287
2 for 25	.215	.269	.281	.286	.289
3 for 25	.231	.274	.284	.289	.291

The upshot is that whenever we are looking at a batter who has recently had a very good (or very poor) stretch, we know that his batting average has been upward (or downward) biased by that very fact. The problem is that we don't know by how much because we never know what any player's "true" underlying skill level is. We can only estimate it.

My general approach to addressing this problem is to use season-long batting statistics (for 2004, of course), *sometimes* removing the "hot" or "cold" stretch from the data. No data is removed when we measure whether batters have more "hot" streaks" (e.g., 2/2, 6/10, 13/25) or "cold" streaks (e.g. 0/10, 2/25) than would be expected by chance; we base the estimate on the player's season-long batting average. However, when examining how a player does in the aftermath of a "hot" or "cold" stretch (e.g., the 11th at-bat following a 7/10 stretch or the 26th at-bat following a 2/25) I take their season-long batting results and eliminate the hot/cold stretch as well as that next at-bat (11th or 26th). The rationale for the latter may not be obvious so let me flesh it out via an example.

Smith has a good year, batting .300 – 150 hits in 500 at-bats. However, during one stretch he was "red hot" – going 15/25. How do we suppose he batted in that 26th at-bat immediately after going 15/25? If we think batting is a random process should we think that his chances of getting a hit in that 26th at-bat are 30%? No, because we KNOW (now that the season is over and that his totals were 150/500) that in his other 475 at-bats he got "only" 135 hits for a .284 batting average. Thus, if he (and other players in the same situation) bat .284 in those (collective) 26th at-bats, then that would be consistent with random (i.e., non-streaky) batting performance. IF he bats .300 (or even .295) it would suggest that Smith is, indeed, hot because he's batting better in the 26th at-bat than in his "average" at-bat over the entire *remainder* of the season. Indeed, because the 26th at-bat *might* be part of the hot streak (we won't know until we've completed our study), the result for that at-bat should also be excluded from the "null" batting average that we expect Smith to produce in his streak-related at-bats.

***Which* Players Should Be Examined for Streakiness?**

It seems obvious that our analysis should be confined to "everyday" players. After all, how does one "get hot" if one plays every other day? Thus, I decided that I would

restrict myself to players with 450 or more official at-bats plus Barry Bonds. (In 2004, Barry Bonds had only 373 official at-bats but he walked a record-setting 232 times. His total plate appearances – 617 – which includes hit-by-pitch and sacrifices, placed him 84th in the Major Leagues and very near the midpoint of my 160 "everyday" players.)

More than 92% of the players with 450 official at-bats played in at least 80% of their teams' games and only one player (Kazuo Matsui – the horrible second baseman for my beloved NY Mets – who *rarely* met a pitch he would not swing at) is the only player to make the list with fewer than 120 games played.[35] (120 games is 74% of the season's 162 games.)

How Many Players Will Look Streaky if Streakiness Does NOT Exist?

It seems pretty obvious, but I'd better point out that if something is expected to happen *by chance* about 5% (or 1%) of the time, then *that's what you should expect.* In other words if something is unlikely (even very unlikely), it should still happen *occasionally.* Good hitters are *supposed* to get four hits in a row about 1% of the time.[36] When they do – whether it's a bit more than 1% or a bit less – it does NOT mean that they are streaky. Rather, it confirms the absence of streakiness (and presence of randomness).

Incidentally, if we examined a sample size of 1,000 for such a batter, we would expect to see four hits in a row about 10 times and the question of whether a higher number was high *enough* to be significant would be a statistical question. For the reader's information, in a statistical sample of 1000, we would need to see 17 (instead of 10) 4/4s to conclude (with 95% confidence) that streakiness is present. Increase the sample size and the required excess increases with the square root of the sample size. (In a sample of 10,000 where 100 4/4s are expected, 120 would be sufficient to indicate streakiness.)

The upshot is this: If we examine 160 batters (which we will) in nine different ways (which we will), we would expect to find (approximately) 36 cases where they *appear* to be streaky merely by chance (because 160 x 9 x .025 = 36). The reason we multiply by .025 instead of .05 is that if batting is truly random then approximately 5% of batters will *appear* to be NOT-random, BUT half of those will actually appear to be the opposite of streaky – i.e., they would be more likely to make an out after getting a hit and more likely to get a hit after making an out. Thus, only 2½% would have a performance consistent with streakiness. If there are more than 2½% – and the excess number is statistically significant – then we can conclude that we have some evidence that streakiness exists.

And Now…

…with *all* of the foregoing in mind, let us (finally) try to discover whether baseball players do or do not have hot streaks.

---

[35] *This* Matsui – as opposed to the *good* Matsui – Hideki, who plays for the hated Yankees – had 460 awficial at-bats in 114 games.

[36] In this example, it's a .316 batter who would be expected to get 4 hits 1% of the time.

Serial Correlation in Two Successive At-bats

The simplest possible test for streakiness (albeit surely somewhat complicated for people not very familiar with statistics) is to measure the correlation between (all) pairs of successive at-bats. If, for example, a .300 batter got three straight hits followed by seven straight outs, followed by three hits and seven more outs, etc., eight of every ten at-bats would produce the same result as the prior at-bat. To spell it out…

At-bats #s
1 & 2 two straight hits
2 & 3 two straight hits
3 & 4 one hit followed by one out
4 & 5 two straight outs
5 & 6 two straight outs
6 & 7 two straight outs
7 & 8 two straight outs
8 & 9 two straight outs
9 & 10 two straight outs
10 & 11 one out followed by one hit
Etc.

In a season of 550 at-bats, the "serial correlation" between his successive at-bats would be 0.53. However, for a batter with 550 at-bats to show pretty good evidence of streakiness we do not require a correlation nearly that high. In fact, a correlation a bit higher than .08 would suffice to be statistically significant (for streakiness). By the same token, a correlation a bit lower than minus .08 would indicate "anti-streakiness" – i.e., a tendency for that-batter to alternate between getting hits and making outs.

How did my universe of 160 batters do on this test in 2004?[37] As it turns out…

- Slightly more than half of the players (83 to be exact) had *negative* correlations – i.e., the "wrong" sign for those expecting to see a tendency towards streakiness;

- The mean correlation for the group was -.003 (i.e., zero for all practical purposes); and

- The median correlation for the group was -.002 (ditto).

But even if the group as a whole did not look streaky, perhaps some of the players did. (Many who believe streakiness exists think only *some* players are streaky; they may be distinguished from those who are "steady.") Recall that we expect about 2½% of all batters (i.e., four, if the universe is 160) to have positive correlations and a t-stat above 1.96 and another 2½% to have negative correlations and a t-stat above 1.96. About ½% each (1 or 2 batters overall) are expected to have t-stats above 2.58. The results shown

[37] Note the correlation required for statistical significance drops as a function of the square root of the number of at-bats.

below are for the six batters whose performance in 2004 was more than 1.96 standard deviations "off of random."

Table 3 – Players Whose Serial Correlations Were "Statistically Significant"

	Serial Correlation	At-Bats	T-stat	
Cintron, Alex	-.11	563	2.59	Anti-streaky (very)
Clayton, Royce	-.11	573	2.58	Anti-streaky (very)
Erstad, Darin	-.125	494	2.78	Anti-streaky (very)
Hunter, Torii	-.09	519	2.06	Anti-streaky
LoDuca, Paul	.09	534	2.15	Streaky
Uribe, Juan	.10	501	2.15	Streaky

As you can see, only two of the six players (LoDuca and Uribe) have correlations that are both positive and strong enough to make them appear to be "streaky." By pure chance we would have expected four such players. Meanwhile, four players had fewer stretches of multiple hits in a row than would be expected from a random process and, indeed, the first three listed had markedly fewer streaks. [The fact that they have the "wrong sign" – i.e., have the opposite correlation than we would expect if batters were streaky – is underscored by the use of shading. Whenever you see shading during the rest of this report, the data have this wrong sign.] The overall results here give no indication that streakiness exists – at least not for pairs of at-bats.

Of course, this was just the first of our statistical tests and along with results for pairs of at-bats where batters face the same pitchers (and which we report on in Appendix A – pp. 41-47), it seems pretty clear that even if players have days where they believe they "see the ball real well" or are "able to focus and stay relaxed" these mental states do not help them get two hits in a row more often than mere chance would suggests.

Of course, many (indeed, most) people who believe streakiness exists do NOT think it arises out of *pairs* of at-bats. Rather, they thing it takes a larger number of at-bats for players to get hot or fall into slumps. We move to those longer stretches of at-bats in the section after next but before moving on to these other tests, we have one more question related to the interpretation of statistical results that deserves to be addressed.

How Many High T-Stats Should One Expect to See From a Random Process?

The expectation that if batting is random that about 2½% of batters will appear to be streaky and another 2½% will appear to be the opposite of streaky are just "best guesses." If our sample were really, really large (say 160 batters for each of 100 seasons for a sample size of 16,000) then the results would probably be within 0.1% of 2.5%. But, since we will be seeking evidence of streakiness looking at results for "only" 160 players, we need to gain an appreciation of how much those 2.5%s might vary, merely by chance. To give the reader a "feel" for what this might look like, I ran 20 simulations, each using a sample size of 160 players and each with 547 at-bats (the average for my sample). I used a random number generator to simulate sequences of hits and outs. The question

was: "if we expect 4 out of 160 in each group to have a positive correlation and a t-stat of 1.96 or higher and 4 more to have a negative correlation and a t-stat of 1.96 or higher, how many will we actually get?" Put another way – if we see only 2 in one group and 7 in the other group, is that significant? The results for 20 simulations are shown below:

Table 4 – Simulated T-Stats that *Appear To Be* Statistically Significant

Simulation	Positive Correlation and T-stat Above 1.96	Negative Correlation and T-stat Above 1.96
1	6	1
2	2	8
3	5	1
4	5	3
5	4	4
6	5	9
7	5	4
8	7	3
9	3	3
10	6	2
11	3	4
12	6	2
13	2	6
14	7	3
15	4	7
16	5	3
17	3	6
18	3	4
19	9	5
20	2	3
Total	92	80

As it turns out, two simulations (6 and 19) had 14 (instead of the expected eight) "significant" t-stats while one had as few as five (simulation 20). The number of negative correlations with strong t-stats was *exactly* what one would expect (80 = 20 X 4) but the number of positives was 12 greater than expected (92 vs. 80). Nothing about these results are unusual for a random process. If the positive (or negative) correlations had totaled 98 or more, one might *begin* to wonder if the process was non-random.[38]

10 At-Bats: How Do the Number of Hits Vary?

How often will a batter go 0/10, 1/10, 2/10….9/10, 10/10? If we assume that players' underlying abilities are known (as given by their season-long batting averages) and that the batting process is random, we can calculate the correct expectation using the binomial expansion. Here are the results for selected batting averages, including the lowest (Jose Valentin) and the highest (Ichiro Suzuki) posted within our group of 160 everyday players.

[38] Of course, we would have *known* in this case, even if there had there been 100 in each column, that those results were still random because we simulated them using a random number generator.

Table 5 – Expected Distribution of Performance in Stretches of 10 At-Bats

	For a Batting Average of			
Hits/10 at-bats	.216	.250	.300	.372
0 for 10	8.8%	5.6%	2.8%	1.0%
1 for 10	24.2	18.8	12.1	5.7
2 for 10	30.0	28.2	23.3	15.1
3 for 10	22.0	25.0	26.7	23.8
4 for 10	10.6	14.6	20.0	24.7
5 for 10	3.5	5.8	10.3	17.5
6 for 10	0.8%	1.6%	3.7%	8.7%
7 for 10	0.1	0.3	0.9	2.9
8 for 10	0.01	0.04	0.1	0.7
9 for 10	0.0008	0.003	0.01	0.1
10 for 10	0.00002	0.0001	0.001	0.01

How do these expectations compare to actual results when we take into account each batter's batting average and exact number of at-bats? The table below focuses on performance that would be indicative of being hot (5/10 or better) or being in a slump (0/10). As may be seen, there were *fewer* extreme performances than one would expect merely by chance in six of the seven stretches of at-bats that we examined and in none of these seven cases were the deviations statistically significant.

Table 6 – Stretches of 10 At-Bats: Extreme Performances Versus Expectation

	0/10	5/10	6/10	7/10	8/10	9/10	10/10	Total 5/10 or better
Expected	3,232.6	7,613.2	2,614.0	626.2	100.3	9.7	0.43	10,963.8
Actual	3,147	7,607	2,585	604	107	9	0	10,912
Deviation from Expected	-85.6	-6.2	-29.0	-22.2	+6.7	-0.7	-0.43	-51.8
Standard Deviation[39]	176.4	263.4	159.2	78.8	31.6	9.8	2.1	97.8
T-stat	0.49	0.02	0.18	0.28	0.21	0.07	0.21	0.53

10 At-Bats: What's Happens in the 11th At-Bat After a Good or Bad Stretch?

Perhaps stretches of 10 at-bats are too short to *reflect* streakiness but are not too short to *create* streakiness. In other words a batter who is 0/5 or 3/5 may not be hot or cold in a meaningful psychological sense but after going 0/10 or 6/10 he will be. If so, the 11th at-bat following a particularly poor or strong performance will show that such batters are, on balance, "slumping" or "in a groove." Indeed this idea was behind the poll questions I discussed on pp.13-15. What do we find in these (collective) 11th at-bats?

[39] See Appendix B (p. 47) regarding how to estimate standard deviations for samples with overlapping data.

Table 7 – Performance in At-Bat #11 After Extreme Performance During Prior 10

	0/10	5/10	6/10	7/10	8/10	9/10	Total 5/10 or better
Number of Stretches[40]	3,137	7,592	2,579	604	107	9	10,891
Expected Batting Avg.	.281	.287	.289	.291	.299	.291	.288
Actual Batting Avg.	.275	.295	.282	.310	.280	.222	.292
Expected Hits	881.3	2,180.9	745.3	176.0	32.0	2.6	3,136.7
Actual Hits	862	2,238	727	187	30	2	3,184
Deviation from Expected	-19.3	+57.1	-18.3	+11.0	-2.0	-0.6	+47.3
Standard Deviation	25.2	39.4	23.0	11.2	4.7	1.4	47.3
T-stat	0.77	1.45	0.79	0.99	0.43	0.45	1.00

Results are mixed inasmuch as sometimes there are more hits than chance would suggest and sometimes there are fewer. But again no data are statistically significant – i.e., there is still no evidence that streakiness exists. Some examples:

- After going 0/10, players with a collective batting average of .281 did bat a little lower than one would expect (.275) – i.e., in the correct slump-like direction – but the result was less than one standard deviation below expectations.

- Among batters who had good performance over stretches of 10 at-bats (5/10, 6/10, etc.) there were also extra hits (collectively a batting average of .292 vs. an expectation of .288) but, once again, the result was not statistically significant.

- Three cases had the "wrong" sign (6/10, 8/10 and 9/10) and the situation that was closest to indicating streaks exist (11th at-bats following 5/10s) did not attain the required t-statistic of 1.96 or higher.

- Specifically with regard to the question that we had asked in our poll which had to do with performance in 11th at-bats following stretches of 5/10, the average batting average was .295 vs. (a) a random expectation of .287 and (b) an expectation implicit in the responses to our poll that such batters would hit something between .312 and .326. [When we asked what a .300 batter would bat after going 5/10, the mean response among those who believe streaks exist was .339 and the median was .325. One answer was .039 higher than the batter's underlying average; the other was .025 higher.]

[40] If you are wondering why the numbers in this row are smaller (in some cases) than the numbers in the second row of Table 6, the answer is that occasionally batters were 0/10 or 5/10 or 6/10 etc. in their *final* 10 at-bats of the season. Whenever that happened, there *was no 11th at-bat.*

Dear Sports Fan: You know the feeling you get when a batter comes to the plate who is on a "tear" – and you feel SURE he's gonna get a hit? Or the guy who has been flailing at pitches for three games and you're SURE he's a basket case? Well, this data suggest the batting average in the 11th at-bat will be only about .004-.006 in the direction you would expect. *Indeed, it's only .004-.006 in the direction you **should** expect – namely the average he had for the entire season **not counting** the streak itself.* The fact is (take a look once again at Tables 2A and 2B on pp. 24-25) that streaks distort player batting averages and most fans' expectations are probably influenced by that. A .300 "underlying" batter whose average is boosted to .315 by some hot streak, is probably expected to hit *better* than .315 in his 11th at-bats. Our data suggest he might hit better than the .300 he was batting *before* the hot streak, but they also suggest he will hit lower than .315 (current) batting average.

Of course it is possible that 10 at-bats are insufficient for batters to get "in the zone" or fall into a slump. Maybe these will become noticeable if we examine stretches of 25 at-bats (and the 26th following good or bad stretches). The sections below examine these questions.

25 At-Bats: How Do the Number of Hits Vary?

How often will a batter go 0/25…3/25, or 11/25 or better? Table 8 is the 25-at-bat-equivalent of Table 5 (leaving out the performances that are not extreme):

Table 8 – Expected Distribution of Performance in Stretches of 10 At-Bats

	For a Batting Average of			
Hits/25 at-bats	.216	.250	.300	.372
0 for 25	0.2%	0.1%	0.01%	0.001%
1 for 25	1.6	0.6	0.1	0.01
2 for 25	5.2	2.5	0.7	0.1
3 for 25	11.0	6.4	2.4	0.4
11 for 25	0.7%	1.9%	5.4%	12.5%
12 for 25	0.2	0.7	2.7	8.6
13 for 25	0.1	0.2	1.1	5.1
14 for 25	0.01	0.1	0.4	2.6
15 for 25	0.003	0.02	0.1	1.1
16 for 25	0.0005	0.004	0.04	0.4
17 for 25	0.0001	0.0006	0.008	0.1
18 or more	0.00001	0.0001	0.002	0.04

How do these expectations compare to actual results when we take into account each batter's batting average and exact number of at-bats? To help preserve legibility, we will subdivide the results into two tables – the first covers superior batting stretches and the latter covers poor batting stretches. The results shown in Table 9A and 9B are a bit stronger than the 10-at-bat-equivalent that we showed in Table 6, but just like in Table 6, almost all of the results have the "wrong sign." Our 160 batters had 83,641 collective

opportunities to get 11 hits (or more) in stretches of 25 at-bats during the 2004 season. Given their respective batting averages, "mere chance" would have suggested they would have accomplished this feat 6,472 times. If batters were genuinely streaky, one would have expected many more than 6,472 such stretches; instead, there were 283 fewer.

Table 9A – Stretches of 25 At-Bats: Superior Performances Versus Expectation

	11/25	12/25	13/25	14/25	15/25	16/25	17/25	18+	Total 11/25 or better
Expected	3,553.6	1,752.9	754.3	283.4	92.8	26.4	6.5	1.7	6,471.6
Actual	3,465	1,726	622	260	85	29	2	0	6,189
Deviation from Expected	-88.6	-26.9	-132.3	-23.4	-7.8	+2.6	-4.5	-1.7	-282.6
Standard Deviation[41]	291.7	207.1	136.7	84.0	48.1	25.7	12.8	6.4	327.1
T-stat	0.30	0.13	0.97	0.28	0.16	0.10	0.35	0.26	0.86

Given the foregoing, we should not be surprised when Table 9B also shows fewer-than-expected extremely poor stretches. The reason this is not a surprise can be illustrated with a somewhat oversimplified example. Consider someone who bats .250 for three games. He might go 1 for 4 in each game – i.e., no superior games and no poor games. However, if he's 3/12 overall and has one game where he gets two hits, then one of the other games must be a poor game (i.e., 0/4). The arithmetic over a full season is not this ironclad, but with the results shown in Table 9A it was extremely likely that Table 9B would show the complementary result – namely, fewer poor stretches than mere chance would suggest.

In any event, the key point is that we have yet to see any data that is statistically significant on either side of the argument: either supporting streakiness or its opposite.

Table 9B – Stretches of 25 At-Bats: Poor Performances Versus Expectation

	0/25	1/25	2/25	3/25	Total 3/25 or worse
Expected	28.1	252.8	1,106.5	3,137.1	4,524.5
Actual	44	246	1,005	2,936	4,231
Deviation from Expected	15.9	-6.8	-101.5	-201.1	-293.5
Standard Deviation[42]	26.5	79.4	165.2	274.7	327.1
T-stat	0.60	0.09	0.61	0.73	0.90

[41] See Appendix B (p. 47) regarding how to estimate standard deviations for samples with overlapping data.
[42] Ditto.

25 At-Bats: What Happens in the 26th At-Bat After a Good or Bad Stretch?

Tables 10A and 10B are the 26th-at-bat-equivalent(s) of Table 7. As you can see in 10A, there is absolutely nothing (of note) going on after batters do well for stretches of 25 at-bats. During the 26th at-bats they average .293 instead of an expected .292.

Table 10A – Performance in At-Bat #26 After Superior Stretches of 25 At-Bats

	11/25	12/25	13/25	14/25	15/25	16/25	17/25	Total 11/25 or better
Number of Stretches	3,459	1,722	622	259	85	29	2	6,178
Expected Batting Avg.	.290	.291	.297	.300	.301	.303	.325	.292
Actual Batting Avg.	.291	.292	.302	.324	.259	.276	.000	.293
Expected Hits	1,004.5	501.3	184.5	77.8	25.6	8.8	0.7	1,803.2
Actual Hits	1,008	503	188	84	22	8	0	1,813
Deviation from Expected	+3.5	+1.7	+3.5	+6.2	-3.6	-0.8	-0.7	+9.8
Standard Deviation	26.7	18.9	11.4	7.4	4.2	2.5	0.7	35.8
T-stat	0.13	0.09	0.31	0.84	0.85	0.32	0.98	0.28

Moving to 26th at-bats following poor stretches, we find that-batters hit just .266 vs. an expected .278. However, the t-stat of 1.73 suggests this is not (quite) enough of a deviation to be statistically convincing that this reflects anything but mere chance. Still, for what it's worth, this is the closest we have come to finding evidence of streakiness. Tables 10A and 10B together suggest hot streaks don't exist but slumps *might.*

Table 10B – Performance in At-Bat #26 After Poor Stretches of 25 At-Bats

	0/25	1/25	2/25	3/25	Total 3/25 or worse
Number of Stretches	44	243	1,002	2,929	4,218
Expected Batting Avg.	.272	.277	.279	.278	.278
Actual Batting Avg.	.182	.251	.255	.273	.266
Expected Hits	12.0	67.4	279.4	815.7	1,174.5
Actual Hits	8	61	256	799	1,124
Deviation from Expected	-4.0	-6.4	-23.4	-16.7	-50.5
Standard Deviation	3.0	7.0	14.2	24.3	29.2
T-stat	1.34	0.91	1.65	0.69	1.73

Taking Stock

So far, our results, by being mixed and, on balance, a bit on the "anti-streaky" side, strongly (if oxymoronically) suggest that randomness is the order of the day. For our group of 160 we have found the following results with the "wrong" sign:

1. A (tiny) negative serial correlation from one at-bat to the next (-.003 on average);

2. Fewer 0/10s than one would expect by chance (t-stat 0.49);

3. (Slightly) fewer 5/10s or better than one would expect (t-stat 0.17);

4. Fewer stretches of 3/25 or worse than one would expect (t-stat 0.90):

5. Fewer stretches of 11/25 or better than one would expect (t-stat 0.86);

On the other side of the coin, we did find some results indicating a some tendency toward streakiness but again none of them rose to the level of statistical significance:

6. A batting average of .292 vs. an expected .288 in the 11th at-bat following 5/10s or stretches better than 5/10 (t-stat 1.00);

7. A batting average of .275 vs. an expected .281 in the 11th at-bat following 0/10s (t-stat 0.77);

8. A marginally better batting average of .293 vs. an expected .292 in the 26th at-bat following 11/25s or stretches better than 11/25 (t-stat 0.28);

9. A batting average of .266 vs. an expected .278 in the 26th at-bat following 3/25s or stretches worse than 3/25 (t-stat 1.73);

Considering that in a random situation you would expect about half the data to look a bit streaky, about half to look "anti-streaky" and none (or hardly any) to reach the level of statistical significance, the data from 2004 offers no support for the idea that baseball players (taken collectively) are streaky.

Multi-Year Efforts To Uncover Streakiness

In a final attempt to uncover evidence of streakiness I decided to extend my analysis to multiple years in the two areas that "looked streakiest" in 2004 – namely (a) prolonged slumps and (b) those individual players who were the streakiest in 2004.[43]

[43] I thought this would have a better chance of uncovering streakiness than looking at all starting players for additional years. Note that since statistical reliability rises only with the square root of sample size, I would have had to do three more years to double the "information" on streakiness for all players. (Analyzing 160 players for a single year took me more than 100 hours.)

Maybe Hot Streaks Don't Exist But Slumps Do

I would certainly have expected hot streaks and cold streaks to be two sides of the same coin – i.e., if streakiness exists, then *both* hot and cold streaks might be expected to occur more often than mere chance would suggest and if streakiness does not exist, then neither should be expected to occur more often. If, for example, slumps exists (beyond chance) then we know (with certainty) that the collective at-bats outside of slumps, not only evince higher than normal batting averages, but that the batting average is boosted further by the fact that there are "extra" slumps (i.e., more than chance would produce). On the other hand, it is conceivable that the distribution is skewed in such a way that the number of very hot stretches simply do not mirror the number of very cold ones. Indeed, points 8 and 9 on the previous page (which are drawn from Tables 10A and 10B on page 34) suggest this might be true. The T-statistic for how batters performed in their 26th at-bat following stretches where that went 11/25 or better is close to zero whereas 26th at-bats following stretches in which batters got three hits or fewer is not all that far from being statistically significant. If the results for 2004 were repeated year after year it would become statistically significant (for slumps).

To test this hypothesis I obtained data on all "slumps" for eight separate seasons – 1996-1998, 2000-2003 and 2005.[44] I looked at every batter who had a stretch of 3/25 or worse, examined whether he got a hit in the 26th at-bat and compared the results to what we would have expected given each player's season-long batting average not counting those 26 at-bats. This is precisely the same kind of test that we reported on for 2004 in Table 10B. Here are the results for all players with 450 or more at-bats in those eight seasons. As you can see, these batters were *not* slumping in their 26th at-bats.

Table 11 – Performance in At-bat #26 After Poor Stretches of 25 At-Bats
(1996-1998, 2000-2003, 2005)

	0/25	1/25	2/25	3/25	Total 3/25 or worse
Number of Stretches	193	1,663	7,796	23,277	32,929
Expected Batting Avg.	.272	.274	.277	.278	.277
Actual Batting Avg.	.275	.272	.274	.277	.276
Expected Hits	52.5	455.9	2,161.4	6,463.2	9,133.1
Actual Hits	53	452	2,134	6,446	9,085
Deviation from Expected	+0.5	+3.9	-27.4	-17.2	-48.1
Standard Deviation	6.2	18.2	39.5	68.3	81.2
T-stat	0.08	0.21	0.69	0.25	0.59

[44] At the time I requested the data, 2006 was not yet available and my supplier (yes, "supplier" is the right word to use here; baseball *is* addictive) had never been able to obtain the data for 1999.

Since I had the data in convenient form, I also examined players with fewer than 450 at-bats (i.e., all players, including all part-timers and pitchers – a poor-hitting group) and the results were essentially the same. In the aftermath of these very bad stretches they batted a collective .233, just .003 lower than the .236 average they posted in all other at-bats.

Earlier I pointed out that if you do enough statistical tests on random data, some percentage will inevitably appear to be non-random. Indeed, 5% of the results should look non-random at the 95% confidence level. What I've done in this section was take the data from 2004 that happened to be most consistent with the notion that streaks *might* exist (if only on the slump-y side), and tested a batch of data that was eight times as large. The result: *bubkes*.

One last point on this subject: Both for 2004, when the collective batting average of slumpers was .012 lower than it "should have been" (.266 vs .278) and the additional eight years when the excess was much less, but it was still a shortfall (.276 vs .277) someone *desperate* to believe in the existence of streaks might say that "well, at least they hit lower than they 'should' after going 3/25 or worse. Maybe, if we tested 50 years of data, we'd find statistically significant evidence that slumps exist." Perhaps. But, a batting average differential of .012 (let alone .001) is nothing like what baseball fans have in mind when they talk of hot streaks or slumps *and because we have excluded the poor stretches of 25 at-bats, most players turn out to hit higher in those 26th at-bats than the batting average they sport when they come to the plate.* (See Table 2B on page 25 if you've forgotten why this is true.)

<u>But (Finally) Maybe *Some* Baseball Players Have Hot Streaks</u>

If ballplayers collectively do not evince streakiness, perhaps that is because the group data represent a mix of two opposite kinds of players: (1) some who are very steady and (2) others who are quite streaky. But looking at individual players poses some thorny statistical problems. (For those of you interested in this discussion see Appendix B, beginning on page 47.)

As it turns out, we did not find any players who looked convincingly streaky in 2004. But – as might be expected from a random process, about half looked less streaky than chance would suggest and the other half looked streakier. The latter group ranged from those who were only marginally streakier than random, to those who looked streaky enough that if those result were reproduced for multiple seasons we would have pretty convincing evidence that *some* players *are* streaky.

I selected 18 players (of my group of 160), nearly one-fourth of those players on the "streaky side" and obtained their batting data for all years (other than 1999) between 1996 and 2005. I then established some criteria they needed to meet before I would include them in the additional study:

- I would not include any player without at least one other year of 450 at-bats (this eliminated two players – one who never had another season with more than 368

at-bats and another who never had more than 298;

- I would only include seasons with at least 300 at-bats.

The latter rule had the effect (I believe) of coming very close to the ideal of including all the years in which these players were everyday players. Many players have years early in their careers where they are being platooned (or even just pinch-hitting and subbing for injured players). All such years are excluded. Often a player becomes a "starter" mid-way through his first successful season. When you see a player with 85 at-bats one year, 380 the next and 540 the next, it is often true that he became a starter part-way through the middle season. (Admittedly, sometimes he's being platooned – but even so, he's playing about 2/3s of the time and I decided to include those seasons.) Finally if you see a succession of seasons, with, say, 550, 499, 325, and 522 at-bats, the depressed year almost always reflects a severe injury. In other words the player *was* an everyday player, *except* during the period he was out with the injury.

In the end, I examined the performance of 16 players over 68 seasons (36,261 at-bats). Since I know you're dying to know who the streakiest players were during 2004, here's the list, along with the number of seasons *other* than 2004 that I studied:

Angel Berroa	2 Seasons
Bret Boone	8
Jeromy (*sic*) Burnitz	7
Orlando Hudson	2
Raul Ibañez	3
Derek Jeter	8
Mark Kotsay	6
Paul Lo Duca	4
Victor Martinez	1
Hideki Matsui	2
Lyle Overbay	1
Jay Payton	5
Juan Pierre	4
Joe Randa	8
Juan Uribe	3
Jason Varitek	4

I studied the players individually (by season), in aggregate over each of their careers, and collectively (all 68 seasons put together). After all, if a bunch of players are a *little* streaky and I add up all of their seasons, a very large sample size might yield statistical significance. I looked for hot and cold stretches within groups of 10 at-bats and 25 at-bats as well as how they performed in the 11^{th} or 26^{th} at-bats following such stretches. What did I find? From a *group* standpoint, I found more *bubkes*. To summarize the main findings:

- This group (of 16 players over 68 seasons) should have (by mere chance) gone 0/10 a total of 1,389 times but did so only 1,329 times. (That's the "wrong sign" for those who believe in streaks);

- The group should have gone 5/10 or better 4,438 times but did so just 4,284 (again the wrong sign);

- They should have gone 3/25 or worse 1,993 times and did so 2,010 times (which was more than chance would predict but insignificantly more);

- They should have gone 11/25 or better 2,577 times but did it just 2,380 (wrong sign);

- After going 0/10 these guys were expected to bat .2755 but-batted .2757 (wrong sign by a tiny amount, i.e., they were supposed to get 364.7 hits but actually got 365);

- After going 5/10 or better, they were supposed to bat .285 but-batted .283 (wrong sign);

- After going 3/25 or worse, they were supposed to bat .273 in their 26th at-bats but-batted .263 instead (this is the largest margin on the "streaky" side thus far but the t-statistic is only 0.99 – i.e., it's not close to significant);

- After going 11/25 or better, they were supposed to bat .286 in their 26th at-bats but batted .292 instead (another "correct" sign but a t-statistic of only .56).

I could add that the serial correlation between successive at-bats for the group was zero to four decimal places and more players were negative than positive. The upshot is that if you take 16 players who look pretty streaky over one particular season and examine their entire careers, you are very unlikely to find that the are streaky (as a group).

BUT, maybe it's not the group that is streaky. Maybe it's just a couple of the players! Do any of them fill the bill?

With regard to the number of 0/10s, 5/10 (or better), 3/25 (or worse), and 11/25 (or better), by far the streakiest player (over the period studied) was Juan Uribe. In the three seasons where he had enough at-bats to qualify, he was on the streaky side with regard to all four of those measures.[45] (Although he was slightly on the UNstreaky side in 2003, he was quite streaky 2002 and 2005 – to go with the streakiness demonstrated in 2004.) None of the t-statistics reached the threshold of significance but all were above 1.00 (they ranged from 1.07 to 1.50). Maybe he's the one (the Dalai Lama of streakiness?). Indeed,

[45] The one area where he looked the opposite of streaky over those three years (2002, 2003 and 2005 combined) was serial correlation. Over the full five years where he has been an everyday player (including all years from 2002-2006), his serial correlation is positive (.024), but not significantly so (t-stat 1.14).

when I examined his performance in the 11th and 26th at-bats he was again on the streaky side over the three-year period. He got more hits after going 5/10 or better or 11/25 or better than his un-streak-related at-bats would have suggested and he got fewer hits in his 0/10s or stretches that were 3/25 or worse than randomness would say is likely. Two of the t-stats were in the 1.1-1.3 range, one was exactly 1.96 (right at that 5% confidence interval) and the last one (performance in 11th at-bats after going 0/10) had a t-stat of 2.08.

Of course, if you look at 160 players, some are bound to look streaky, just by chance. And, if you pick 16 of the streakiest for a single season and look at their careers the chances are decent (albeit less than 50%) that one of them will look streaky. That Juan Uribe is just short of streaky for the four seasons ended 2005, might mean he's genuinely streaky – and we can follow the rest of his career to see how he pans out.

However, I reviewed his performance for 2006 and I'm afraid it does not augur well for believers in streakiness. Indeed, while Uribe showed up as being a little streaky on successive at-bats (serial correlation), in every one of the eight other measures I examined (the number of 0/10s, 5/10s or better, 3/25s or worse, and 11/25 or better and also how he did in the 11th and 26th at-bats following such stretches), Uribe was less streaky in 2006 than mere chance would have predicted. Finally, as a result of the inclusion of the 2006 season in his *career* statistics, Uribe is no longer close to statistically significant on any of these measures.

Do Baseball Players Get Hot (or Cold)?

No – not even Juan Uribe.

Appendix A
Estimating the Impact of Fundamental Factors on Streakiness

Recall that there were seven fundamental factors cited on pages 17-18 including (1) home-away effects, (2) ballpark effects, (3) opposing pitcher variability, (4) opposing pitcher declining effectiveness, (5) variability in defensive skills, (6) injury, (7) stress unrelated to baseball. The final three factors are impossible (for us) to measure, but the first four are quite tractable. For these purposes I am going to rely exclusively on pairs of at-bats because they are the ones that are easiest to model and to explain.

Home-Away and Ballpark Effects

An approximation of the *combined* effects for these two factors may be calculated using a few reasonable assumptions plus the data in the table below which report home-vs.-away batting averages for each of the 30 major league teams for 2004.[46] Below you can see the 18 teams that had higher batting averages in their own parks; on page 42, you can see the 12 teams that hit better on the road. As may be seen, only eight of the 30 teams posted a Home-away difference greater than .016. The mean absolute difference (ignoring

	Home Batting Average	Away Batting Average	Home Advantage/ (Disadvantage)
Colorado Rockies	.303	.246	.057
Boston Red Sox	.304	.260	.044
Texas Rangers	.285	.246	.039
San Francisco Giants	.284	.257	.027
Arizona Diamondbacks	.266	.240	.026
Houston Astros	.277	.257	.020
Chicago White Sox	.276	.260	.016
Chicago Cubs	.274	.262	.012
St. Louis Cardinals	.284	.272	.012
New York Mets	.254	.244	.010
Toronto Blue Jays	.264	.256	.008
Minnesota Twins	.268	.263	.005
Pittsburgh Pirates	.263	.258	.005
Oakland Athletics	.272	.268	.004
New York Yankees	.270	.267	.003
Milwaukee Brewers	.249	.247	.002
Detroit Tigers	.273	.272	.001

[46] The main assumption we use has to do with the variability of ballpark effects that each team encounters on the road. We have relied on the standard deviation of home ballpark batting average (for each league) to estimate this effect. For a more accurate study, one would want to break down "away batting average" into averages for each individual ballpark visited, adjusted correctly for the exact number of games played in each park. That would require collecting much more data for only a slight improvement in accuracy. I should add that all statistics estimated in this section take account of the exact number of home and away at-bats that each team has – i.e., it does *not* assume that 50% of at-bats occur at home and 50% on the road.

Kansas City Royals	.259	.258	.001

	Home Batting Average	Away Batting Average	Home Advantage/ (Disadvantage)
Anaheim Angels	.282	.283	(.001)
Philadelphia Phillies	.266	.268	(.002)
Baltimore Orioles	.280	.283	(.003)
Florida Marlins	.262	.265	(.003)
Los Angeles Dodgers	.260	.263	(.003)
Montreal Expos	.246	.251	(.005)
Tampa Bay Devil Rays	.254	.262	(.008)
Atlanta Braves	.265	.274	(.009)
Cleveland Indians	.270	.281	(.011)
Cincinnati Reds	.242	.258	(.016)
Seattle Mariners	.255	.284	(.029)
San Diego Padres	.256	.288	(.032)
Average for 30 Teams	.269	.263	.006

whether home was greater than away or vice versa) was .014 and the median absolute difference was.009. Both figures, however, are greater than the .006 home-away difference for the major leagues as a whole and, thus, it is obvious that ballpark effects are quite a bit more important than the home-vs.-away advantage.[47] How much can this variability in batting averages be expected to contribute to the frequency with which batters get two hits in a row?

Let's first consider the arithmetic of zero variability. If all batters came to the plate 100% of the time as .266 hitters (which is the average for all major leaguers), they would get two straight hits 7.08% of the time (.0266 x 0.266 = .0708). If, instead, they came to the plate as .269 batters half of the time (the "home" batting average for the majors) and as .263 batters half of the time (the "away" average) the resulting two-for-twos would be the very same 7.08% (.0269 x .0269 plus .0263 x .0263 divided by two still equals .0708). If we increase the number of decimal places, however, the two probabilities will not be identical. A constant .266 batting average produces streaks of two 7.0756% of the time while .269 at home and .263 on the road produces 7.0765%. So, there is a difference, but it is worth only than 1½ two-for-twos for the entire major leagues over a full season.[48]

[47] Nevertheless, the .006 home-away difference is statistically significant owing to the huge number of at-bats in a major league season.

[48] Total at-bats in 2004 were 167,353. The number of players with at least one at-bat was 959 while the number of players with at least two at-bats was 880. The 79 players (almost all pitchers) who had just one at-bat could not possibly get two straight hits while the other 880 each had one fewer chance than their number or total at-bats. The upshot is that the opportunities to get two straight hits were 167,353 minus 79 minus 880 or 166,394. If we multiply 166,394 by .070756 we get 11,773.4 expected two-for-twos. If we multiply 166,394 by .070765 we get 11,774.9 expected two-for-twos – a difference of 1.5. (Statistic hounds will recognize that even this estimate is not quite right because of the transitions that happen every time a home-stand ends or a road-trip venue shifts from one ballpark to another. But the impact of this factor – one which *reduces* the number of streaks is very small. I estimate that it offsets about one-seventh

If 1½ extra two-for-twos seems trivial, the reader is correct. For streakiness to be statistically significant we would need to find at least 205 extra two-for-twos. Assuming 7.0756% two-for-twos is the right expectation, we would have expected 11,773 during 2004 with a standard deviation of 105. This means that our best guess was 11,773 but that the fluctuations that are *merely due to chance* were expected to boost OR reduce the number of two for twos *outside* the 11,668-11,878 range almost one-third of the time. Multiply the 105 standard deviation by 1.96 (the number of standard deviations required for *statistical significance*) and we would want to see an extra 205 two for twos or more than 100 times the extra two for twos that result from the home-away difference of .269 vs. .263 before we were willing to conclude that home-away differences should clearly be expected to contribute to streakiness.

But what if we take account of the full variety of home-away differences and ballpark effects that was shown in the tables on pp. 41-42? Earlier I did not mention that the impact of these differences tend to mushroom as a function of the *square* of the proportion of the differences. Thus, if Colorado bats .057 higher at home than on the road – or about 10 times the .006 difference for the major leagues as a whole – then the impact on that team's streakiness will be about 100 times as great (!) as one might surmise from the aggregate data. Note, that it doesn't matter whether a team hits higher at home or on the road; either will add to their streakiness. (Because the affect is a function of the square of the home-road differential, just four teams account for about two-thirds of the all ballpark effects in the major leagues. Colorado alone accounts for 28%, while Boston contributes 17%, Texas 13% and San Diego 9%. However, when we aggregate the results for all the teams, taking account of (a) their particular home and away batting averages and also (b) (an estimate of) how their away batting averages will vary because of ballpark effects in different cities my best estimate is that teams should expect about 7.10% two for twos versus the 7.08% that would expected if all teams batted .266 regardless of the ballpark they played in.[49] Worked out to all the decimal places at our disposal, the number of extra two for twos is 28 – i.e., 11,801 instead of 11,773. The difference is just over one-quarter of one standard deviation and not even close to statistically significant.

Note that a .266 league-wide average both at home and away might indicate (to take two extremes) that every team hits .266 everywhere or, say, half of the teams average .296 on the road and .236 at home and the other half do the opposite. If all teams always batted .266, they would get two hits in a row 7.08% of the time. If half hit .296 one place and .236 in the other (whether home or on the road is irrelevant) they would get two hits in a row 7.17% of the time. The difference between those two percentages would be worth 150 two for twos or almost three-quarters of what is needed reach the threshold of statistical significance. But, except, in Colorado and, to a lesser extent, a few other

---

of the streak boosting factor here. Put another way, we do not really expect 1.5 extra two-for-twos because of these factors in a typical major league season; instead we expect about 1.3 more streaks of two-for-two.

[49] The astute reader may realize that the fact that the various players on each team have different batting averages from their teammates has an impact on the expected number of two for twos. But assuming that home-away and ballpark effects affect all players similarly, the impact of this would be negligible.

places, the home-away and ballpark effects are not close to this size. The upshot: This factor will NOT affect the statistics on streakiness to any material degree.

Pitcher Variability and Pitcher Fatigue (Declining Effectiveness)

On page 20 I noted that the fact that some pitchers are above average while others are below average "is surely the most important factor *reducing* the amount of streakiness we see." While that is true for stretches of at-bats (say four or more) where one faces more than one pitcher, it will *increase* streakiness when measured for really, really short stretches – namely stretches of two at-bats. Meanwhile, pitcher fatigue should always reduce streakiness.

I thought the best way to begin to analyze how these factors interact would be to focus on data related to the top five starting pitchers for each team. These starters collectively represented fewer than 24% of all pitchers who appeared in games in 2004 but they nevertheless started 82% of all games.[50] In the games that they started, our 150 top starters pitched approximately 24,220 innings (56% of the innings pitched by all pitchers in 2004), representing an average of slightly more than 6 innings per game and accounting for about 2/3 of all batters faced in those games.[51]

Although, starting pitchers are somewhat less variable in their performance than relievers, the worst relievers (who account for the excess variability of all relievers) pitch very few innings. The upshot is that I figured that a simulation based on the performance of these principal starters would give us a "decent" portrayal of how pitcher variability might affect streakiness. Later, when trying to measure the effects of pitcher fatigue on streakiness, we will again use starting pitchers (because relievers rarely face batters more than once per game) but we will do so by matching all starters (not just the top 150) against the 160 batters whom we analyzed for streakiness.

In reviewing these data, we have two questions we would like to answer:

---

[50] You might think that five pitchers on each of 30 teams would total 150… but, heh, heh…you would be wrong because Freddy Garcia started 16 games for the Chicago White Sox and 15 games for Seattle, making the top five list for both teams. I will refer to these 149 pitchers as "150 starters" throughout. These 150 starters, started 3994 games or an average of 26.6 apiece and 133 per team. All teams use 5-man rotations but there are very few teams where five starters each start about one-fifth of the games for the entire season. Injuries, "demotion" of ineffective pitchers (either to the minors, or by sending them to the bullpen), or simply passing over the least effective starter when there are days off in the schedule are very common. The average distribution of starts in 2004 for starters 1-5 were 33.3, 30.9, 27.4, 23.7 and 18.3 respectively. Every team had at least one starter with 30 or more starts and 23 of 30 teams had two such starters. "Genuine" 5-man rotations – i.e., where the 5th starter had 24 or more starts were uncommon – there were only five such teams. Twenty of the 30 "fifth starters" started between 14 and 19 games while there was one each starting 10, 13, 20 and 21.)

[51] I know that these starters pitched 24,524 innings out of a total of 43,257 pitched in the majors. I also know that they (collectively) appeared in relief 200 times. Assuming each relief outing was one-fourth as long as each start (something I'm guessing), the relief appearances averaged 1.515 innings and each start averaged 6.06 innings. Meanwhile, the average number of innings per game is 8.93 because the extra innings that result from extra-inning games are slightly fewer than the "missing" innings that result from the fact that the bottom of the ninth is not played/finished when the home team is ahead and also as a result of innings lost to rain-shortened games.

1. What is the effect on batting streakiness when batters face one starting pitcher one day and then another starting pitcher (of *different* skill) the next day? And…

2. What is the effect on batting streakiness as batters face the *same* pitcher more than once in the same game?

Somewhat surprisingly (to me) our 150 starters had a "batting average against" (BAA) of .266 which is the same as the aggregate batting average for the major leagues in 2004.[52] Question: How many two-for-twos would this group of pitchers have yielded if *every single batter* batted .266 against *every single pitcher* AND how would that compare to the number of two-for-twos these same pitchers would have would have been expected to allow given their actual BAAs?

Note, that some of these 150 pitchers are superb while others are mediocre (to say the least). The estimates I derived in the simulation took into consideration not only BAA for each pitcher, but also their (a) walks per inning and (b) innings pitched per game started (this number is estimated but the margin of error should be quite small – see note 49 at the bottom of p. 44). I also made allowances for hit-batsmen, sacrifice hits, safe-on-error, double plays, caught stealing, outfield assists and unforced infield assists that were based on averages for the major leagues.[53]

In my estimation, there were nearly 104,000 plate appearances against the 150 starters and I will guesstimate that there were approximately 57,000 occasions where batters had a chance to get two straight hits in two official at-bats in the same game.[54]

If every pitcher had a .266 BAA, then one would have expected 4,033 two-for-twos by opposing batters (.266 x .266 x 57,000 = 4,033).[55] However, given the heterogeneity of BAAs and the other factors mentioned above, the number of expected two-for-twos is higher – *but not that much higher* – only 4,070 – a difference of 0.9%. How can this be, given that some pitchers are so effective while others are so ineffective? [56]

---

[52] I would have thought it would be lower but it isn't – probably because fifth starters (and, even, fourth starters) are not very good and starters face more batters when they are tired than relievers do. Also, relievers get to face same-handed batters a higher percentage of the time than starters do.

[53] The impact of these factors on the estimate of the number of batters each pitcher faced two or more times was quite small and researching the exact data here would have been an immense job. In the case of hit batsmen, I used an algorithm that took account of the number of walks/inning for each pitcher.

[54] Note that this need not take place in two straight plate appearances. One might face a pitcher three times in a game. When that occurs and you walk on your second plate appearance, you have one chance for a "two-for-two" – i.e., you can get a hit during both your first and third plate appearances.

[55] Actually, this also assumes every batter hits .266 against every pitcher. To the extent that players' batting averages vary, the number of two-for twos will be higher, but that applies to ALL of these estimates.

[56] Johan Santana of the Minnesota Twins had the lowest BAA in the major leagues in 2004, holding opposing batters to a .192 average while his teammate – Terry Mulholland – allowed the major-league high (among "top" starters) of .327.

The reason the difference is so small is that fully half of all pitchers were in the relatively narrow range for BAA of .252 to .284 and relatively few were at the extremes (only about 8% each were below .229 or above .302). The difference between .252 and .284 is more than enough to make the difference between a winning pitcher and a losing pitcher, but it will increase the number of 2/2s in a season by less than 0.4%![57]

Recalling that these 150 pitchers started "only" 82% of all games in 2004 and allowing for greater variability when we bring the "lesser" pitchers into the picture, one might expect 1.0% or, conceivably, 1.1% extra two-for-twos given the variability in skills of starting pitchers, which would mean that there would be about 4,075-ish (not 4,070) two-for-twos instead of 4,033. Unfortunately, to meet the standard statistical significance hurdle, one would need approximately 4,153 two-for-twos.

Meanwhile, pitcher fatigue cuts the other way – i.e., it reduces the number of two-for-twos one can expect to see. To estimate this effect I did something very simple: I looked at every case where the 160 batters examined for streakiness had two official at-bats against the opposing starter.[58] The sample size for this was 20,465 (for a total of 40,930 at-bats). Overall, our 160 batters hit .2836 against these pitchers, but in their first at-bats they hit .2733 while in their second at-bats, they averaged .2939. This difference is highly statistically significant – more than 3.2 standard deviations – and demonstrates clearly that pitchers become less effective every successive time they face an opposing batter. How much can this be expected to reduce the number of two-for-twos? The answer is scarcely anything at all! Here's the arithmetic:

.2836 x .2836 x 20,465 = 1,646.0

.2733 x .2939 x 20,465 = 1,643.8

The upshot is that the fact that pitchers become less effective each successive time they face a batter is expected to reduce the number of two-for-twos by barely more than 0.1%.

Mindful of the fact that *theoretically* we expect a 0.9% boost to the number of two-for-twos when we examine actual results for *particular batters facing particular pitchers* (say +1.0% from opposing pitcher variability minus 0.1% from opposing pitcher declining effectiveness), what do we find in practice?

Again I resorted to the data for my 160 batters in their first two official at-bats against all starters that they faced for two official at-bats (or more). Here I took account of the actual batting averages posted by each player. Given that, if (a) every pitcher had identical skills to every other pitcher and (b) if declining pitcher effectiveness was not a

[57] .252 squared plus .284 squared divided by two is only 0.36% higher – i.e., about 1/3 of one percent – than .268 squared. Even if the situation were much more extreme – i.e., if every pitcher in baseball was a .229 or .302 pitcher – it would increase the number of 2/2s by only 1.9% compared to the number of 2/2s that would result from every pitcher being half-way between .229 and .302.

[58] Here I was able to use all starting pitchers because Rick Kaye was able to deliver the information to me in that form.

factor, one would have expected my 160 batters to have 1,668 two-for-twos against starting pitchers in 2004. Allowing for pitcher variability and fatigue this number would have been boosted to approximately 1,683. The actual result: 1,648 is nearly one standard deviation below what one would expect given all of the above factors and chance deviations.

Question 1: What might we expect when we examine stretches of at-bats where batters face two different pitchers over two consecutive at-bats (which is what happens in the large majority of at-bats after the sixth inning and in 100% of pairs of at-bats that span the final at-bat in one game and the first at-bat in the next game)? For sure one should expect to find fewer streaks of two-for-two because the variability in skill from one (random) pitcher to the next is much greater than the decline in effectiveness that occurs when batters face particular pitchers more than once.

Question 2: What might we expect when we examine stretches of at-bats that are long enough (say four or more) to ensure that the skills of opposing pitchers will vary? Fewer streaks for sure should be expected, although modeling this is extremely difficult (at least for me). My guess is that the reduction in streakiness resulting from this will be rather less than the 1% boost in streakiness (specifically, the expected number of two-for-twos) when the same batter faces the same pitcher twice in a row.

Appendix B
Statistical Problems Associated with Measuring Streakiness

In examining stretches of 10 at-bats, one could measure results in (at least) two different ways. We could examine (a) non-overlapping at-bats – 1-10, 11-20, 21-30, 31-40, etc or (b) overlapping at-bats – 1-10, 2-11, 3-12, 4-13, etc. The problem with option (a) is that the sample size ends up being very small – i.e., it is only about one-tenth the size of option (b). But option (b) has a problem as well – namely, the use of overlapping stretches produces distortions in the apparent streakiness (or lack thereof) for individual players. This problem also affects certain group statistics (see below).

To illustrate the problem with option (a), consider a .300 batter with a really weird season. He always gets his hits in the middle 4 or 8 at-bats of each cluster of 20 at-bats. Thus his hits come in at-bats 7-14, 29-32, 47-54, 69-72, 87-94, 108-112, etc. If we use option (a) we will see him getting either four or two hits in *every single* stretch of 10 at-bats during the season – i.e., he will not have any stretches where he goes 5/10 or better or 1/10 or worse. This player will appear to be the complete opposite of streaky. Of course, in truth, this batter is the epitome of streakiness. And, indeed, if we use option (b) we will find one 8/10, two 7/10s, two 6/10s and two 5/10s in every group of 40 stretches of 10 at-bats. That is *much* streakier than would be expected by mere chance.[59]

[59] Comparing this hypothetical with the percentages shown in Table 5 on page 30 we would find that the total number of 5/10s or better is only slightly higher 17.5% vs. 15.0%, but the number of very streaky performances (7/10 or 8/10) are much higher – 10.0% vs. an expected 1.0%.

Meanwhile, as we said, option (b) has problems too. Table 6 (page 30) showed that among our group of 160 players we should have expected to find approximately 10 occasions during the 2004 season where batters went 9/10. The actual result – nine such stretches – was close to our expectation. But when a 9/10 happens it, *of necessity*, generates so many *other* good stretches that it will tend to make the individual player look very streaky. Even if the five at-bats on either side of the 9/10 produce no hits at all, there will be nine additional stretches of 5/10 or better – two each of 5/10s, 6/10s, 7/10s, and 8/10s and one more stretch of at least 5/10.[60] Inasmuch as batters usually do get additional hits just before and/or after their 9/10s, it is not a surprise that 9/10s will usually spawn 13 or more stretches at least as good as 5/10. Here are the 17 stretches of 5/10 or better that Kevin Millar's back-to-back 9/10s created in 2004:

Table B1 – 5/10s or Better Associated with Kevin Millar's 9/10s During 2004

At-bat	Hit	5/10	6/10	7/10	8/10	9/10	5 or more
1							
2	yes						
3	yes						
4	yes						
5							
6							
7							
8	yes						
9							
10	yes	yes					yes
11	yes		yes				yes
12	yes		yes				yes
13	yes		yes				yes
14	yes		yes				yes
15	yes			yes			yes
16				yes			yes
17	yes				yes		yes
18	yes				yes		yes
19	yes					yes	yes
20	yes					yes	yes
21					yes		yes
22				yes			yes
23	yes			yes			yes
24			yes				yes
25		yes					yes
26		yes					yes
Total	15	3	5	4	3	2	17

The upshot is that when we use overlapping stretches of 10 (or 25) at-bats in our analysis, the streakiness (*or* anti-streakiness) of individual players becomes augmented which causes the t-stat to be inflated. The reason this happens is that "nearby" stretches share

[60] The ninth stretch will be either 5/10, 6/10, 7/10, 8/10 or 9/10 – the outcome depending on which at-bat during the 9/10 was not a hit. Each of the five possible outcomes will happen one-fifth of the time.

at-bats with one another and are, thus, correlated with one another. If a particular stretch happens to contain many (or very few) hits, it's "neighbor" will as well.

Consider a batter who is 2/10 in at-bats 1-10. What are his chances of being 5/10 or better in at-bats 2-11? His chances are obviously zero. Indeed we are guaranteed there will be no 5/10 or better until at least his 13th at-bat and at that point a 5/10 is possible *only* if his at-bats 11, 12 and 13 are all hits *and* his at-bats 1 and 2 were not hits. Similarly a .300 batter who goes 7/10 to start the season is not only guaranteed to be 5/10 or better in his next two stretches (at-bats 2-11 and 3-12), but he has an almost 90% chance of being 5/10 during at-bats 4-13 and is a favorite to be 5/10 or better for at-bats 5-14 and 6-15 as well.

We can correct for this problem by multiplying the estimated standard deviation by the square root of the overlap – i.e., by 3.162 in overlapping stretches of 10 at-bats and by 5.000 in overlapping stretches of 25. This has the effect of dividing the resulting t-statistic by the same factors.

Note that this problem (and correction) are used only for Tables 6, 9A and 9B, because of their use of overlapping sets of 10 or 25 at-bats where adjacent sets "share" at-bats. For the results of Tables 7, 10A and 10B, no such adjustment is necessary because there we are only looking at *single* at-bats (either the 11th or 26th following a good or bad stretch). Put another way, if a player is 6/10 in his first ten at-bats of the season, the chances are 100% that he will be 5/10 or better in at-bats 2-11. But if we are just looking at at-bats 11 and 12, his performance in at-bat 11, in no way controls his performance in at-bat 12 – they are totally "independent" events.