Response to "Mapping the Fog" – Phil Birnbaum, July, 2005

In a famous 1977 clutch-hitting study, Dick Cramer took 122 players who had substantial playing time in both 1969 and 1970. He ran a regression on their 1969 clutch performance versus their 1970 performance. Finding a low correlation, he concluded that clutch performance did not repeat, and that, therefore, this constitued strong evidence that clutch ability did not exist.

> Bill James, in his recent essay "Underestimating the Fog," disputes that the Cramer study did indeed disprove clutch hitting.

My essay, "Clutch Hitting and the Cramer Test," explicitly disagreed with the second of these points, and implicitly with the first. In "Measuring the Fog," Bill James criticized aspects of my essay, and reasserted that his position was correct.

But I still believe that Bill is not correct.

Bill's position can be summarized by these two quotes, from "Underestimating the Fog:"

> "… even if clutch-hitting skill did exist and was extremely important, [Cramer's] analysis would still reach the conclusion that it did, because it is not possible to detect consistency by the use of this method [regression on this year's clutch performance against next year's]."

> "… 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."

To which I respond:

1. Yes, random data *on its own* proves nothing. But combined with evidence that your test *would have found an effect if it existed*, the random data is evidence that the effect doesn't exist.

2. It *is* possible to detect clutch-hitting consistency (at reasonable, non-trivial levels) by the use of the Cramer test.

3. It *is* possible to show what effects the Cramer test is capable of finding, and, therefore, to what extent a "finding of no effect" disproves clutch hitting.

On number 1, Bill charges me with a fallacy – the fallacy of believing that, if a test finds no evidence of clutch hitting, this means that clutch hitting does not exist. I agree with Bill that this logic would be seriously incorrect – but I neither stated it nor implied it. My point was that if a test finds no evidence of clutch hitting, *and you can show that the test would have found clutch hitting if it existed*, well, then, and only then, are you entitled to draw a conclusion about the non-existence of clutch hitting. Either Bill misread what I said, or I didn't say it clearly enough.

Point number 2 is the most important, because it's the point of greatest contention between Bill and myself. Bill thinks you can't detect clutch hitting by the use of the Cramer test. I believe you can.

The reason for the difference is *that we're using different tests*.

Bill's test, in essence, consists of looking at players in consecutive years, and assigning each player one of four symbols. He gets a "+ +" if he was a clutch hitter both years; "- -" if he was a choke hitter both years; and "- +" or "+ -" if he was split. Bill then counts the number of consistent players (+ + or - -), and compares it to the number of inconsistent players (+ - or - +). If clutch hitting existed, there would be significantly more consistent players than inconsistent.

My test – which is the same test that Cramer used (but with Bill's measure of clutch rather than Cramer's "(a)" measure, as Bill calls it), uses the actual numbers, and runs a regression. So if player A was 50 points higher in the clutch one year and 10 points higher the next, I add the pair (+50, -10) to my sample. I then run a (standard STAT101) regression on all the pairs, and look for a significance level.

The point is that Bill's test is much, much weaker than mine. I think Bill is correct that with his test, "even if clutch-hitting skill did exist and was extremely important," the test would be incapable of finding it.

By using only the signs, James lost a huge amount of information – he kept the consistency aspect, but not the *amount* of consistency. To James, a hitter who hits one point better in both years gets the same weight as a player who hits 50 points better in both years.

(As an aside, I'd bet that if Bill threw out all datapoints except those where the absolute value of clutch hitting was over 25 points both seasons, the test would be much more likely to find signficance. But that's not important right now.)

By analogy, suppose that team A wins three games against the Brewers all by scores of 5-4, while team B wins three games against the same Brewers all by scores of 10-1. Bill's test treats the teams the same, scoring them both as "+ + +", and is incapable of noticing that team B is actually much better than team A.

But to my test (and Cramer's), the amount of clutch hitting is considered. And so the Cramer test *is* capable of finding significant clutch effects.

My first test asked this question: if clutch hitting were normally distributed with standard deviation of 30 points, would the Cramer test find it?

It would and it did. The second row of my table (at the top of page 10 of "Clutch Hitting and the Cramer Test"), contains the results of 14 simulations of a season where clutch hitting was normally distributed with an SD of 30 points. Of those 14 simulations, the Cramer test found the effect, with statistical significance, in 11 of those 14 seasons. Seven of those 14 were *extremely* significant, rounding to .00.

Now, you could argue that 11 out of 14 isn't enough – the test is only powerful enough 79% of the time. 21% of the time, the test will fail.

And that's true if you only run the test on one season's worth of data. But I ran it on 14 seasons. If clutch hitting at the .030 level should be caught 11 out of 14 times, and the real-life data (top row of the same table) showed significance 0 out of 14 times, does that not "reasonably suggest" (Bill doesn't like this expression) that clutch hitting at .030 does not exist?

In my essay, I stopped there, but I could have done a more formal calculation. It looks like there's about a 21% chance of failing to find significance for a single season. Let's up that to 30% just to be conservative. We found 14 of those in a row. What's the chance of a 30% shot happening 14 times in a row? 1 in 21 million.

What is Bill's response to this test in "Mapping the Fog"? He doesn't dispute the method or conclusion. Rather, he argues that .030 is a massive SD for clutch hitting (I implied that it was moderate; Bill is correct – it is massive). Of course this method can find an SD of 30 points, Bill says. "Stevie Wonder could find it."

Bill writes, "maybe [the SD is] … 12, or 14, or 6, or 2. It sure as hell isn't 30."

Which is fair enough. But my original essay actually does go on to repeat the same test for 20 points, then 15 points, then 10 points, then 7.5 points – using exactly the same method, which Bill doesn't dispute (and uses himself, as we will see shortly). Bill does not mention these subsequent tests at all – nor does he mention my conclusion that the Cramer test (with 14 seasons of data) is "doubtful" with a standard deviation of 10 points, and that I agree with him that it "fails" if the SD of clutch hitting is actually only 7.5 points.

In "Measuring the Fog," Bill suggests a different distribution – he supposes 80% of the population has no tendency for clutch hitting whatsoever, and the 20% vary uniformly (ie, a flat curve rather than a bell curve) between -50 points clutch and +50 points clutch. He goes on to do a very similar simulation to what I did. He finds (and I agree) that the test is very weak, and will almost always fail to find an effect.

But Bill used his "signs" test rather than the Cramer regression, and that's why he failed to find any effect.

To prove that, I repeated my regression, but used the James distribution rather than my normal distribution. (James says the SD of his distribution is .011, but I found .013.)

My results: out of my 56 simulated seasons, 11 showed statistical significance at the .05 level in a positive direction. If the data were random, it should have been 2.5% of 56, or 1.4.

Again I didn't do this in the essay, but what is the probability of getting exactly 11 positives out of 56, where the chance of each positive is 2.5%? If I've done the calculation right, it's about 1 in 8.6 million. (It's higher for 11 or more, but I'm too lazy to run the normal approximation to binomial right now. It's definitely less than 1 in a million, in any case.)

(By the way, I think the 11 successes might have been a random fluke. But even if we got only 6 successes, I (lazily) believe that would still significant at the 1% level.)

In point form, then:

-- Under Bill's distribution, the simulated Cramer Test succeeded in finding positive significance about 19% of the time in 56 tries.
-- Random data would, by definition, find positive significance 2.5% of the time.
-- The chance of the 19% happening by chance in 56 tries, where the real probability is 2.5%, is much less than 1 in a million.

On that basis, I would conclude that the Cramer test over 14 single seasons "reasonably suggests" that a level of clutch hitting as described by Bill's distribution does not exist.

But I guess there are really two conclusions:

-- With 14 separate seasons worth of data, the Cramer test "works" in that it identifies the existence of clutch hitting at the Bill James distribution;
-- As an aside, the real-life data do provide reasonable basis to conclude that if clutch hitting does indeed exist, it does so at a lower level than the Bill James distribution.

So, now going back to James' original two quotes:

> 1. "… even if clutch-hitting skill did exist and was extremely important, [Cramer's] analysis would still reach the conclusion that it did, because it is not possible to detect consistency by the use of this method [regression on this year's clutch performance against next year's]."

It seems to me that Bill believes this because he used a much weaker signs test, rather than a full regression. (Although, to be fair, I don't know whether the Cramer test succeeds using Cramer's own measure of clutch hitting. It might, or it might not.) I believe that the data and logic fully support the conclusion that for a large enough effect (such as Bill's distribution) and enough seasons of data (say, the 14 that I used), the Cramer test quite easily detects consistency.

> 2. "… 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."

As I argued earlier, I believe this is not true – random data, *combined with a powerful enough test*, is legitimate evidence of "nothingness" – or at least, a small and bounded amount of somethingness.

And, judging by Bill's response, I don't think he believes this quote himself. His own test of whether the signs test would pick up an effect proves that. If he really believed that random data proved nothing, what would be the point of checking if the test could produce non-random data? Answer: he really means that random data proves nothing *only if random data would come out in any case*.

And so I wonder if by this quote, Bill actually meant what I meant – that *taken alone*, random data is not sufficient proof of nothingness – and just overstated his case.

Having said all this, my overall impression is that James and I do, in fact, substantially agree, and that a large part of our disagreement stems from the fact that James used a test that *doesn't* work, whereas I used a test that *does* work. James correctly concludes that you can't disprove clutch hitting from his test, and I (believe I) correctly conclude that you can disprove a certain level of clutch hitting from my test.

James writes that "I take no position whatsoever about whether clutch hitting exists or does not exist." But he does acknowledge that if clutch hitting exists, it must have a standard deviation that doesn't even approach 30 points ("or Stevie Wonder could find it"). My position is similar – I don't know whether it exists or not either -- but I believe that if it *does* exist, the simulations prove that the Cramer test has lowered the possible SD down to 10 points, or even less.

Our only large disagreement, I think, is that Bill argues very strongly, in absolute terms, that the Cramer method can't work. I argue that the absolutist formulation is wrong. The Cramer method is as legitimate as any other statistical method. With enough data – exactly how much data depends on the size of the effect you're looking for -- the test is powerful enough to provide good evidence for the lack of the effect.