Mapping the Fog – Bill James, July, 2005

1. My model

In issue number 33 of the Baseball Research Journal, I published an article entitled "Underestimating the Fog". The thesis of this article is that we in sabermetrics have been relying on a method which doesn't actually work, under closer scrutiny, and we should stop relying on this method. "This method" is the practice of attempting to determine whether some characteristic within the game is "real" or a statistical artifact by comparing whether the players who do well in this area in one year also do well in the same performance category the next year, as one would expect them to if the skill under study was "real". I hope that made sense. . . .I'm a little confused myself, and, speaking of myself, I certainly was not suggesting that other researchers were guilty of this but I wasn't. I was more guilty than anyone. I had misled the public on a series of issues due to my own failure to think clearly about this one matter, and I felt it was important for me to stand up and take responsibility for that.

What I did not do in that article, however, was to establish that what I was saying was true. I argued that this was true, and I discussed the implications of that truth, but I did not attempt, in that forum, to demonstrate that this method does not in fact work reliably. This is the first order of business in this article: to demonstrate that what I was saying before was true (or, more carefully stated, to show you how you can demonstrate this to your own satisfaction.)

Let us take the issue of clutch hitting, which is the most controversial of the many peripheral subjects entangled in the debate. Dick Cramer argued the following in 1977:

1) If clutch hitting really exists, one would expect that the players who were clutch hitters in 1969 would be clutch hitters again in 1970.
2) I have studied who was a clutch hitter in 1969 and who was a clutch hitter in 1970.
3) The lists do not correlate to any notable extent.
4) Ergo, clutch hitting does not exist.

I accepted this argument for about a quarter of a century, but eventually it began to trouble me. When it began to trouble me enough, I posed a counter question to myself: is it possible to create a model in which clutch hitting clearly exists, but goes undetected by this type of analysis?

It is, in fact, possible. Let us create a “model league” based on the following assumptions:
1. The league consists of 100 batters.
2. Each batter has 600 at bats.
3. Of those 600 at bats, 150 are in clutch situations, 450 are not.
4. The average hitter will hit .270.
5. Individual batting averages can range from .170 to .370, but are normally distributed (bell-shaped curve) and are clustered around .270.
6. Eighty percent of the players will have the same expected batting average in clutch situations as in ordinary situations
7. However, the other 20% may hit significantly better or significantly worse in clutch situations than they do overall.

In clutch situations, the batting average of the other twenty percent was re-calculated as

Their regular batting average,
Minus 50 points,
Plus a random number between zero and one, divided by 10.

Thus, a .280 hitter in non-clutch situations can be a .230 hitter in clutch situations, or a .330 hitter in clutch situations, or anywhere in between, and any one figure is as likely as any other—for those players who did have a “clutch element” in their makeup. The average clutch effect, for those players who have one, is 25 points positive or negative.

You may or may not agree that this model represents a fair test of the clutch thesis. If you agree that it does, end of subject. If you would argue that it does not. …Dick Cramer, in his 1977 article, stated that “I have established clearly that clutch-hitting cannot be an important or general phenomenon.” I would argue that if 20% of the hitters have clutch effects averaging 25 points, that is quite certainly an important and general phenomenon. Further, in several respects, this model exaggerates the impact of clutch hitting, which should make it easier to detect whether or not a clutch hitting ability is an element of the mix. In this league there were 60,000 at bats, which were neatly divided into 600 at bats each for 100 players. In the real American League in 1969—one of the leagues included in Cramer’s study—there were 65,536 at bats, but there were only 25 players who had 550 or more at bats, the rest of the at bats being messily distributed among players who had 350, 170, 80 and 4 at bats. This would make it much easier to detect the presence of clutch hitters in the model than in real life.

In the real leagues studied by Cramer, there were many players who had 520 at bats one year but 25 the next, making those players—and those at bats--essentially useless as a basis for year-to-year comparison. In my model, all 100 players had 600 at bats each year, with no one dropping out or coming in. This, again, would make it vastly easier to have meaningful year-to-year comparisons, in my model, than it would be in real life.

In my model, one-fourth of all at bats are designated as "clutch" at bats. In real life, it seems unlikely that the number of true "clutch" at bats would be that large. In real life, a player probably has 50 or 75 high-pressure at bats in a season. In my model, he had 150. This would make it vastly easier to detect clutch performers in the model than it would be in real life.

In my model, all at bats are cleanly delineated as "clutch" or "non clutch". In real life, it is extremely difficult to say to what extent any at bat is "clutch" or "non clutch". Again, this would it make it much, much easier to detect the presence of clutch hitters in this model than it would be in real life.

If you object to the fact that only 20% of the players in this study had some clutch ability:
a) what if only 20% of players in real life have some clutch ability?, and
b) it isn't crucial anyway. The conclusion wouldn't change if it was 40% or 50%.

Having constructed this model, I then simulated on a spreadsheet 600 at bats for each player—450 in non-clutch situations and 150 under clutch conditions—and figured for each player his batting average in "clutch" situations and his batting average in non-clutch situations. I did this for two seasons for each of the 100 players, creating a "clutch differential" for each player in each season. Each player's intended batting average changed from season to season, but his "clutch differential" remained the same. The spreadsheet on which this experiment was conducted is named "Clutch Consistency.XLS", and I will e-mail a copy of this spreadsheet to anyone who asks. At first glance it just looks like a vast collection of random numbers, but I think you can figure it out with a little effort.

This method does not exactly mirror Cramer's method, in his 1977 article which I was using as a kind of whipping boy in Underestimating the Fog. What I have described as "Cramer's method" is in fact two methods—an (a) method which was used to determine whether a player was a clutch hitter in any given season, and a (b) method which was used to determine whether those players identified as clutch players were consistent from season to season. I was interested entirely in the questions raised by the (b) method. The subject of my article could be stated as "Will Cramer's (b) method work reliably under real-life conditions, if we assume that his (a) method works?"

The (a) method I never discussed at all, for three reasons—
1. That this was not relevant to my article,
2. That his (a) method is much more complicated, and much harder to replicate in a model, than the method I preferred, and
3. I'll tell you later.

Anyway, in my model, we know that clutch hitting does exist, and that it does exist at what seems to me a very significant level. Yet when I compared the "clutch differentials" of the 100 players in the two seasons, the year-to-year consistency was far, far below the level at which any conclusion could be drawn from the data. Despite all of the steps I took to make clutch ability easier to spot in the model than it would be in real life, it remains essentially invisible.

In the study, a player's clutch contribution was labeled as "consistent" if he hit better in clutch situations than he did overall in both simulated seasons, or if he hit worse in both seasons. His clutch contribution was labeled as "inconsistent" if he was better one year and worse the other.

As you would expect, 50% of the players who had no actual clutch differential were consistent, and 50% were inconsistent. But of the players who did have actual clutch differentials, 62.2% were consistent, while 37.8% tested as inconsistent, given these conditions.

Overall, then, 52.4% of the players in the study showed consistency in their clutch contribution. If 52.44% of the players in a group are consistent from year to year and there are 100 players in the group, what is the random chance that 50 of them or fewer will show up as consistent in one test?

It's 35%. Thus, no conclusion whatsoever can be drawn from the apparent lack of consistency in the data. Even when we know that the clutch effect does exist within the data, even when we give that effect an unreasonably clear chance to manifest itself, there is still a 35% chance that it will entirely disappear under this type of scrutiny.

What if 40% of the players have an "actual clutch effect", rather than 20%?

At 40%, there is still a 14% chance that 50 or fewer of the 100 players will have positive year-to-year consistency—which means that we are still in a position where no conclusion can be drawn from the lack of documented consistency. Even if 50% of the players have an actual clutch effect, there remains a 9% chance that this would not show up in a test of 100 players.

2. Random Observation

Part of the problem with measuring “agreement” is that “agreement” narrows the odds, and thus profoundly changes the percentages. Suppose that half of the players in a group are good clutch hitters, and half are poor clutch hitters. Suppose that you have a test of clutch ability which is 80% accurate. Under those conditions, how many players will measure as consistent, meaning that they measure the same both years?

68%. 64% will measure as “consistent” accurately--.80 times .80—and 4% will measure as “consistent” due to a repeated inaccuracy. If the measurement is 80% accurate, in a two-year period 64% of the players will have two accurate measurements, and 4% will have two inaccurate measurements.
If the test of clutch ability is 70% accurate, then, it will test as 58% accurate (.49 + .09). If the test of clutch ability is 60% accurate, it will test as 52% accurate (.36 plus .16).

Thus, in order to achieve 62% agreement, as we did in the model above, you have to have a test which is 75% accurate. This is actually more of a problem in the catcher-ERA studies than it is in the clutch hitting studies.

3. Reaction to Underestimating the Fog

In the first few weeks after “Underestimating the Fog” was published, I got reactions which were all over the map. However, the one thing that nobody said, in the first few weeks—at least, nobody said it where I happened to see it—was that what I was saying was not correct. Thus, I felt no pressure, in those opening weeks, to demonstrate that what I was saying was correct.

However, in the February, 2005 edition of By the Numbers—which I think came out in June, 2005, go figure—there were two articles which touched on the veracity of my central claim, and thus prompted me to put my supporting work on record.

These two articles tend to broaden the debate, and raise a number of points that I wanted to comment on. In the first of those two articles (Comments on “Underestimating the Fog”), Jim Albert writes:

I was interested in a statement that James made in this article regarding the existence of individual platoon tendencies. This was counter to the general conclusions Jay Bennett and I made in Chapter 4 of Curve Ball.

However, Dr. Albert doesn’t say what the statement was that he disagreed with, and, pardon my obtuseness, but I’m not able to figure it out. I’ve read his comment three or four times, but my math skills are limited, and I just can’t figure out what it is I said that he disagrees with. My ability to respond is thus impaired.

With this exception, I think that the rest of Dr. Albert’s comments, including those critical of the article, seem to me to be fair and well-considered, and I have no response to them.

The following article, however, the Phil Birnbaum article entitled “Clutch Hitting and the Cramer Test”, contains a number of statements that I wanted to comment on.
1) For the sake of clarity, the issue that I was discussing in Underestimating the Fog is peripheral to Birnbaum’s article, and the issue that Birnbaum is discussing in his article was on the periphery of my article. I was writing about whether Cramer’s (b) method works. Birnbaum is writing about whether Clutch Hitting could exist. These are not articles discussing the same subject.

2) I don’t think that Birnbaum himself is confused about this (point 1), but he appends to his article a head-note which seems to suggest that he is responding directly to my article, and follows this by quoting two or three things I had said and responding to them. This creates the impression, to the reader, that we are writing about the same central issue. The longer his article goes, the more it drifts away from being a response to Underestimating the Fog.

3) In my article I had written that “random data proves nothing—and it cannot be used as a 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.”

In response to this, Birnbaum says that “This is certainly false. It is true that when you get random data, it is possible that ‘your study has failed.’ But it is surely possible, by examining your method, to show that the study was indeed well-designed, and that the random data does indeed reasonably suggest a finding of no effect.”

Reasonably suggests? We’re not talking about reasonable suggestions here; we’re talking about valid inferences from the data. Cramer didn’t say that his data “reasonably suggests” the absence of clutch hitters; he said—incorrectly—that his data “established clearly that clutch hitting cannot be an important or general phenomenon.” Joe Morgan, Tim McCarver, and generations of sportscasters before them have reasonably suggested that some players may have a special ability to rise to the occasion. The task in front of us is not to reasonably suggest the opposite, it is to find clear and convincing evidence one way or the other.

In the process of doing this, studies resulting in random data show only that the study has failed to identify clutch hitting ability. I stand by my statement without any reservation.

4) What Birnbaum means by “The Cramer Test” in his title is also Cramer’s (b) method; he doesn’t actually use Cramer’s (a) method, either. In point of fact, nothing in Birnbaum’s article examines the effectiveness either of Cramer’s (a) method OR his (b) method. Birnbaum’s article examines not whether Cramer’s method works, but whether his conclusion—that clutch hitting doesn’t exist at a significant level--is true. He poses this question:

Bill James’ disputes this result, writing that “it is simply not possible to detect consistency in clutch hitting by the use of this method.” Is he correct? If clutch hitting were a consistent skill, would the Cramer test have been powerful enough to pick it up?

But he never actually addresses this question. His subsequent research has to do with whether Cramer is correct, and has nothing at all to do with whether his method works. He drops Cramer’s (a) method, and performs a test of statistical significance on the (b) method, the results of which, in my opinion, he misinterprets.

5) For the sake of clarity, I take no position whatsoever about whether clutch hitting exists or does not exists. I simply don’t have any idea.

6) Either Birnbaum or myself is profoundly confused about the difference between “no evidence of effect” and “evidence of no effect.” Birnbaum writes:

The results: a correlation coefficient ® of .0155, for an r-squared of .0002. These are very low numbers; the probability of an f-statistic (that is, the significance level) was .86. Put another way, that’s a 14% significance level—far from the 95% we usually want in order to conclude that there’s an effect.

But this data—and all of Birnbaum’s data—actually doesn’t indicate that there is no effect. In fact, it shows that there is some evidence that there may be such an effect, but that this evidence merely is far too weak to say for sure one way or the other. This is a very, very different thing—and one absolutely may not segue from one into the other in the way that Birnbaum is attempting.

Why? For this reason. Suppose that you took a ten-at-bat sample of Stan Musial's career, and asked "does this ten at bat sample provide clear and convincing evidence that Musial was an above-average hitter?"

Of course the answer would be "no, it doesn't." In the ten at bats Musial might go 4-for-10 with 2 homers, but in a ten-at-bat sample, A. J. Hinch might go 4-for-10 with 2 homers. You would conclude, by Birnbaum's method, that this provided very, very little evidence that Musial was in fact an above-average hitter.

Suppose that you broke Musial's 1948 season down into a series of 61 ten-at-bat sequences, and tested each one for evidence that Musial was an above-average hitter.

You would certainly fail, 61 times in a row—indeed, in many of those ten-at-bat samples, Musial would appear to be a below average hitter.

By Birnbaum's logic, this would provide overwhelming evidence that Stan Musial in 1948 was not really an above-average hitter, since he had failed 61 straight significance tests.

But wait a minute. . .the real-life problem is worse than that. Suppose that you took each ten-at-bat sample of Musial's season, and you buried it in a pile of one thousand at bats by ordinary hitters, and you then tested the significance of the 1010-at-bat composite. This would make the f-statistic (significance level) much higher, while making the correlation coefficient even lower. You quite certainly would find no evidence whatsoever that Musial was pushing the group to be above average.

This is the real-life problem that we confront here. The clutch hitting contribution, if it does exist, is buried in large piles of random and confusing data, with very little marking the clutch contribution to enable us to dig it out and examine it.

I'm not saying it can't be done; there are lots of clever people in the world, and it probably can be done, eventually. But the problem is a hell of a lot harder than Birnbaum realizes.

7) Birnbaum writes "Let's suppose a clutch hitting ability existed, that the ability was normally distributed with a standard deviation (SD) of .030 (that is, 30 points of batting average.)"

But the scale proposed here is massive. The standard deviation of batting average itself isn't thirty points. The standard deviation of batting average, for all players qualifying for the batting title in the years 2000 to 2004, is 28 points (.0277).

Birnbaum's argument is "if a clutch hitting ability existed on this scale, this analysis would find it." But if a clutch hitting ability existed on anything remotely approaching that scale, Stevie Wonder could find it. If a clutch hitting ability existed on anything like that scale, we wouldn't be having this discussion.

If the standard deviation of clutch ability was 30 points, there would be a very significant number of players who hit 50 points better in clutch situations, throughout their careers. If that was the case, we would have known it 20 years ago. If the standard deviation of clutch ability was 30 points, there would be one or two players in each generation who would improve their performance in clutch situations by 100 points. We could find that without doing any of this stuff.

8) No one has ever suggested that clutch hitting operates on that scale. Listen to the things that Tim McCarver says about clutch hitting, or Joe Morgan, or any of those druids. What they are saying is not that EVERYBODY has some huge clutch effect, but rather, that there are some few players—some tough, veteran players who have real character, and who might someday even go on to become TV broadcasters—who are able to come through in the clutch. Sometimes.

In my model of the problem, I envisioned this as 20% of the players, having a clutch effect of 25 points (.025). That creates a standard deviation, for the group as a whole, of eleven points.

Maybe it's not eleven; maybe it's 12, or 14, or 6, or 2. It sure as hell isn't 30.

9) Let us talk for a moment about Cramer's (a) method.

Cramer's (a) method—his method of determining whether a player was or was not a clutch hitter—was to contrast two measurements. One was an estimate of the player's presumptive win contribution, based on his total batting statistics. A home run is a home run. If a player hit a home run in the ninth inning of a 12-1 ballgame, that was the same as if he hit a walk-off homer in the bottom of the ninth. The other was an event-by-event assessment of what the player had contributed to his team's wins. If a player hit a home run in the ninth inning of a 12-1 ballgame, that would essentially be a non-event, whereas if a player hit a David Ortiz shot, that might be worth 100 times as much.

If a player ranked much better in the second evaluation than in the first, Cramer's (a) method designated him a clutch hitter. If he ranked much better in the first evaluation, Cramer designated him as a non-clutch player.

Neither Birnbaum nor I, in discussing Cramer's article, made any effort to replicate or to examine this method, what I have been calling Cramer's (a) method. We both tested his (b) method, but replaced his (a) method with something more straightforward. I had three reasons for not doing so, two of which I explained before.

My third reason for skipping this system is that I wanted a system which I knew would work. I wanted to test whether or not Cramer's (b) method would work if we assumed that his (a) method worked reasonably well. I therefore substituted an (a) method that I knew would work, demonstrated that it did work, and moved forward from there.

This leaves unexamined the question of whether or not Cramer's (a) method would work. Could one, in fact, identify clutch hitters by contrasting a player's overall offensive work with his win contribution, figured from the sequence of events?

I don't know. I'm skeptical. I doubt that it would work. The problem, it seems to me, is that the method might be heavily liable to random influences.

Here's how we could tell if the method works or not. . ..I'll get around to doing this eventually, I suppose, if nobody beats me to it. Construct a "model universe", as I did in my study, and designate 15 or 20% of the players as clutch hitters, as I did in my study. Then simulate games, and evaluate the output by the method Cramer used to evaluate the real-life events.

One would then be in a position to ask "do the players who are ACTUALLY clutch hitters, in the underlying codes, show up as clutch hitters in the output?" By random chance, 50% of them would show as better clutch hitters than neutral-case hitters. By the method I used, 75% of the clutch hitters were identified as clutch hitters. I would be very, very surprised if Cramer's (a) method would match that. I would guess you would get 53, 55% accuracy, somewhere in there.

Why? Too much weight on too few outcomes. I am guessing—but I don't really know—that in Cramer's (a) method, 50% of the variance between the player's situation-neutral win contribution and his situational win contribution will be determined by 30 at bats by fewer (if the player plays regularly). Thus, the player's ranking in this system would seem to be heavily influenced by random deviations in performance in a small number of at bats, and thus the players who were “truly” clutch hitters, in the model, might very often not be identified as clutch players.

10) Again for the sake of clarity, I am not suggesting that my “clutch indicator” systems works, either. My system worked, in my model, only because I set up the model to enable it to work within the model. It wouldn’t work worth a crap in real life.

Also, my system was 75% accurate only in the sense of agreeing that a clutch hitter was a clutch hitter if we already knew that he was. But my system would also identify as clutch hitters a large number of players who actually weren’t coded to hit well in the clutch, but who had merely done so at random.

Ultimately, what we need is a system which can reliably identify a clutch hitter, if one exists. That doesn’t seem to me like an impossible problem. But we’re nowhere near to having such a thing.

11) Birnbaum did attempt to demonstrate that his (a) method worked; he just did a couple of things that, in my opinion, undermine his attempt.

Look, what I was trying to say in Underestimating the Fog is “You can’t assume that your system works. You have to prove that it works. You have to demonstrate that it works, detail by detail.”

We are no closer to that now than we were a year ago. Cramer’s article remains immensely important, for this reason: that it proposed a road map through a wilderness. That was a wonderful thing; I appreciated that 28 years ago, and I appreciate it now.

But the first maps drawn of America showed huge waterways cutting through the Rocky Mountains—and that was after the explorers finally realized they weren’t in India. Maps drawn of the moon even fifty years ago were comically inaccurate.

I don’t know how accurate Cramer’s (a) method really is. But the limitations of his (b) method are such that, even if his (a) method was 100% accurate, that might not be enough to justify the conclusions he thought he had reached. . .the conclusions that we thought he had reached. I doubt that the (a) method works, either.

It is my opinion that there is an immense amount of work to be done before we really begin to understand this issue.