By the Numbers

Volume 15, Number 1 The Newsletter of the SABR Statistical Analysis Committee February, 2005

Summary

Academic Research: Baseball Players and Superstition

Charlie Pavitt

An academic study that investigates superstition among professional baseball players finds some interesting differences between fielding, hitting, and pitching superstitions, as well as between American and Japanese players.

This is one of a series of reviews of sabermetric articles published in academic journals. It is part of a project of mine to collect and catalog sabermetric research, and I would appreciate learning of and receiving copies of any studies of which I am unaware. Please visit the Statistical Baseball Research Bibliography at its new location www.udel.edu/communication/pavitt/biblioexplan.htm . Use it for your research, and let me know what is missing.

Jerry M. Berger and Amy L. Lynn, Superstitious Behavior Among American and Japanese Professional Baseball Players, Basic and Applied Social Psychology, 2005, Volume 27 Number 1, pp. 71-76

I must admit that I questioned whether this paper deserved my attention in the first place, but when I saw that the authors had cited a BTN study (Rob Wood's *How Often Does the Best Team Win the Pennant?*, from Volume 10, Number 2), I gave it the benefit of the doubt. Berger and Lynn framed their study as a test of the idea that superstitious behavior is more likely to occur when people feel that the outcomes of their efforts are determined by forces they cannot control. This idea led them to two baseball-relevant hypotheses: that superstitious behavior would be correlated with the belief that chance or luck has an effect on what occurs during a game, and that this belief would be stronger for batting and pitching than for fielding, because the outcomes of the former two are less certain. In addition, they hypothesized that cultural differences between the "individualistic" Americans and "collectivist" Japanese implies that the Japanese would be less superstitious (as their culture emphasizes that outcomes are a result of effort) and more likely to perform superstitious behavior for the good of the team rather than the self.

The authors distributed surveys to players on five American teams (Angels, Red Sox, Indians, Giants, and Devil Rays) and three Japanese teams (Marines, Hawks, Fighters), and received responses from 50 Americans and 27 Japanese. They found 74.3 percent listed at least one superstition they sometimes engaged in and 53.3 percent admitted to at least one every game.

Although the players' reports of how often they engaged in superstitious behavior and the extent to which they felt it had an impact both correlated between .3 and .4 with their reported belief in the impact of luck on outcomes, the overall mean score of an item asking how much effect their superstitious behavior has on their outcomes was only 2.37 on a 1-to-5 scale. This lead the authors to speculate that much superstitious behavior may be performed not because the players feel it improves their luck but rather because it is normative (everyone else does it) or as part of comforting rituals. Although the differences among them did not reach statistical significance, the percentage believing in the efficacy of superstitious behaviors was indeed less for fielding (15.9%) than

In this issue

for batting (22.8%) and pitching (31.7%). The cross-cultural hypotheses were mostly supported; the Americans were more likely to perform superstitious behaviors (although, surprisingly, they were less likely than the Japanese to believe they had an impact), and the Americans were also more likely to believe these behaviors impacted their personal outcomes -- but less likely to believe they impacted their team's outcomes -- than the Japanese.

Charlie Pavitt, 812 Carter Road, Rockville, MD, 20852, chazzq@udel.edu ♦

Submissions

Phil Birnbaum, Editor

Submissions to *By the Numbers* are, of course, encouraged. Articles should be concise (though not necessarily short), and pertain to statistical analysis of baseball. Letters to the Editor, original research, opinions, summaries of existing research, criticism, and reviews of other work are all welcome.

Articles should be submitted in electronic form, either by e-mail or on PC-readable floppy disk. I can read most word processor formats. If you send charts, please send them in word processor form rather than in spreadsheet. Unless you specify otherwise, I may send your work to others for comment (i.e., informal peer review).

If your submission discusses a previous BTN article, the author of that article may be asked to reply briefly in the same issue in which your letter or article appears.

I usually edit for spelling and grammar. If you can (and I understand it isn't always possible), try to format your article roughly the same way BTN does.

I will acknowledge all articles within three days of receipt, and will try, within a reasonable time, to let you know if your submission is accepted.

Send submissions to:
Phil Birnbaum
88 Westpointe Cres., Nepean, ON, Canada, K2G 5Y8
birnbaum@sympatico.ca

Comments on "Underestimating the Fog"

Jim Albert

In a recent BRJ, Bill James wrote that many accepted sabermetric studies are methodologically flawed, arguing that they were incorrectly designed to be able to reach the conclusions they did. Here, the author considers and comments on James' arguments, and, as a bonus, presents a result on whether the platoon advantage varies among players.

I enjoyed Bill James' recent article, *Underestimating the Fog*[1] since it relates to one of the major issues facing sabermetricians. We observe interesting patterns in baseball data. Are these patterns meaningful or informative about the abilities of players or teams? Or are these patterns just a by-product of chance variation? Using James' terminology, distinguishing between transient and persistent phenomena is one of the difficult tasks we face. (Indeed one of the main points in our book *Curve Ball* was to emphasize the important role that chance plays in the variation of baseball data.) I agree with James that the chance variation in baseball data is analogous to a dense fog and we have to look carefully through the fog if we wish to make accurate statements about the qualities of teams and players. Although I agree with James' general conclusions, unfortunately I think that he is unclear and sometimes wrong in some of his statements about chance variation. Hopefully I can clarify some of his comments and at the same time describe a recent short study of mine that confirms one of his statements about individual platoon tendencies.

What Does it Mean for a Statistic to be 50% Luck?

James says that one baseball statistic can be 30% luck and another statistic can be 80% luck. Since he doesn't define luck percentage in his article, what does this mean? Let me try to explain my interpretation of this concept using batting averages, although this explanation can apply to any statistic. Suppose that we have a collection of 100 batting averages from the 2004 baseball season. Obviously we will see a lot of variation – we will see averages ranging from .200 to .350. What is causing this great variation in AVGs? Actually, there are two principal causes. The batting averages vary partly because players have different batting abilities, and partly due to the inherent chance variation in a player's batting average. This chance variation is basically the same type of random variation we get when we flip a coin many times and record the number of heads. We are interested in how much of the variation in batting averages is due to the difference in batting abilities and how much is due to chance or luck. If we divide the variation due to luck by the total variation (variation due to luck and due to difference in abilities), we get the luck percentage. If you prefer to see a formula, here it is:

$$LUCK\ PCT = 100 \times \frac{LUCK\ VARIABILITY}{LUCK\ VARIABILITY + ABILITY\ VARIABILITY}$$

In a recent paper (to appear in *STATS*, a statistics magazine for students), I looked at several hitting statistics from this luck perspective. The variation in batting averages is heavily influenced by luck. In fact, if you have a collection of players each with 500 at-bats, then about 50% of the variation in the batting averages is due to luck, and 50% is due to difference in the players' hitting abilities. Other statistics, such as a player's strikeout rate (SO/AB), are much less affected by luck variation. If you collect the strikeout rates for many players with 500 at-bats, then it turns out that only 10% of the variation is due to luck.

I looked at many basic rate statistics from this viewpoint. Here is an ordering of the statistics from most ability-driven to most luck-driven. The table also gives a range of 90% of the abilities (not the performances) of current MLB nonpitchers for each statistic.

[1] *Baseball Research Journal 33* (2005)

	Rate Statistic	90% of "true" rates fall between
More Ability	Strikeout rate (SO/AB)	0.094 and 0.278
	In-play home run rate (HR/(AB-SO))	0.012 and 0.088
	Walk rate (BB/AB)	0.042 and 0.143
	OBP	0.278 and 0.384
	In-play AVG (H/(AB-SO))	0.288 and 0.365
	AVG	0.235 and 0.301
More Luck	In-play singles rate (1B/(AB-SO))	0.186 and 0.245

Does the Luck Percentage Depend on the Sample Size?

It is important to emphasize, as James does, that the luck in a statistic depends on the sample size. Batting averages for entire seasons are about 50% luck and 50% ability. But batting averages based on 100 at-bats are about 80% determined by luck. It is silly to draw any conclusions on how players perform after only one month of the season.

Luck in Comparison Offshoots

James later in the article talks about the luck dimension of what he calls comparison offshoots. As a simple illustration, suppose you collect, for each player, the AVG against pitchers of the opposite arm and the AVG against pitchers of the same arm. You compute the difference AVG (opposite arm) – AVG (same arm). This is what James calls a comparison offshoot. James talks about the luck aspect of these offshoots:

> "Suppose you take two statistics, each of which is 30% luck, and you add them together. The resulting new statistic will still be 30% luck. 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), the resulting new statistic – the comparison offshoot – may be as much as 60% luck."

To be honest, this is nonsense, since the luck in a comparison offshoot really depends on what comparison you are looking at. The luck in a comparison has nothing to do with the luck in the individual statistics that you are subtracting or adding.

Let's illustrate this point using batting averages. Suppose that there are no individual platoon tendencies with respect to pitchers of the same versus opposite arm. In other words, there is a general tendency for players to hit better against pitchers of the opposite arm, but this tendency is the same for all players. What will be the luck of a comparison offshoot defined by AVG (opposite arm) – AVG (same arm)? The answer is 100% since we have assumed that there is no individual platoon tendency.

This answer of 100% is independent of the luck involved in an individual's batting average. We know from above that an individual's batting average is roughly 50% luck (if a hitter has 500 at-bats), but we would get the same answer if an individual's AVG was 80% luck or if his AVG was 30% luck. Actually, by subtracting two AVGs for the same player, we have removed the luck aspect of an individual's AVG from the data. (This is a common way of reducing variation of a response variable in an experimental design.)

Do Individual Batters Have Individual Platoon Tendencies?

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*. So I thought I'd do a brief investigation of this issue.
I thought I would work with a statistic that is primarily ability-driven (instead of luck-driven), so I focused on strikeout rate (SO/AB). For the 2003 season, I collected (1) the number of AB and SO against pitchers of the opposite arm and (2) AB and SO against pitchers of the same arm for all players with at least 100 AB against both types of pitchers. (Obviously, switch-hitters were excluded since they always have the advantage against the pitcher.) Define the platoon effect as:

```
EFFECT = SO/AB (against same arm pitchers) - SO/AB (against opposite arm pitchers)
```

I wanted to address the question: Is there evidence to suggest that batters have different platoon effects?

My approach for answering this question is to (1) fit a simpler model that says that there are no individual platoon effects, (2) simulate data from this simpler model, and (3) check if our 2003 season data is consistent with the simulated data. My simpler model says that players do have different strikeout probabilities described by a normal curve with mean .173 and standard deviation .056. (I got these values by fitting the model to data from a previous year.) Also this model says that players do have a greater tendency to strike out against pitchers of the same arm, but the effect is a constant value of .033 for all players. (Again I estimated the size of this effect from the data.) I used this model to simulate a group of strikeout rates – afterwards I computed the spread (standard deviation) of the rates. After I repeated this simulation process 1000 times, I got a distribution of spreads of strikeout rate effects from my simpler model. When I compared the actual 2004 strikeout rate effects to my simulated distribution, what I found is that the 2004 data was not consistent with my model. There was more variation in the strikeout effects than I would predict if the simpler model was true. So it appears that there is some evidence in this case for individual platoon effects. I confirmed this conclusion by use of a different analysis – I collected the strikeout rate effects for a number of players for two consecutive seasons and I found a small, but significant, positive correlation value of .2.

On the Incorrectness of a Method to Show a Wide Range of Conclusions in Sabermetrics

This analysis brings me to the major point of James' article. He seems to imply that sabermetricians have been using the wrong method for years, but his clarification is just the familiar logic (at least familiar among statisticians) in performing a statistical test. Suppose that I found that the strikeout effects for the 2004 season were consistent with my simpler model. Does that prove that individual platoon effects for strikeout effects do not exist? No! In statistics, a model is basically a simple representation of the variation in baseball data. Rarely is a model "true". Generally, models are approximations of reality since real-life is too complicated to be described by a model. But models can be useful since they can make reasonably accurate predictions in a process like a baseball season.

When we assess the goodness of fit of a model, there are two possible conclusions: either there is significant evidence to reject the model or there is insufficient evidence. Saying there is not enough evidence to reject the model doesn't say the model is true, but it does say that we just can't provide evidence to say that it is false. Statisticians don't prove models are true – that's why we are careful to say things like "we have insufficient evidence to reject a model."

Let me illustrate this point with the hot-hand phenomenon, which is the focus of Chapter 5 of *Curve Ball*. Let's say that we wish to investigate if a team's sequence of wins and losses is streaky. We decide on an appropriate measure of streakiness, say the longest run of wins or losses, and decide that the team is streaky if the longest run is very unlikely if the pattern of wins and losses was just due to chance. What if we don't decide that the team is streaky – does it mean that the hot-hand phenomenon does not exist? No! What it simply means that we haven't found enough evidence to reject the chance hypothesis. My opinion about the hot-hand phenomenon (or the clutch-hitting phenomenon) is that it may indeed exist, but it is likely a small effect, and very difficult to pick up with a statistical test.

"The Fog May Be Many Times More Dense Than We Have Been Allowing For"

In closing, I really like the fog analogy used by James. In modern statistical analyses of baseball data, we are working more with situational data or statistics derived from play-by-play accounts of games. There might be interesting insights to be learned from these data. But as James says, the danger of looking at this situational data is that measures of performance are based on small sample sizes, and these data are extremely chancy or heavily influenced by luck variation. One can easily lose any significant information or signal in the fog of chance randomness that surrounds these data. Maybe this will mean that teams will rely on statisticians more in the future to learn from this situational data. I'm waiting for the Phillies to call.

I would like to thank Jay Bennett and Michael Schell who contributed helpful suggestions to the writing of this article.
Jim Albert, albert@bgnet.bgsu.edu ♦

Get Your Own Copy

If you're not a member of the Statistical Analysis Committee, you're probably reading a friend's copy of this issue of BTN, or perhaps you paid for a copy through the SABR office.

If that's the case, you might want to consider joining the Committee, which will get you an automatic subscription to BTN. There are no extra charges (besides the regular SABR membership fee) or obligations – just an interest in the statistical analysis of baseball.

The easiest way to join the committee is to visit http://members.sabr.org, click on "my SABR," then "committees and regionals," then "add new" committee. Add the Statistical Analysis Committee, and you're done. You will be informed when new issues are available for downloading from the internet.

If you would like more information, send an e-mail (preferably with your snail mail address for our records) to Neal Traven, at beisbol@alumni.pitt.edu. If you don't have internet access, we will send you BTN by mail; write to Neal at 4317 Dayton Ave. N. #201, Seattle, WA, 98103-7154.

Receive BTN by Internet Subscription

You can help save SABR some money, and me some time, by downloading your copy of *By the Numbers* from the web. BTN is posted to http://www.philbirnbaum.com in .PDF format, which will print to look exactly like the hard copy issue.

To read the .PDF document, you will need a copy of Adobe Acrobat Reader, which can be downloaded from www.adobe.com.

To get on the electronic subscription list, visit http://members.sabr.org, go to "My SABR," and join the Statistical Analysis Committee. You will then be notified via e-mail when the new issue is available for download.

If you don't have internet access, don't worry – you will always be entitled to receive BTN by mail, as usual.

Clutch Hitting and the Cramer Test

Phil Birnbaum

Bill James recently asserted that Dick Cramer's famous 1977 clutch-hitting study, which purportedly demonstrated that clutch talent is a myth, was fatally flawed. James argued that the study's finding that year-to-year clutch hitting looks random was not enough to show non-existence. Here, the author uses statistical methods to try to determine whether James' argument is indeed correct.

In the 1977 *Baseball Research Journal*, Dick Cramer published a now-classic study on clutch hitting. He looked at the best clutch hitters in 1969, and found that, on average, they reverted to normal in 1970. Cramer concluded that since the 1970 performances looked random, the lack of persistency showed that the 1969 performance was simply random luck, and, therefore, that clutch hitting as an ability does not exist.

Twenty-seven years later, in the same publication, Bill James disputed Cramer's conclusion. In his essay "Underestimating the Fog" (critiqued by Jim Albert elsewhere in this issue), James wrote:

> "… 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." [emphasis in original]

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.

But you have to look at the specifics of the study. For some studies, you'll find that the study has indeed "failed" – that even if a substantial effect existed, the study would still have found random data. But, contrary to Bill James' assertion, many studies will indeed be powerful enough to find a relationship if one exists – and for those, a finding of random data is powerful evidence of a non-effect.[1]

The Cramer Test

Cramer's study 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 clutch ability was not shown to exist.

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 this correct? If clutch hitting were a consistent skill, would the Cramer test have been powerful enough to pick it up?

To check, I repeated a variation of Cramer's study for the 1974-75 seasons. For each of the 137 players having at least 50 "clutch" at-bats both years,[2] I calculated his batting average difference between clutch and non-clutch for both seasons. I then ran a linear regression on the 1974 vs. 1975 data.

[1] There is a strict philosophical sense in which it can be said that no test can ever be *certain* that an effect exists or doesn't exist. That is, no matter how many data points in our sample, and no matter how evident it *appears* that there is an effect (or no effect), it is always possible, although perhaps vanishingly unlikely, that the observed differences (or lack thereof) were caused by chance. (For instance, it cannot be said for sure that Barry Bonds is a better hitter than Mario Mendoza, because there is an infinitesimally small probability that Mario is actually better, but just had really bad luck.) However, this interpretation does not appear to be what James had in mind.

[2] For "clutch," I used the Elias "Late Inning Pressure Situations" definition – 7th inning or later, tied or down by 3 runs or less, unless the bases are loaded, in which case down by 4 runs was included. Thanks, as always, to Retrosheet. Note that my data occasionally slightly differs from Elias, and perhaps other sources – my best guess is that Retrosheet's raw data differed from Elias's, causing Elias to count some at-bats as clutch which I didn't, or vice-versa. A full set of raw data is available from the author, or, at time of writing, at http://www.philbirnbaum.com/clutch.txt.

The results: a correlation coefficient (r) of .0155, for an r-squared of .0002. These are very low numbers; the probability of the 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 there's an effect.

So this study reaches the same preliminary conclusions as Cramer's – players with good (or poor) clutch hitting in 1974 showed no tendency to repeat in 1975. Standard statistical operating procedure would have us conclude that there's no evidence of a clutch effect.

But does this constitute good evidence that clutch hitting does not exist? We can't answer that question yet. It could be that the study simply isn't strong enough to find such an effect if it exists. If that's the case, then, as Bill James suggested, we'd get random data whether an effect existed or not, and the study could be said to be a failure.

But not a complete failure. What the study *does* clearly tell us is the relationship between a player's 1974 performance and his 1975 performance. It tells us that relationship is very weak.

We may not be able to say for sure, yet, that the statement "clutch hitting talent exists" is false. But the statement, "Because Wade Boggs hit 110 points better in the clutch in 1989, he must be a good clutch hitter" is definitely false, because we have shown that even if clutch hitting exists, there is no season-to-season consistency.

Specifically, the correlation coefficient of .0155 tells us that if a player was X standard deviations above average in 1974, he would be expected to be only 1.55% of X standard deviations above average in 1975. Roughly speaking, a player who was 110 points above average in 1974 would be *about 2 points above average* in 1975.

Yes, this study has not necessarily proven that clutch hitting doesn't exist. But it *does* show that if clutch hitting *did* exist in 1974-75, it was in very, very small quantities.

Other Years

I repeated this study for all pairs of years from 1974 to 1990 (excluding pairs involving 1981). See Table 1, below.

The results are pretty consistent: none of the pairs shows anything close to statistical significance, except one – 1978-1979. But that one shows a *negative* correlation; players who hit well in the clutch in 1978 hit *poorly* in the clutch in 1979. Since we have no reason to believe that clutch hitting one year turns you into a choker the next, that season is probably just random noise.

I have no explanation for why 1979-80 is so highly negative, or for why the correlation is positive in 12 of the 14 years. In any case, we do have confirmation of Cramer's findings in all fourteen of those seasons – in none of them did clutch performance this year significantly lead to clutch performance next year.

We still have not shown whether clutch performance exists or not. But we have shown that for each of the fourteen seasons in the study, a good clutch performance over a season's worth of at-bats does *not* reliably indicate a good clutch hitter.[3]

Table 1 – Clutch hitting correlation between pairs of consecutive seasons

	r	r-squared	P(f statistic)
74-75	.0155	.0002	.86
75-76	.0740	.0055	.37
76-77	.0712	.0051	.40
77-78	.0629	.0040	.44
78-79	-.1840	.0339	.02
79-80	.0038	.0000	.96
82-83	-.0250	.0006	.75
83-84	.0456	.0021	.60
84-85	.0222	.0005	.79
85-86	.0728	.0053	.38
86-87	.0189	.0004	.82
87-88	.0034	.0000	.97
88-89	.0829	.0069	.33
89-90	.0373	.0014	.67

[3] And by extension, if a full season's worth of clutch hitting doesn't identify a good clutch hitter, neither can a good clutch post-season (e.g., Reggie Jackson).

Power of the Single-Year Cramer Test

Back to the question – if clutch hitting *did* exist, would this study have been able to find it? Bill James seemed to imply that the question could not be answered, and we therefore must assume the answer is no. But we *can* answer it.

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).

Statistically, that would mean that, of the 137 players,

- 47 players, or about 34%, would be clutch hitters of 0-30 points.
- 47 players, or about 34%, would be choke hitters of 0-30 points.
- 19 players, or about 14%, would be clutch hitters of 30-60 points.
- 19 players, or about 14%, would be choke hitters of 30-60 points.
- 3 players, or about 2%, would be clutch hitters of 60-75 points.
- 3 players, or about 2%, would be choke hitters of 60-75 points.
- One half player would be a clutch hitter of 75 points or more.
- One half player would be a choke hitter of 75 points or more.

(This adds up to 139, instead of 137, because of rounding.)

If that were the case, we'd probably agree that clutch hitting is a reasonably important part of the game. But, still, two-thirds of all players are only 30 points different in the clutch, and only one regular player in the league has a 75 point difference (and it might go either way).

We'd probably say that clutch hitting is reasonably significant -- something for the player's manager to keep in mind, along the lines of a platoon differential, for instance.

So, back to the original question: would the Cramer test be able to pick up this distribution of clutch talent?

To check, I took the 137 players in the 1974-75 sample, and assigned each of them a "clutch talent" based on this normal distribution.[4] Then, I threw away their real-life "hits" columns, and simulated their actual number of clutch and non-clutch hits based on their clutch and non-clutch expected averages.

I repeated this test 14 times, for easy comparison to the 14 years of data shown above, in table 1.

The results: in 11 of the 14 cases, the Cramer test had absolutely no trouble picking up the clutch hitting at the .05 significance level:

r	r-squared	P(f statistic)
.2385	.056	.005
.1872	.035	.03
.3589	.128	.00
.2196	.048	.01
.0889	.007	.30
.0646	.004	.45
.3006	.09	.0004
.3928	.15	.00
.2945	.086	.0005
.2711	.073	.0014
.1425	.02	.10
.1826	.033	.03
.3321	.11	.0001
.2715	.073	.0013

[4] Technical note: to generate a clutch difference from this normal distribution, I created a "fake" .250 hitter. I ran 208 random at-bats for him, giving him a 1 in 4 chance of getting a hit each at-bat. I then computed his batting average over those 208 AB, and subtracted it from .250. According to the normal approximation to the binomial distribution, that random BA should be normally distributed with mean .000 and standard deviation .030. So that BA difference for the "fake" player becomes a clutch talent for one of the simulated players.

To make things clearer, here are the probabilities (third column of the table) summarized in table form. The top row is for the real-life data (from Table 1); the bottom row is for our simulation (table above). Statistically significant samples (.05 or lower) are in bold, and "Neg" indicates a negative correlation:

.86	.37	.40	.44	Neg	.96	Neg	.60	.79	.38	.82	.97	.33	.67
.01	**.03**	**.00**	**.01**	.30	.45	**.00**	**.00**	**.00**	**.00**	.10	**.03**	**.00**	**.00**

There is no question that the results in the second row are highly significant, and the results in the first row are largely random. For a clutch SD of 30 points, the Cramer test finds the clutch hitting. Since it did not, we can be almost certain that if clutch hitting talent does exist, its standard deviation must be less than 30 points.

Let's repeat the experiment, but this time with an SD of 20 points and 28 simulated seasons instead of 14. I'll leave out the full results, and go right to the table of significance levels:

.86	.37	.40	.44	Neg	.96	Neg	.60	.79	.38	.82	.97	.33	.67
.41	**.04**	Neg	**.00**	.26	.10	**.00**	**.00**	**.03**	**.047**	.41	.11	**.00**	.54
.02	Neg	.33	**.00**	.46	.15	**.02**	.19	**.02**	.16	**.03**	.41	.47	.55

Overall, 12 of the 28 simulated runs found statistical significance, many of the others were low, .2 or less. Again, there's no comparison between the real-life data and the simulation, and we can again conclude that if clutch hitting existed at the 20-point SD level, the Cramer test would have found it.

So, let's now try an SD 15 points, or .015. This is half what we started with, which would mean that only one player in both leagues would be as much as 45 points better (or worse) in the clutch.

Since the significance levels will start to vary more, I'll run more of them. Here are four rows of simulated data, again with the real-life data in the top row:

.86	.37	.40	.44	Neg	.96	Neg	.60	.79	.38	.82	.97	.33	.67
.18	**.02**	.08	.91	.14	.11	**.04**	**.03**	Neg	Neg	**.01**	.41	.22	.41
Neg	Neg	.25	Neg	.57	**.03**	.21	.18	.53	.12	**.046**	Neg	.75	.17
.09	.09	**.01**	.09	.27	.82	.69	**.03**	.26	.74	.90	Neg	**.008**	**.02**
.98	.18	Neg	Neg	Neg	.76	.38	Neg	.67	.46	.12	.13	**.002**	Neg

We're down to only 11 significant simulations out of 56. But there are still a few close calls and the numbers lean towards the smaller, which suggests that the test is picking up a real effect. Still, the test seems to have gotten significantly weaker.

Let's go down another step to .010:

.86	.37	.40	.44	Neg	.96	Neg	.60	.79	.38	.82	.97	.33	.67
.10	.08	.09	.42	.10	.35	Neg	.44	.37	.65	Neg	Neg	.21	Neg
.81	.52	.99	.87	**.02**	.14	.26	.35	Neg	Neg	Neg	Neg	.29	.90
.80	.75	.27	.09	Neg	.68	Neg	.10	**.001**	.61	Neg	Neg	Neg	**.02**
.08	Neg	.70	**.04**	.78	.26	.52	.34	Neg	**.002**	.31	.60	.95	Neg

There are five significant findings out of 56 – about 9%, whereas 2.5% would be expected by chance – but the numbers are getting bigger. The real-life row of data still doesn't quite fit in, but it's getting pretty close. We certainly can't say that an SD of .010 is out of the question, but it does still look a bit doubtful.

To make sure, we'll go down one more step to .0075:

.86	.37	.40	.44	Neg	.96	Neg	.60	.79	.38	.82	.97	.33	.67
.35	.19	.99	Neg	.62	.99	Neg	.10	**.01**	**.006**	.60	Neg	.41	.38
Neg	Neg	Neg	Neg	Neg	.59	Neg	.44	.68	Neg	Neg	.66	.52	.34
Neg	.22	.71	Neg	Neg	.77	Neg	.51	.71	.36	.27	Neg	Neg	.34
.06	.79	Neg	.61	.68	Neg	Neg	.46	.48	.20	.88	Neg	Neg	.95

Although there are still a couple of significant results, our row of real-life data fits right in among the other rows. It's fair to say that the Cramer test does indeed "fail" when the standard deviation of clutch hitting is as low as .0075.

A Formal Statistical Power Test

Instead of by using the simulation, we can produce an actual measurement of the Cramer test's power by using a freeware software tool called GPOWER[5].

Suppose clutch hitting existed, but with a real-life correlation of .25. That would mean that roughly speaking, a player who hit 110 points above average in the clutch this year would hit about 27 points in the clutch higher next year. What are the chances the 1974-75 Cramer test would find a statistically significant result? GPOWER's answer: 91.3%. That means, on average, we should have found 13 significant real-life seasons out of the 14. We actually found zero.

With a correlation of .2, the Cramer test would succeed 77% of the time – 11 out of 14 instead of zero. At r=.15, we'd find significance 55% of the time – 6 out of 14. At .1, the success rate is still 32%, or 4 out of 14. Finally, at .05, the Cramer test would succeed only 14% of the time, which is still 2 out of 14.

It's fair to say that at the most optimistic, season-to-season clutch hitting might exist with a correlation of .1 or less. That means that a player with 110 points of clutch hitting this year would show only 11 points next year. And again, that's being the most optimistic.

The Multi-Year Cramer Test

If we take all 14 years of the Cramer Test, and put them in one large regression, that would save us having to look at fourteen different regressions. It might also find a real effect where it couldn't before, since 14 seemingly random results often to combine to give a very significant one. For instance, when a player goes 2-for-5, he might just be an average player having a good day; but when he goes 2-for-5 over a whole season, he's Ted Williams.

In the 14 seasons in the study, there were a total of 2,057 season-to-season comparisons. Running a regression on those seasons gives:

	r	r-squared	P(f statistic)
14 seasons	.0142	.0002	.52

This is, decidedly, an insignificant result. If clutch hitting were truly and completely random, on average we would get a significance level of .50 – and we got a very close .52. This is very far from the .05 we would want to see to conclude clutch hitting is real.

What about the simulation? Let's start with a standard deviation of 15 points, where the one-season Cramer test was successful only about 20% of the time. I ran 15 copies of each 1974 player, for a total of 2,055 player-seasons:

	r	r-squared	P(f statistic)
14 simulated seasons	.0529	.002	.02

The simulation turns out significant at the 2% level. But it could be a fluke – let's run it a few more times, and compare it to the real life .52 (not shaded):

.52	**.02**	**.004**	**.005**	**.003**	.06	**.00**	**.003**	**.005**	**.01**
.12	**.03**	**.007**	**.02**	**.01**	.12	**.00**	**.01**	**.01**	**.01**

The 14-seasons-combined Cramer test seems to find clutch hitting at a standard deviation of .015 almost all the time. And, in fact, our .52 "real-life" significance level is very much out of place in this list.

[5] GPOWER is available from *http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/*. Thanks to Charlie Pavitt for letting me know about GPOWER, as well as for several suggestions that significantly improved this paper.

Let's try .010:

.52	.82	.11	.60	.49	**.002**	.27	.65	.10	.30
.32	.26	.14	.18	**.01**	.24	.07	.08	**.005**	.11

Three of the 19 simulations came up with significant results; several others were very close; and most of the numbers are low. However, there were a few higher numbers, and the real-life .52 isn't completely out of place here. We can say, then, that the 14-season-combined Cramer test begins to "fail" at a standard deviation of about 10 points.

GPOWER again

Running our complete 14-season study (2,055 pairs) through GPOWER shows:

If there were a correlation of .15 or higher, the 14-season Cramer test would have found it 100% of the time (or at least more often than 99.9995%).

If the correlation were .10 or higher, we would have found it 99.8% of the time.
If the correlation were .08 or higher, we would have found it 97.7% of the time.
If the correlation were .05 or higher, we would have found it 73% of the time.

Finally, if the correlation were .05, what are the chances we would find a significance level of .52 or better? *99%*. That is, if there was indeed clutch hitting at a 5% correlation – which means 110 points this season translates into only 5.5 points the next – 99 times out of 100 we would have seen more significant results than we did.

Conclusions

We can draw two sets of conclusions from all this: one set about the Cramer test, and one set about clutch hitting.

- The Cramer test provides important evidence on the clutch-hitting question. While it cannot completely disprove the existence of clutch hitting, it puts a strong lower bound on the magnitude of the possible effect.

- If clutch hitting talent is distributed among players with standard deviation of at least .030, the single-season Cramer test (with 137 player-pairs) is sufficiently powerful to find the effect. Between .020 and .010, the single-season test might find the effect, but probably will not. Below .010, the single-season test will not find the effect.

- Using the Cramer test 14 separate times on 14 separate seasons, we should note an obviously above-average number of significant seasons down to an SD of .015. At .010, we will notice more significant seasons than expected, perhaps 9% instead of 2.5%. But at .0075, the Cramer test will almost certainly fail to find the effect.

- By using the Cramer test on one large sample of 14 seasons combined, we are almost certain to notice an effect all the way down to an SD of .015. But at .010, a finding of significance is unlikely.

- Finally, if the correlation of the large sample is .08 or higher, the Cramer test on the combined data should spot the correlation almost 98% of the time.

As for clutch hitting itself:

- The Cramer tests show that clutch hitting appears to not exist at all. The question, though, is whether clutch hitting is still *possible* given the limitations of the Cramer Test.

- If clutch hitting exists despite the Cramer tests failing to find it, its standard deviation among players is almost certainly less than 15 points of batting average; that is, at least two-thirds of all players are expected to clutch hit within .015 of their non-clutch batting average. It is likely even less than 10 points of batting average.

- If clutch hitting does exist at the 10-point level, that means that less than one player in 200 can hit more than 25 points better in the clutch – a very small effect, even in the best case.

- If clutch hitting does exist with a between-seasons correlation of .05, there would have been about a 99% chance that the Cramer test would have found more significance than it did.

- If clutch hitting does exist with a between-seasons correlation of .05, that means that a player who hits 110 points better in the clutch one year can be expected to hit only 5.5 points better in the clutch next year.

- The results confirm other studies that failed to find clutch hitting, including Pete Palmer's multi-year study (BTN, March, 1990).

- In summary – the Cramer test doesn't prove that clutch hitting does not exist. But it does prove that if clutch hitting does exist, for the vast majority of players, it's at levels almost too small to be detectable or important, and previous clutch-hitting performance is not a reliable predictor of future clutch-hitting performance. This suggests that in any case, it is probably impossible to distinguish good clutch hitters from bad.

Phil Birnbaum, birnbaum@sympatico.ca ♦

Informal Peer Review

The following committee members have volunteered to be contacted by other members for informal peer review of articles.

Please contact any of our volunteers on an as-needed basis - that is, if you want someone to look over your manuscript in advance, these people are willing. Of course, I'll be doing a bit of that too, but, as much as I'd like to, I don't have time to contact every contributor with detailed comments on their work. (I will get back to you on more serious issues, like if I don't understand part of your method or results.)

If you'd like to be added to the list, send your name, e-mail address, and areas of expertise (don't worry if you don't have any - I certainly don't), and you'll see your name in print next issue.

Expertise in "Statistics" below means "real" statistics, as opposed to baseball statistics - confidence intervals, testing, sampling, and so on.

Member	E-mail	Expertise
Ben Baumer	bbaumer@nymets.com	Statistics
Jim Box	jim.box@duke.edu	Statistics
Keith Carlson	kcsqrd@charter.net	General
Dan Evans	devans@seattlemariners.com	General
Rob Fabrizzio	rfabrizzio@bigfoot.com	Statistics
Larry Grasso	l.grasso@juno.com	Statistics
Tom Hanrahan	Han60Man@aol.com	Statistics
John Heer	jheer@walterhav.com	Proofreading
Dan Heisman	danheisman@comcast.net	General
Bill Johnson	firebee02@hotmail.com	Statistics
David Kaplan	dkaplan@UDel.Edu	Statistics (regression)
Keith Karcher	karcherk@earthlink.net	Statistics
Chris Leach	chrisleach@yahoo.com	General
Chris Long	clong@padres.com	Statistics
John Matthew IV	john.matthew@rogers.com	Apostrophes
Nicholas Miceli	nsmiceli@yahoo.com	Statistics
Duke Rankin	RankinD@montevallo.edu	Statistics
John Stryker	johns@mcfeely.interaccess.com	General
Tom Thress	TomThress@aol.com	Statistics (regression)
Joel Tscherne	Joel@tscherne.org	General
Dick Unruh	runruhjr@dtgnet.com	Proofreading
Steve Wang	scwang@fas.harvard.edu	Statistics

Beyond Player Wins – Calculating Individual Player Pennants Added

Dan Levitt

Most sabermetric studies attempt to quantify player performance in terms of wins However, it's not really wins that teams are looking for, but, rather, pennants. What is the relationship between the number of wins a player contributes, and the number of pennants? And does it depend on whether the wins are evenly distributed among many seasons, or concentrated among few?

When attempting to reduce a player's value to one number, what is it we really want to know? I think most of us presume we are trying to discover the player's contribution to winning. Which leads to the question, what do we mean by winning? Not surprisingly, baseball analysts typically denominate player value in one of two metrics: runs or wins. And I believe that the ability to express ballplayer value in these two measures marks one of the most important (if not the single most important) breakthroughs by sabermetrics.

But one could further argue that the real goal of the regular season is not simply wins for wins' sake, but to capture the pennant (or since 1969, to qualify for the postseason). For the evaluation of players over multiple seasons, this leads to the question of whether all player "wins" are created equal. In other words, are five pretty good seasons of, say, two wins added above what an average player would have generated (WAAv), more or less or of equal value to two excellent seasons of five wins above average?[1]

From 1972 through 1974, Johnny Bench created 13.4 WAAv, an average of about 4.5 per year. Over the next seven years, from 1975 through 1981, Bench created a similar cumulative total of 12.9 WAAv but only around 1.8 per season. On a collective basis are Bench's three strong seasons approximately equivalent to the seven more ordinary seasons at the end of his career in terms of helping his team reach the postseason? That is, if one adds a 4.5 WAAv season to a randomly selected team, does he increase that teams chances of winning the pennant 2.5 times as much (4.5/1.8) as adding a 1.8 WAAv player?

In terms of winning pennants, Bill James argued that, in fact, a few excellent seasons interspersed with a few poor ones are actually more valuable than an equivalent number of more commonplace seasons (see *The Politics of Glory*, pp. 81-87). In the 2002 Baseball Prospectus (pp. 470-475), Michael Wolverton took another look at the question; he too concluded that a few excellent seasons added more pennants than a number of mediocre ones, even if both added an equivalent number of wins. Wolverton calculated what he called Pennants Added, the player's value in expected pennants added less "the probability that his team would have made the playoffs without him."

I decided to relook at calculating pennants added using a slightly different approach than Wolverton's. Later in this essay I outline my method in detail, but a summary may be appropriate here.

I first calculate the player's runs created and outs used. The player is then effectively "placed" on all possible teams and the runs scored by each of those teams recalculated by adding the impact of the subject player's runs created to the team total. James' Pythagorean formula is then used to determine the winning percentage of each possible team, now with the player added in. Based on that winning percentage, the probability of each possible team reaching the postseason is calculated. The pennants added by the subject player equals the sum of the probability of each possible team reaching the pennant multiplied by the probability of that particular runs scored and allowed pair. Finally, the pennants added by an average player is subtracted from the player's pennants added to generate Pennants Added Above Average (PAAAv). That is, pennants added above average effectively equals the increased (or decreased) probability of a randomly selected team reaching the post season when substituting the subject player for a typical player on that randomly selected team.

My research supports the findings of James and Wolverton. For example, Bench's three exceptional seasons of 13.4 WAAv resulted in .234 pennants added beyond what an average player would have produced (his 1972 year of 6.7 WAAv correlates to increasing the Reds post season probability by .139). Bench's subsequent seven seasons which resulted in a similar WAAv of 12.9 resulted in only .184 PAAAv.

Why are extraordinary seasons worth more than an equivalent number (in terms of wins added above average) of less exceptional ones? Assume that a team needs 93 wins to make the postseason. A player with, say, three WAAv will put an otherwise 90 win team over the top,

[1] "Wins Above Average" is used as a comparison throughout this paper. The WAAv calculation is similar to other wins above average metrics such as "Adjusted Batting Wins" used in *The Baseball Encyclopedia* and *Total Baseball*. The specific methodology is detailed in *Paths to Glory*, which I co-authored with Mark Armour.

while a player with six WAAv will put an otherwise 87-win team into the postseason. The key point is that there are more teams that win around 87 games than 91. From 1901 through 2002 (excluding 1981 and 1994) 132 teams won the equivalent of between 90 and 92 games (based on a 162 game season) and 171 won between 87 and 89. Thus, for illustrative purposes only (because the math doesn't quite work out like this), a player who records two seasons of three WAAv could be seen as propelling 264 (2 x 132) teams into the postseason, while a player who has a single season at six WAAv moves 304 (132 + 171) teams into the postseason.

Table 1 lists the top 25 players in career PAAAv through 2002. Note that because of the computational intensity of the calculation, the list is limited to those players elected to the Hall of Fame through 2002.

Ordering by PAAAv changes the overall ranking slightly but does not materially affect our understanding of the top players. The nineteenth century players appear slightly higher because they accumulated their WAAv in a shorter season; thus each player win above average is slightly more valuable relative to winning the pennant. Several twentieth century players (not shown) who end up higher in the rankings using PAAAv (as opposed to WAAv) include Ross Youngs, Frank Chance and Home Run Baker. Three who fall a little under this metric include Billy Williams, George Brett and Bill Dickey.

To quantify the relationship between wins and pennants added, the final column in Table 1 presents the wins above average necessary to add one pennant above average. Because they had fewer games over which to spread their WAAv, nineteenth century players typically generate one pennant per approximately 50 WAAv or less. More recent players may require nearly 60 WAAv to produce one additional pennant above average. More interesting are the differences among players of the same general era: Lou Gehrig created one additional pennant for every 52 wins above average, while Jimmie Foxx needed 59 WAAv to produce one additional pennant.

The goal of any team over the regular season is to win the pennant, or since 1969, to qualify for the postseason. The relationship between player wins above average and the value of his production in terms of winning the pennant is not linear. A fewer number of peak seasons is more valuable than a larger number of ordinary seasons even if both sets of seasons total an equivalent number of wins added above average. Unfortunately the calculation is quite computationally demanding making it questionable whether the additional information offered by PAAAv is worth the effort.

Table 1--Top 25 Hall of Famers Ranked by PAAAv

Name	PAAAv	WAAv	W/PA
Ruth, Babe	3.056	153.9	50.4
Cobb, Ty	2.699	132.2	49.0
Mantle, Mickey	2.270	118.2	52.1
Gehrig, Lou	2.176	116.3	53.4
Aaron, Hank	2.093	115.7	55.3
Williams, Ted	2.088	115.4	55.3
Hornsby, Rogers	2.018	101.9	50.5
Wagner, Honus	1.989	96.5	48.5
Mays, Willie	1.973	110.7	56.1
Musial, Stan	1.855	104.7	56.4
Speaker, Tris	1.844	97.3	52.8
Ott, Mel	1.750	97.1	55.5
Collins, Eddie	1.716	88.0	51.2
Robinson, Frank	1.532	89.3	58.3
Lajoie, Nap	1.518	75.3	49.6
Brouthers, Dan	1.497	71.3	47.6
Mathews, Eddie	1.461	80.3	55.0
Connor, Roger	1.398	65.7	47.0
Morgan, Joe	1.394	80.5	57.7
Foxx, Jimmie	1.314	77.8	59.2
Delahanty, Ed	1.273	66.0	51.8
Crawford, Sam	1.240	63.3	51.1
DiMaggio, Joe	1.234	70.8	57.4
McCovey, Willie	1.226	72.1	58.8
Burkett, Jesse	1.152	61.2	53.1

For an analysis of only more recent seasons one can put together a shortcut formula to simplify the calculation. Figure 1 shows the relationship between WAAv and PAAAv for the three years 2000 - 2002. The formula on the graph provides a direct approximation formula to convert from WAAv to PAAAv.

Based on the formula, two 5 WAAv seasons are 20% more valuable than five 2 WAAv seasons, and one ten WAAv season is in turn 19% more valuable than two 5 WAAv seasons. In sum, a superstar-type season improves a team's probability of reaching the postseason beyond what a typical sabermetric win valuation metric suggests.

In most analytical circumstances, however, moving beyond wins added above average to pennants added above average may not be worth the added computational effort; moreover using wins as a metric is more intuitive than pennants. None the less, it is instructive to verify that peak seasons are in fact more valuable as opposed to a larger number of statistically equivalent seasons.

The Methodology

In summary, the methodology involves the following steps.

Step 1

For each potential number of runs scored by a team in the subject year, back out the number of runs scored in the subject player's number of outs (using runs per out for the subject team); then add back the subject players runs created.

Step 2

Select each possible number of runs allowed by a team in the subject year.

Step 3

Using James' Pythagorean theorem (I use an exponent of 1.86), calculate the winning percentage of a team for each combination of runs scored and allowed in Steps 1 and 2, above.

Step 4

Calculate the probability of reaching the postseason for a team with the Pythagorean winning percentage generated in Step 3. For this I use the formula:

```
PrPost = 44.596 - 235.68 * WLPct +
406.18 * (WLPct ^ 2) - 226.35 *
(WLPct ^ 3)
```

which is derived from all seasons 1901 through 2002 (exclusive of 1981 and 1994). I arbitrarily did not include the nineteenth century because of the wide variation in team and league quality. In any case, the results would not have differed materially. There is probably a more elegant formula for the likelihood of reaching the postseason, but this formula fit well, and moreover it is buried in the overall calculation negating any benefit from a more elegant one.

Aside on the Probability of Reaching the Postseason Based on Runs Scored and Allowed

I examined the probability of a team qualifying for the postseason based on both its actual winning percentage and its Pythagorean winning percentage (a winning percentage estimate based on runs scored and allowed). Because estimating winning percentages based on runs scored and allowed is quite accurate, the two probabilities are fairly similar as can be seen in Figure 2.

However, there does exist a difference, and the direction of this difference shifts at a winning

Figure 1

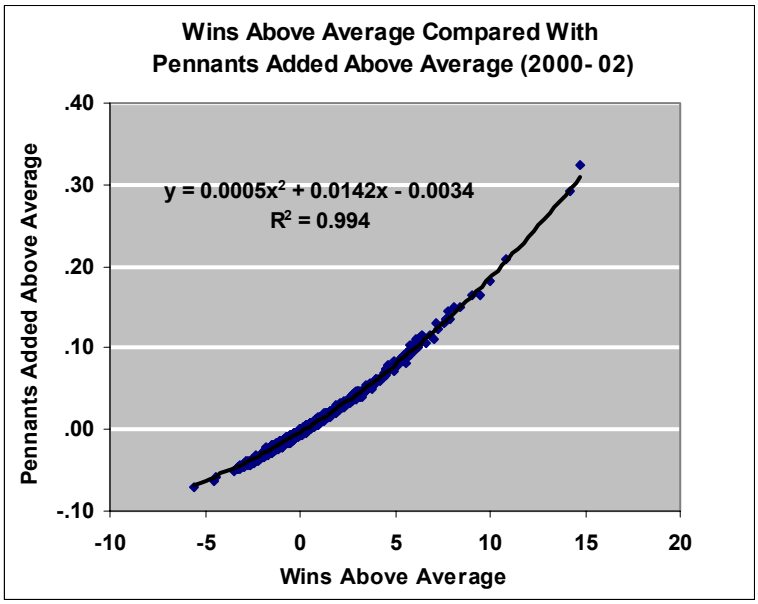
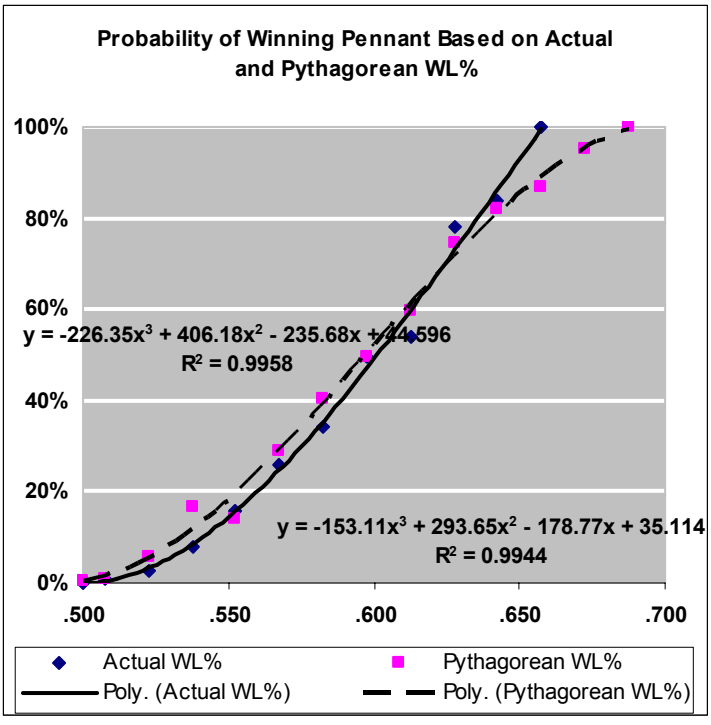


Figure 2



percentage of approximately .620. For equivalent actual and Pythagorean winning percentages below about .620, the team with the Pythagorean winning percentage is more likely to make the postseason than the team with the actual winning percentage. Above, the opposite holds; a team as measured by its actual winning percentage is more likely to qualify for the postseason than one considered by its Pythagorean winning percentage.

Upon reflection, this is not particularly surprising. The Pythagorean winning percentage is an estimate of actual winning percentage based on runs scored and allowed. For a number of reasons, including simply random luck, teams will often either outperform or underperform their estimate. At relatively lower winning percentages where a team typically may not be eligible for the postseason, a team that outperforms its Pythagorean estimate may do so by enough to qualify for the postseason. The reverse occurs at relatively higher winning percentages. Here a team will likely make the postseason, but a team that underperforms its Pythagorean estimate might not. At these relatively higher winning percentages, outperforming its estimate is not as significant given that the club will likely qualify for the postseason in any case.

I have summarized below how I derived the regression equations shown on Figure 1 to estimate the probability of qualifying for the post season based a particular winning percentage.

I looked at the record of all teams between 1901 and 2002 (excepting 1981 and 1994) and whether they made the postseason. I left the nineteenth century out of the derivation because of the generally unsystematic procurement of players which lead to a wide disparity of talent, and the oft-changing number of leagues and teams. I split team winning percentages into bins of .015 and counted the number of teams in each bin and how many qualified for the post season. I ran the regression on the midpoint of the bins as the x-value and the percentage of teams making the post season as the y-value. Table 2 summarizes the data.

Up until winning percentages around .650, the data shows a smooth upward progression of the likelihood of making the postseason. At that point, however, the data becomes less regular, owing mainly to the smaller sample sizes and the fact that the probability of qualifying for the post season based on a given winning percentage changed over time. Therefore, I had to make some manual adjustments to the data to reflect an upper limit. The percentages in italics indicate my manual editing of the data to provide a systematic upper limit.

Table 2 – Probability of Qualifying for the Postseason

WL%	Act Post%	Act Count	Act Post	Pytha Post%	Pytha Count	Pytha Post
.5000	0.0%	971	0	0.3%	950	3
.5075	0.9%	111	1	0.8%	127	1
.5225	2.6%	117	3	5.8%	139	8
.5375	8.1%	148	12	16.6%	145	24
.5525	15.8%	120	19	14.2%	141	20
.5675	25.9%	112	29	29.1%	117	34
.5825	34.2%	76	26	40.2%	82	33
.5975	49.5%	111	55	49.4%	77	38
.6125	54.1%	61	33	59.7%	62	37
.6275	78.3%	46	36	74.4%	43	32
.6425	83.8%	37	31	81.8%	33	27
.6575	*100.0%*	15	15	*87.0%*	19	15
.6725	80.0%	15	12	*95.0%*	13	10
.6875	100.0%	12	12	*100.0%*	4	3
.7025	100.0%	4	4	100.0%	5	5
.7175	100.0%	4	4	50.0%	2	1
.7325	NA	0	0	100.0%	1	1
.7475	100.0%	1	1	100.0%	1	1
.7625	100.0%	1	1	100.0%	1	1

Step 5
Multiply the probability of reaching the pennant for each runs and allowed combination by the probability of that runs scored and allowed combination.

Step 6
Sum all the results in Step 5 which provides the overall probability of a player placed on a randomly selected team of reaching the pennant.

Step 7
Subtract the probability of an average team from reaching the postseason.

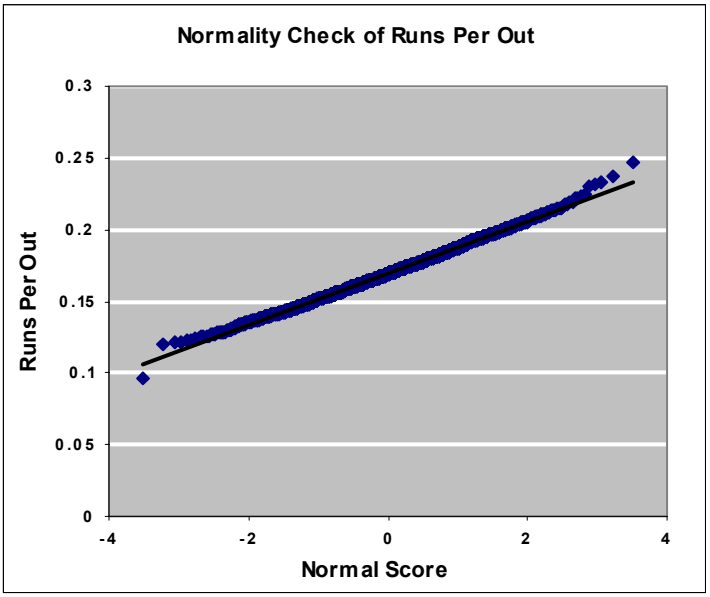
The seven-step calculation process described above further requires an explanation of several nuances and assumptions:

To select the range of a team's runs scored involves several steps. First of all one needs to calculate the league average and standard deviation of runs per out based on that season's team runs per out averages. One then needs the cumulative probability of reaching the postseason based on all possible runs scored (after substituting the subject player's statistics) and allowed. This would be an extremely computationally demanding calculation. There may be some closed form for this calculation, but I essentially brute force it while making some approximations.

To simplify calculating each individual runs scored and allowed possibility, I break the runs scored and allowed into 10 equally likely bins and take the midpoint number of runs per out in each bin. Then I sum up the likelihood of reaching the postseason over the 100 (10 x 10) equally likely different runs scored and allowed scenarios.

A couple of nuances need to be noted. First, I needed to confirm that runs scored were in fact approximately normally distributed. As Figure 3 indicates, except at the very extremes, runs scored per out can be considered normally distributed.

Figure 3



Second, in general, higher-scoring seasons often exhibited larger standard deviations of runs scored; the correlation between average runs per out and the standard deviation of runs per out for each league season is .35. Therefore, using a simple average of all the annual standard deviations to calculate the probability of different potential number of runs scored would have biased the data. That is, for high or low scoring years, the dispersion of runs scored would have been different than generally reflected by the data. To adjust for this range of standard deviations, the standard deviation used in the calculation is approximated from the average runs per out (derived from OLS regression to generate the formula):

```
RpOStd = -0.00665 + RpOAvg * 0.15098
```

One more adjustment needs to be made to the standard deviation. If the probability of winning a pennant formula is going to work, an average player should have a postseason probability equal to the historical average of the percentage of teams that actually qualify for the postseason. From 1901 through 2002 (again excluding 1981 and 1994) 15.0 percent of teams made the postseason. Using the above formula to derive the standard deviation for use in the pennants added function gives the probability of a league average player reaching the postseason materially less than .150.

The cause of the unsuitability of using the historical average of the runs scored standard deviation is quite interesting. Runs scored and allowed are not independent; it turns out they are inversely correlated. That is, teams that score a lot of runs are also more likely to give up fewer. After normalizing for league averages, the correlation between runs scored and allowed per out is -.27. On reflection, this should not be particularly surprising. A well-capitalized and/or intelligent baseball organization would likely develop both a quality offense and defense. Conversely, a more poorly run baseball organization will likely be weaker on both offense and defense. This means, for example, that when a team has a good offense, if one independently matches up a random defense, it will have a tendency to be worse than what shows up in practice. Thus, such a team would show up with a lower pennant probability than its historical information would suggest.

The impact on a pennants added formula of this inverse correlation is that one cannot simply use the overall calculated standard deviation on runs scored and allowed per out to generate a random team's probability of winning a pennant. To address this I edited the above standard deviation formula so that calculating the pennants added by an average player equaled .15 (i.e. for an average player the PAAAv = 0). The revised standard deviation formula is:

```
RpOStd = -0.004 + RpOAvg * 0.15098
```

The average runs scored per out over the history of baseball is approximately .17. Plugging this average into the pre-adjusted standard deviation formula gives a standard deviation of .019; after adjusting the formula, the standard deviation is .022 when using an average runs per out .17. The increased standard deviation is needed in order to correct for the fact that runs scored and allowed are not independent.

Excel Code

For those with some Excel macro experience, it might be simpler to follow the methodology through the code for the function itself. Additionally, some might wish to apply the function in their own analysis.

Readers interested in the code can obtain it from the author, or, at time of writing, from http:/www.philbirnbaum.com/paaavCode.txt .

Thanks to Phil Birnbaum for a couple of methodological suggestions. *Dan Levitt, danrl@attglobal.net* ♦