

## Underestimating the Fog

If this was a real scientific journal and I was a real academic, the title of this article would be *The Problem of Distinguishing Between Transient and Persistent Phenomena When Dealing with Variables from a Statistically Unstable Platform*. But I was hoping somebody might actually read it.

I have come to realize, over the last three years, that a wide range of conclusions in sabermetrics may be unfounded, due to the reliance on a commonly accepted method which *seems*, intuitively, that it ought to work, but which in practice may not actually work at all. The problem has to do with distinguishing between transient and persistent phenomena, so let me start there.

If you make up a list of the leading hitters in the National League in 1982 (or any other year) and check their batting averages in 1983 (or the follow-up year, whatever it is) you will quite certainly find that those hitters hit far better than average in the follow-up season. If you look at the stolen base leaders in the National League in 1982, you will find that those players continued to steal bases in 1983. If you look at the Hit By Pitch Leaders in 1982, you will find that those players continued to be hit by pitches in 1983. That is what we mean by a persistent phenomenon—that the people who are good at it one year are good at it the next year as well.

If the opposite is true—if the people who do well in a category one year do *not* tend to do well in the same category the next year—that’s what we mean by a transient phenomenon. Here today, gone tomorrow.

All “real” skills in baseball (or anything else) are persistent at least to some extent. Intelligence, bicycle riding, alcoholism, income-earning capacity, height, weight, cleanliness, greed, bad breath, the ownership of dogs or llamas and the tendency to vote Republican . . . all of these are persistent phenomena. Everything real is persistent to some measurable extent. Therefore, if something *cannot* be measured as persistent, we tend to assume that it is not real.

There are, in sabermetrics, a very wide range of things

**BILL JAMES** has been a member of SABR for many years, and is the author of more baseball books than anybody really needs. He is now Senior Baseball Operations Advisor for the World Champion Boston Red Sox.

which have been labeled as “not real” or “not of any significance” because they cannot be measured as having any persistence. The first of these conclusions—and probably the most important—was Dick Cramer’s conclusion in the 1977 *Baseball Research Journal* (SABR) that clutch hitting was not a reliable skill. Using the data from the “Player Win Averages” study by E. G. Mills and H. D. Mills of the 1969 and 1970 seasons, Cramer compared two things—the effectiveness of all hitters in general, and the impact of hitters on their team’s won-lost record, as calculated by the Mills brothers. Those hitters who had more impact on their team’s won-lost record than would be expected from their overall hitting ability were clutch hitters. Those who had less impact than expected were . . . well, non-clutch hitters, or whatever we call those. There are a number of uncomplimentary terms in use.

“If clutch hitters really exist,” wrote Cramer, “one would certainly expect that a batter who was a clutch hitter in 1969 would tend also to be a clutch hitter in 1970. But if no such tendency exists, then ‘clutch hitting’ must surely be a matter of luck.” Cramer found that there was no persistence in the clutch-hitting data—therefore, that clutch performance was a matter of luck. “I have established clearly,” wrote Cramer, “that clutch hitting cannot be an important or a general phenomenon.”

The argument triggered by this article continues to boil, and has now reached the point at which even *Sports Illustrated* is willing to discuss clutch hitting as an open question, at least for one article. But I am not writing about clutch hitting; I am talking about the method. Cramer’s article was very influential. Subsequent to this article, I used a similar method to “demonstrate” that a wide variety of supposed “skills” of baseball players were actually just random manifestations of luck, and many other people have done the same. The list of conclusions which have been bulwarked by this method would be too long to include here, but among them are:

1. There is no such thing as an “ability to win” in a pitcher, as distinguished from an ability to prevent runs. A pitcher who goes 20-8 with a 3.70 ERA is no more likely to win 20 games in the following season than a pitcher who goes 14-14 with a 3.70 ERA on the same team.
2. Winning or losing close games is luck. Teams which win more one-run games than they should one year have

- little tendency to do so the next year.
3. Catchers have little or no impact on a pitcher's ERA. Whether a pitcher pitches well with a given catcher or does not appears to be mostly luck.
  4. A pitcher has little or no control over his hits/innings ratio, other than by striking batters out and allowing home runs. A high hits/innings ratio, if the pitcher has a normal strikeout rate, is probably just bad luck.
  5. Base running, like clutch hitting, has no persistent impact on a team's runs scored, other than by base stealing. If a team scores more runs than they ought to score based on their hits, home runs, walks, etc., it is probably just luck.
  6. Batters have no *individual* tendency to hit well or hit poorly against left-handed pitching. There is a very strong *group* tendency for all right-handed hitters to hit well against left-handed pitchers, but individual deviations from the group tendency have a persistence of zero, therefore are not meaningful.
  7. Batters do not get "hot" and "cold." Hot streaks and cold streaks are just random clusters of events.
  8. A quality hitter in the middle of the lineup has little or no impact on the hitters surrounding him. A good hitter will not hit appreciably better with Manny Ramirez in the on-deck circle than he will with Rey Ordonez on deck.

I will revisit these issues later in the article. For now, trying again to keep clear what I am saying and what I am not. I am not saying that these conclusions are false. What I am saying, and will try to demonstrate beginning in just a moment, is that a method used to reach these conclusions is unreliable to the point of being useless—therefore, that some of these conclusions may be wanting in proof. Let me pick up the sixth item listed above, since, as far as I know, I was the only person ever to make this argument, and therefore there is in that case the least chance that someone will take offense when I try to demonstrate the error.

In the 1988 *Baseball Abstract* (pages 9-15), I tried to do a thorough analysis of platoon data—data for left-handed hitters against right-handed pitchers, etc. I asked a series of questions about the platoon differential, and tried to work systematically through the data toward the answers.

One of the conclusions of that article was: "The platoon differential is not a weakness peculiar to some players. It is a condition of the game." I based this conclusion on the following research and logic. Suppose that you identify, in last year's platoon data, two groups of players: those who had the *largest* platoon differentials, and those who hit better the wrong way

(that is, left-handed hitters who hit better against left-handed pitchers, and right-handed hitters who hit better against right-handed pitchers). Suppose that you then look at how those players hit in the *following* season. You will find that there is no difference or no reliable difference in their following-year platoon differentials. The players who had huge platoon differences in Year 1 will have platoon differences in Year 2 no larger than the players who were reverse-platoon in Year 1.

Individual platoon differences are transient, I concluded, therefore not real. Individual platoon differences are just luck. There is no evidence of individual batters having a special tendency to hit well or hit poorly against left-handed pitchers, except in a very few special cases.

As recently as two years ago I still believed this to be true, although (fortunately) I never succeeded in convincing anybody. The observation was useful, in a sense, because many people pay far more attention to platoon splits for individual hitters than is justified by an understanding of the data—but, in a literal sense, I simply was not correct. Individual batters do have individual platoon tendencies, in many more cases than I at first concluded.

Given a few paragraphs, I could explain how I finally realized that I must be wrong, and how I finally demonstrated that I was wrong, but that's a little bit outside the present article. In any case, this forced me to consider seriously where I had gone astray. My conclusion, which is the basis of this article, was that the "zero persistence equals luck" type of study poses much greater risk of error than I had previously understood.

Suppose that we have two players, whom we will call Allen and Bob. Allen and Bob are both right-handed hitters. Allen hits .290 against right-handed pitchers but .340 against left-handers. Bob hits .290 against right-handed pitchers but .250 against lefties.

From this we attempt to derive a third measurement, which is the player's *platoon differential*. Allen's platoon differential is .050 [.340 minus .290]; Bob's is negative .040 [.250 minus .290]. The platoon differential is what we could call a *comparison offshoot*—a measurement derived from a comparison of other measures.

The first problem with comparison offshoots is that they have the combined instability of all of their components. Every statistic in baseball is to a certain degree a measurement of a skill, to a certain degree a statement about the circumstances, and to a certain degree simply a product of luck. A pitcher goes 20-8—he goes 20-8 to a certain degree because he is a good pitcher, to a certain degree because he pitches for a good team, and to a certain degree because he is lucky (or unlucky). There is luck in everything, and baseball fans are always engaged in

a perpetual struggle to figure out what is real and what is just luck.

In the case of any one statistical record, it is impossible to know to what precise extent it reflects luck, but a player usually bats only 100 to 200 times a year against left-handed pitchers. Batting averages in 100 or 200 at-bats involve huge amounts of luck. If a player hits .340 against lefties, is that 20% luck, or 50% luck, or 80% luck? There is no way of knowing—but batting averages in 100-150 at-bats are immensely unstable. Walter Johnson hit .433 one year in about 100 at-bats; the next year he hit .194. Just luck.

It is hard to distinguish the luck from the real skill, but as baseball fans we get to be pretty good at it. The problem is, that .290 batting average against right-handed pitchers—that also involves a great deal of luck.

When we create a new statistic, platoon *differential*, as a comparison offshoot of these other statistics, the new statistic embodies all of the instability—all of the luck—combined in *either* of its components. Suppose that you take two statistics, each of which is 30% luck, and you add them together. The resulting new statistic will still be 30% luck (understanding, of course, that the 30% number here is purely illustrative, and has no functional definition).

But when you take two statistics, each of which is 30% luck, and you *subtract* one from the other (or divide one by the other), then the resulting new statistic—the comparison offshoot—may be as much as 60% luck. By contrasting one statistic with another to reach a new conclusion, you are picking up all of the luck involved in either of the original statistics.

But wait a minute—the problem is actually much, much more serious than that. A normal batting average for a regular player is in the range of .270. A normal platoon differential is in the range of 25 to 30 points—.025 to .030.

Thus, *the randomness is operating on a vastly larger scale than the statistic can accommodate*. The new statistic—the platoon differential—is operating on a scale in which the norm is about .0275—but the randomness is occurring on a scale ten times larger than that. The new statistic is on the scale of a Volkswagen; the randomness is on the scale of an 18-wheeler. In effect, we are asking a Volkswagen engine to pull a semi.

But wait a minute, the problem is still worse than that. In the platoon differential example, I reached the conclusion I did by comparing *one* comparison offshoot with a *second* comparison offshoot—the platoon differential in one year with the platoon differential the next year. Dick Cramer, in the clutch-hitting study, did the same thing, and catcher-ERA studies, which look for consistency in catcher's impact on ERAs, do the same thing; they compare one comparison offshoot with a second compari-

son offshoot. It is a comparison of two comparison offshoots.

When you do that, the result embodies not just all of the randomness in *two* original statistics, but all of the randomness in *four* original statistics. Unless you have extremely stable “original elements”—original statistics stabilized by hundreds of thousands of trials—then the result is, for all practical purposes, just random numbers.

We ran astray because we have been assuming that random data is proof of nothingness, when in reality random data proves nothing. In essence, starting with Dick Cramer's article, Cramer argued, “I did an analysis which *should* have identified clutch hitters, if clutch hitting exists. I got random data; therefore, clutch hitters don't exist.”

Cramer was using random data as proof of nothingness—and I did the same, many times, and many other people also have done the same. But I'm saying now that's not right; random data proves nothing—and it *cannot* be used as proof of nothingness.

Why? Because whenever you do a study, if your study completely fails, you will get random data. Therefore, when you get random data, *all* you may conclude is that your study has failed. Cramer's study may have failed to identify clutch hitters because clutch hitters don't exist—as he concluded—or it may have failed to identify clutch hitters because the method doesn't work—as I now believe. We don't know. All we can say is that the study has failed.

Dealing now with the nine conclusions listed near the start of the article, which were:

1. Clutch hitters don't exist.
2. Pitchers have no ability to win, which is distinct from an ability to prevent runs.
3. Winning or losing close games is luck.
4. Catchers have little or no impact on a pitcher's ERA.
5. A pitcher has little or no control over his hits/innings ratio, other than by striking batters out and allowing home runs.
6. Base running has no persistent impact on a team's runs scored, other than by base stealing.
7. Batters have no *individual* tendency to hit well or hit poorly against left-handed pitching.
8. Batters don't get hot and cold.
9. One hitter does not “protect” another in a hitting lineup.

On [1], it is my opinion that this should be regarded as an open question. While Dick Cramer is a friend of mine, and I have tremendous respect for his work, I am convinced that, even if clutch-hitting skill did exist and was extremely important, this

analysis would still reach the conclusion that it did, simply because it is not possible to detect consistency in clutch hitting by the use of this method.

There have been other studies of the issue (including several by me) which have reached the same conclusion, but these were in essence repeats of the Cramer approach. If that approach doesn't work once, it's not going to work the second time, the third, or the fourth. It just doesn't work. We need to find some more affirmative way to study the subject.

On (2) above (pitchers have no ability to win, which is distinct from an ability to prevent runs), this, I think, has been a very useful observation over the years, and it now has an additional claim to being true, which is: many predictions have been made based on this assumption which later proved to be accurate.

Simple example: in 2002, Dan Wright went 14-12 with a 5.18 ERA for the Chicago White Sox. It's a data mismatch; a 5.18 ERA should not produce a 14-12 record. Anyone in sabermetrics would immediately recognize this as a strong indication that Wright would *not* be able to continue to win in 2003—and in fact he couldn't, finishing the season 1-7. We have made hundreds of observations/predictions of that nature based on this understanding, and most of these have proven correct. I'm not even going to bring up Storm Davis. Therefore, we probably would not wish to abandon the insight simply because the original proof thereof was faulty.

However, I would have trouble now with my original argument that the pitcher has *no* ability to win, other than what is reflected in his runs allowed. There may in fact be *some* ability to win, in the way the old-time baseball guys imagined that there was. There may be some pitchers who have some ability to win games 3-2 and 9-8. Sabermetrics has traditionally discounted the existence of this ability at *any* level. I would now argue that it may exist at some fairly low level.

On (3) above (winning and losing close games is luck) . . . it would be my opinion that it is probably not *all* luck.

On (4) above (catchers have little or no impact on a pitcher's ERA), I don't think that there is a scintilla of evidence that that is true. It is my opinion that it is impossible to evaluate a catcher's defensive contribution by a comparison based on catcher's ERAs.

Many of the pitcher/catcher combinations which have been studied to reach this conclusion worked together for 40 or 50 innings. ERAs in less than 100 innings pitched have immense instability due to randomness. Further, since the catcher's defensive skill is only one of many, many factors in the prevention of runs, the randomness occurs on a scale which must be 20 times larger than the scale on which the catcher's ERA

contribution must be measured—even if you assume that the catcher's defensive contribution is very large.

Obviously, if a catcher makes a defensive contribution, this must result in a lower ERA for his pitchers. It *seems*, intuitively, that this difference would *have* to be visible in the stats at least at some level, that there would at least have to be *some* measurable consistency in the data. That intuitive sense is what misled me, on this issue, for 25 years. But, in fact, it doesn't. There is so much instability in the data that the catcher's defensive contribution simply cannot be isolated in this form.

On (5) above (the Voros McCracken observation), this seems to me different from the others, for this reason. Voros's observation relies on something which is near to a historical constant. When a ball is in play—not a home run, not a strike-out, not a walk—that ball will be turned into an out about 70% of the time. That is the nature of the game. OK, it's 72% for some teams; it's 67% for other teams; it's 69.5% in some years, it's 68.8% in others. But it doesn't vary crazily from team to team or park to park, and it's really about the same now as it was in 1930 or 1960.

This creates something close to a "stable platform" against which to measure the individual variable, and this makes an important difference. What Voros was saying, in essence, was: "When you see a pitcher who gets outs on 75% of his balls in play, he's just been lucky, because *no pitcher can actually do that*. It's not the nature of the game." This may have been overstated by some people sometimes, but I have little doubt that this observation is more true than false.

On (6) above (base running has no persistent impact on a team's runs scored, other than by base stealing), that's probably not true, and that's probably mostly my error, again. Base running can be measured in simple, objective terms—bases gained, base running errors, etc. A much better way to think about the problem is to measure those things and study what impact they have on runs scored, rather than starting with the proposition that they are probably not meaningful.

On (7) (batters have no *individual* tendency to hit well or hit poorly against left-handed pitching), that, as I said, was just wrong. My mistake.

On (8), this almost becomes a brain teaser. Most baseball fans believe that players get "hot" and "cold." Many analysts believe (and a popular web site is devoted to proving) that this is nonsense, that hot streaks and cold streaks are just random clusters.

Everyone agrees that a hot streak is a transient phenomenon. Therefore, why doesn't everyone agree that it is a non-real phenomenon—a random sequence?

Because people believe that there is *some* persistence to

the transient phenomenon—in other words, that the persistence is not zero.

My opinion is that, at this point, no one has made a compelling argument either in favor of or against the hot-hand phenomenon. The methods that are used to prove that a hot hitter is not really hot, in my opinion, would reach this conclusion whether hot hitters in fact existed or whether they did not.

Stated another way, the hot-hand opponents are arguing—or seem to me to be arguing—that the absence of proof *is* proof. The absence of clear proof that hot hands exist is proof that they don't. I am arguing that it is not. The argument against hot streaks is based on the *assumption* that this analysis would detect hot streaks if they existed, rather than on the proven fact. Whether hot streaks exist or do not I do not know—but I think the assumption is false.

On [9] (batting ahead of a good hitter does not ordinarily cause anyone to hit better), I still believe this to be true. While this analysis relies in part on comparison offshoots, it does so in a more tangential way. I believe that a more careful study, steering clear of comparison offshoots, is still likely to dem-

onstrate that hitters perform (essentially) independent of one another, except in a few isolated cases.

In a sense, it is like this: a sentry is looking through a fog, trying to see if there is an invading army out there, somewhere through the fog. He looks for a long time, and he can't see any invaders, so he goes and gets a really, really bright light to shine into the fog. Still doesn't see anything.

The sentry returns and reports that there is just no army out there—but the problem is, he has underestimated the density of the fog. It *seems*, intuitively, that if you shine a bright enough light into the fog, if there was an army out there you'd have to be able to see it—but in fact you can't. That's where we are: we're trying to see if there's an army out there, and we have confident reports that the coast is clear—but we may have underestimated the density of the fog. The randomness of the data is the fog. What I am saying in this article is that the fog *may be* many times more dense than we have been allowing for. Let's look again; let's give the fog a little more credit. Let's not be too sure that we haven't been missing something important.