By the Numbers

Volume 13, Number 2 The Newsletter of the SABR Statistical Analysis Committee May, 2003

Summary

Academic Research: Clutch Hitting and Sudden Slugging

Charlie Pavitt

The author catalogs two more recent studies from Chance magazine. First, another investigation into clutch hitting, and, second, an examination of the probability that large jumps in a player's home runs would arise simply by chance.

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.

Jim Albert, Hitting with Runners in Scoring Position, Chance, Volume 15 Number 4, Fall 2002, pages 8-16

Jim Albert has already explored this area in an important article in the *Journal of the American Statistical Association* back in 1994 (summarized in *Curve Ball*), showing little support for across-the-board differences in several of the offensive "breakdown statistics" listed in the STATS *Player Profiles* series publications. In this study, Albert used 1987 National League raw data compiled by Project Scoresheet to address one of these breakdowns; offensive performance with baserunners in scoring position (vs. not in scoring position). For that year, Albert calculated expected runs scored for the 24 base-out situations, computed the difference in expected runs between contiguous base-out situations (Gary Skoog's still-underutilized "value added" approach; see the 1987 *Abstract*), and then computed the difference in individual players' "value added" offensive performance between plate appearances with and without runners in scoring position. As we almost always find in these sorts of studies, the distribution of the differences was largely random. This is a nice new method leading to more evidence that most breakdown stats have little value.

I would be remiss, however, by not mentioning a potential problem in Albert's analysis. Albert computes the change in run potential between plate appearances. But base-out situations change during at bats due to stolen bases, caught stealing, wild pitches, passed balls, and balks. I expect that the former two mostly cancel one another out in the long run and the latter three may not amount to much, but the impact of events during at bats would throw Albert's value added numbers off a bit. Jim, I trust that you will read this and I invite your response.

In this issue

Scott M. Berry, A Juiced Analysis, Chance, Volume 15 Number 4, Fall 2002, pages 50-53

This is an example of a fun and useful column Berry writes in each issue of *Chance* entitled "A Statistician Reads the Sports Pages," which features a baseball subject once or twice a year. Responding to the idea that steroid use could be responsible for sudden improvements in a batter's home run production, Berry computed the odds of a player's home run percentage in a

specific year given that player's home run percentage the previous three seasons of his career (1000 at bat minimum), weighing (along the lines of Bill James's career assessment method) the most previous season the most and first of the three seasons the least in the computation. The most abrupt increase was Kirby Puckett's 31 in 1986, after 4 combined in his previous two seasons, with a probability of 2 in 100 million.

Berry knows full well that this is no evidence whatsoever that Kirby did steroids. Its value, if any, is as a method for recognizing abrupt performance improvements. My problem with this analysis in that context is that it ignores subsequent seasons, so that it cannot distinguish between a player who has abruptly but clearly raised his performance level and a player who is having a fluke season. Berry lists the 40 most extreme home run seasons since 1923, and one finds examples of both the former (Lou Gehrig in 1927, Hank Greenberg in 1938, Johnny Mize in 1947, Stan Musial in 1948, Carl Yastzremski in 1967, Jeff Bagwell in 1994, Sammy Sosa in 1998) and latter (Tommy Holmes in 1945, Chico Fernandez in 1962, Bert Campaneris in 1970, Davey Johnson in 1973, Enos Cabell in 1977, Wade Boggs in 1987). Puckett's 1986 was a signal of an actual performance jump, as he notched 20 or more homers five more times before his career's abrupt end. Thus, the performance I am personally most impressed with is Brady Anderson's 50 in 1996, with odds of 6½ in 10 million.

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 (but no death threats, please) 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. (But if you want to make my life a bit easier: please, use two spaces after the period in a sentence. Everything else is pretty easy to fix.)

If you can (and I understand it isn't always possible), try to format your article roughly the same way BTN does, and please include your byline at the end with your address (see the end of any article this issue).

Deadlines: January 24, April 24, July 24, and October 24, for issues of February, May, August, and November, respectively.

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
18 Deerfield Dr. #608, Nepean, Ontario, Canada, K2G 4L1
birnbaum@sympatico.ca

Brief Reviews

Phil Birnbaum

The author gives short reviews from recent non-academic sabermetric studies. This issue: a possible log5 alternative, a proposed method to forecast a player's batting average, and a pitch count estimator statistic.

Jai-Alai Analysis Suggests Log5 Alternative

The book *Calculated Bets* details the author Steven Skiena's successful attempt to turn a consistent profit on Jai-Alai pari-mutuel betting (which is legal in Connecticut). In his entertaining book, Skiena describes how he wrote a computer simulation to determine the relative probabilities of winning combinations, and refined his system to calculate the odds on the fly, betting only when the odds were in his favor.

Many of the lessons of Skiena's jai-alai research could apply equally to baseball research; but one finding in particular is of interest to Sabermetricians. Skeina's system requires an estimate of the probability of, say, a .550 Jai-Alai player beating a .475 player. For baseball, Bill James came up with the log5 method two decades ago; here, on page 103, Skiena (who seems unaware of the baseball equivalent) invents an alternative:

$$prob = \frac{1 + [wp(A) - wp(B)]^{\alpha}}{2}$$

Here, "prob" is the probability favorite A beats underdog B. The terms wp(A) and wp(B) are A and B's theoretical winning percentages against a .500 league. The exponent, alpha, is determined empirically – Skiena found that a value of 0.4 works well for his data.

Steven Skeina, Calculated Bets, 2001, Harvard University Press, ISBN 0521009626

Batting Average Forecasting

In an essay on page 7 of the *2003 Baseball Forecaster,* John Burnson discusses a new method for predicting a player's batting average. First, he notes that, by definition,

$$BA = \frac{BallsInPlay}{AB} * \frac{Hits}{BallsInPlay}$$

He then says that Baseball Forecaster research shows that the first term, balls in play per at-bat, is consistent for players from season to season. Therefore, coming up with an estimator for the second term, hits/balls in play, would allow us to estimate BA. A regression leads to a formula for this second term, and then Burnson tests his new estimator, which he calls xBA.

The results: as a predictor, xBA beat the previous year's BA by 55:45. Further, "when the discrepancy between xBA and prior year's BA was large (at least 30 points), our equation's predictive edge grew to almost 2:1 vs. prior year's BA, and xBA predicted the direction of change in 80% of those cases."

However, it should be pointed out that prior year's BA is not a particularly good estimator of current year's BA. Because of regression to the mean, players who had a higher BA last year would be predicted to drop, and players with a lower BA would be predicted to rise. It's possible that a simple formula like "last year's BA, but brought 20% closer to the mean," or even "a weighted average of the player's last three years' BA" would perform better than xBA. But Burnson's approach is interesting, and deserves a closer look.

Ron Shandler (et al), Baseball Forecaster 2003, Shandler Enterprises LLC, ISBN 1891566032

Pitch Count Estimators

If a pitcher throws seven innings, striking out five and walking three, how many pitches has he likely thrown? Prolific sabermetrician Tangotiger, who publishes on various websites, has two ways of addressing the question.

The first model from Tangotiger (who is also known as Tom) is

$$PitchCount = 3.3 * BIP + 4.8 * K + 5.5 * BB$$

This formula is based on averages for every pitcher in the league. Since the average plate appearance that led to a ball in play ended in 3.3 pitches, we just estimate than any such PA consumed 3.3 pitches.

Tom recognizes, however, that the coefficients should be different for every pitcher. Since Nolan Ryan strikes out more batters than average, and walks more batters than average, we would expect he goes deep in the count more than average. And therefore, even when the batter puts the ball in play, he probably did so after more pitches than the average 3.3.

Similarly, the average pitcher, when striking out a batter, does so after 1.8 non-strikes (since there are 4.8 pitches in the average strikeout). But Nolan Ryan, who throws so many balls, might have a higher number of non-strikes per strikeout. The same logic applies to walks.

Tom's new formula tries to estimate new values for the coefficients, based on the characteristics of the pitcher – namely, his ball-in-play rate. The new coefficients are:

$$PitchesPerBIP = 2.5(1 - BallInPlayRate)^{0.08} + 1$$
$$PitchesPerBB = 1.5(1 - BallInPlayRate)^{0.10} + 4$$
$$PitchesPerK = 1.9(1 - BallInPlayRate)^{0.07} + 3$$

Tom writes, "I encourage other sabermetricians to pick up where I left off … I'm reasonably certain that the basis for the model is correct, but this would be better established with actual data."

Any empirical test would have to keep in mind that because managers won't leave pitchers in for high pitch counts, high predictions might be incorrect. For instance, if a pitcher's line predicts he threw 200 pitches, he probably didn't – he likely did not go as deep in the count that day as the formula predicts, because the manager wouldn't have left him in that long if he had.

Website: http://www.baseballstuff.com/tangotiger/pitchCountEstimator.html

The editor encourages readers to submit short reviews so we can make this a regular feature of BTN. Phil Birnbaum, 18 Deerfield Dr. #608, Nepean, ON, Canada, K2G 4L1, birnbaum@sympatico.ca ♦

The Relationship Between Skin Tone and Performance – A Preliminary Analysis

Duke Rankin

Cubs manager Dusty Baker recently sparked controversy when he asserted that black players perform better in hot weather than white players do. Here, the author looks for evidence of whether the assertion is, or isn't, supported by the statistical record.

Introduction

Baseball certainly has its share of controversies. Recently, for example, Dusty Baker stated black players perform better in the heat of summer:

> We were brought over here for the heat, right? Isn't that history? Weren't we brought over because we could take the heat? Your skin color is more conducive to heat than it is to the lighter-skinned people. I don't see brothers running around burnt. That's a fact. I'm not making this up. I'm not seeing some brothers walking around with some white stuff on their ears and nose (http://espn.go.com/mlb/news/2003/0707/1577519.html).

Setting aside the sociological implications of this observation, Baker presents a testable hypothesis: players with dark skin tones should play better in warm weather than players with light skin tones. At the risk of offending large sections of the baseball public, I examined this question by comparing the performance of differently colored players during different periods of the season.[1]

Methods

I selected three teams for the analysis: the Red Sox, Yankees and Braves. I chose these teams because I am reasonably familiar with the racial backgrounds of their players. In addition, the home cities of these teams form a climatic gradient while retaining the seasonal climates uncharacteristic of the more equitable climate of cities such as San Diego. Although certainly not a random sample, I have no reason to believe these three teams represent a biased sample for this analysis.

Next, I divided the MLB season into two climatic periods -- cool months (April, May, September and October) and hot months (June, July and August). The hot months exhibit the three highest mean temperatures for each of the three cities (Table 1).

For each team, I designated each member of the starting lineup as either white or black, based on the photograph of the player on the ESPN website combined with my personal knowledge of the player (Appendix 1). Please note this is a subjective determination of relative skin tone, and not a determination of racial origin. In general, skin tone was readily evident -- Alfonso Soriano's skin tone is clearly darker than Robert Fick's skin tone. For a few players, however, skin tone was intermediate. Derek Jeter, for example, has one black parent and one white parent. I designated these players as intermediate condition, and expanded my hypothesis: if skin tone determines player performance in regards to temperature, then players of intermediate skin tone should exhibit an intermediate response to changes in temperature. I placed many of the Latin players into the intermediate category.

Table 1: Monthly mean temperature (in degrees Fahrenheit) for the three cities in the study (downloaded from www.weather.com).

	Boston	New York	Atlanta
April	48	52	62
May	59	63	70
June	68	72	79
July	74	77	80
August	72	76	79
September	65	69	73
October	54	58	63

[1] I would like to thank three anonymous reviewers for their comments regarding the statistical analysis. I also thank Phil Birnbaum for his insights into the analysis and assistance in completing the manuscript.

For each starter, I downloaded basic hitting statistics from the splits section of www.sports.espn.go.com/mlb/ for the three year period 2000-2002. Pitchers were excluded from the study because the three teams have only one black starting pitcher (Pedro Martinez). Players without complete seasonal data were also excluded. For each player/climate period category, I totaled at bats, hits, walks and total bases, then calculated BA, OBP, SLG and OPS. The study was completed before the recent trades involving Robin Ventura and Raul Mondesi.

Results and Discussion

In general, white players hit for higher average and OPS in cooler months than warmer months (Table 2). Black players, on the other hand, hit better in warmer months. Players of intermediate skin tone exhibited an intermediate response, at least for batting average.

Table 2: Differences in batting performance based on skin tones and climatic period of the MLB season

Skin Tone	Players	Total AB	BA Cool Months	BA Warm Months	BA Difference	OPS Cool Months	OPS Warm Months	OPS Difference
White	13	15426	.288	.282	-.006	.858	.842	-.016
Intermediate	5	7276	.282	.285	+.003	.798	.829	+.031
Black	9	12253	.285	.294	+.009	.860	.885	+.025

For the purposes of the study, Jason Giambi is a good example of a cool-season hitter:

	AB	H	2B	3B	HR	BA
Cool Months	820	283	56	1	72	.345
Hot Months	770	241	54	3	50	.312

Gary Sheffield, meanwhile, is a good example of a warm-season hitter:

	AB	H	2B	3B	HR	BA
Cool Months	701	203	28	3	39	.289
Hot Months	807	271	50	2	65	.335

And Derek Jeter exhibits an intermediate response:

	AB	H	2B	3B	HR	BA
Cool Months	847	262	43	4	23	.309
Hot Months	1004	321	49	3	31	.319

The most enigmatic player was Johnny Damon. On the one hand, I had trouble characterizing his skin tone -- his features strike me as intermediate, but Damon is from Kansas, and, in the absence of better information, I categorized him as white. Damon, however, is one of the most pronounced warm season hitters in the study:

	AB	H	2B	3B	HR	BA
Cool Months	939	255	37	15	19	.272
Hot Months	983	302	73	10	20	.307

The data are quite limited, and the differences almost certainly insignificant.[2] Clearly, the study would be improved by increasing the sample to include all major league players or more effectively randomizing the sample. Nevertheless, the data are consistent with the Baker

[2] The statistical analysis of the data is problematic. One reviewer compared the means in each skin tone treatment using Student's t-test, and concluded none of the differences are statistically significant, including the seasonal differences exhibited by the individual players. A second reviewer used a *z* score patterned after Pete Palmer's test for clutch hitting and found only one potentially significant difference -- the decrease in BA for white players from cool season to warm season. A third reviewer concluded it was not possible, given the data, to determine the statistical differences of the data.

hypothesis: black players tend to perform better in warm months, at least in comparison to their performance in cool months. White players exhibit the opposite trend.

Although the seasonal differences may not be statistically significant, they may be meaningful in a baseball context. Regressing team batting average against runs scored for 2002 suggests a fifteen point increase in team batting average translates into a 50 - 75 increase in runs scored over the course of the year. This would represent, however, the extreme implementation of the seasonal effect -- a team platooning black and white players of equal abilities at all eight positions. The seasonal effect is actually quite small: one hit every 67 at bats, or eight to nine hits per player over the course of the season. The seasonal effect is probably inconsequential in comparison to the overall quality of the players, or more traditional platoon advantages. The average left/right BA differential for the players in the analysis, for example, is fifty points (excluding switch hitters).

Unfortunately, the study creates more questions than it answers. Is the effect real, or simply an artifact of the teams and players chosen? Does the effect extend to pitchers? Is the effect more pronounced in cities with pronounced seasonal changes -- should Boston, for example, use white players preferentially in the spring and fall, while San Diego use intermediate players all season? Is the effect more pronounced across a climatic gradient -- should Detroit, for example, preferentially acquire white players, while Florida acquires black players? Can the climate in which a player is raised reverse the trend -- do white players from warm climates hit better in warm weather than black players raised in cold climates? Fortunately, these represent testable hypotheses that can be addressed by further research.

Appendix

Players in the study, listed by skin tone categories. Categories are based on subjective appraisals of color, and do not necessarily reflect racial background.

White: Damon, Fick, M. Franco, Giambi, M. Giles, N. Johnson, C. Jones, Millar, Mueller, Nixon, Varitek, Ventura, Walker
Intermediate: Castilla, Garciaparra, Jeter, J. Lopez, Posada
Black: J. Franco, Furcal, A. Jones, Mondesi, D. Ortiz, M. Ramirez, Sheffield, Soriano, B. Williams

Duke Rankin, drankin@hiwaay.net ♦

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.

To join, or for more information, send an e-mail (preferably with your snail mail address for our records) to Neal Traven, at beisbol@alumni.pitt.edu. Or write to him 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.

On the Pythagorean Winning Percentage

Jeff Thurston

The Pythagorean Projection is widely used to compute a team's winning percentage based on its runs scored and runs allowed, and a large body of research has shown it to be accurate. But why does it work? Here, the author examines some theoretical implications of run scoring, with a view to casting some light on the success of the Pythagorean formula.

Introduction

The Pythagorean formula is frequently used to compute a team's winning percentage using only the average number of runs the team scores and allows. Widespread application of the formula is attributable to the fact that it consistently generates results that approximate actual winning percentages. However, to my knowledge, there has been no mathematical proof demonstrating why this is so. As a consequence, the reasons for the accuracy of the formula are not well understood. Improved understanding would certainly be welcome, perhaps instructing us in, for instance, assigning meaning to departures of actual results from expected ones.

Originally this study was conceived to bolster understanding of the Pythagorean formula. The plan was to provide a formal derivation of this equation. I optimistically assumed then that the assumptions and approximations necessarily imposed to proceed through the steps of the derivation would provide insight useful for interpreting the numbers produced by the formula. While I do make some progress in this regard, by providing a proof, I fall short in my effort to draw a connection between knowledge garnered from the derivation, and an understanding as to why a team does or does not perform according to the formula's predictions. Nevertheless, all is not lost, as a by-product of the mathematical analysis is two new formulae, one of which is shown to hold some promise for assessing team performance.

I begin this paper by showing that the Pythagorean winning percentage is the probability that a team wins two-run games, assuming that runs are scored by and against the team according to Poisson processes. Studies that have been published in the statistical literature suggest that run scoring is better described by a mixed Poisson (i.e. a negative binomial), rather than by a simple Poisson distribution. Further, there is nothing to suggest that a two-run difference has some special status as a particularly diagnostic margin. As a consequence, there is no immediately obvious reason as to why the Pythagorean formula works as well as it does.

Thus, while my original goal has not been achieved; the derivation does have some value, as it does lead to two new formulae, the first of which makes it possible to calculate the probability of winning as a function of the final run differential (i.e. the conditional probability of victory, given a final margin of victory/defeat). The second equation can be used to compute overall probabilities of winning (i.e. the probability of victory for all margins of victory/defeat). It is important to note that both these equations arise from the assumption that runs occur according to a Poisson distribution. Because this is a statistically accurate assumption in only a limited number of games (Keller, 1994), one might be disinclined to present these formulae as universally reliable predictors of winning percentages. However, the Pythagorean method, which relies on the same assumption, has enjoyed ubiquitous application, and surprising accuracy, and this legacy provides sufficient motivation to assess the ability of these two new closely related formulae to predict winning percentages. I do this with two tests. In the first, I compare the probability of victory (given a final run differential) with actual winning percentages at a specified run differential. This comparison suggests that, on average, teams perform quite closely to what this first equation predicts. In the second test I compare actual and Pythagorean winning percentages with the probabilities computed by the second equation. This suggests that this second equation less accurately models winning percentage than does the Pythagorean formula.

The remainder of the paper is concerned with analysis and an application of the run-differential dependant equation. In the analysis, I show how this formula dispels the notion that good teams (i.e. teams that average more runs than their opponents) should perform their best in close games. The application presents an example demonstrating how this formula can be used for analysis. In this case, I compare expected and actual winning percentages in close games, and discuss potential implications of this for assessing managerial performance.

Before continuing, the reader should be forewarned that, buoyed by the reasonable accuracy demonstrated in the aforementioned test, I routinely refer to the conditional probabilities (i.e. the run-dependant probabilities) in terms of *expected winning percentages as a function of run differential*. This is done in spite of the fact that, as previously mentioned, the underlying assumption (namely that Poisson processes govern run scoring) is likely oversimplified. Nevertheless, the accuracy of the Pythagorean formula, together with initial indications of the accuracy of the new run-dependant equation, suggest that perhaps this oversimplification is close enough for many purposes.

Derivations

For two Poisson processes with averages m_1 and m_2, Skellam (1946) has shown that the probability that the distribution with mean m_1 exceeds the distribution with mean m_2 by exactly r is:

$$P_r(m_1,m_2) = e^{-(m_1+m_2)}\left(\frac{m_1}{m_2}\right)^{\frac{r}{2}} I_r(2\sqrt{m_1 m_2}), \quad (1)$$

where I_r is a modified Bessel function of the first kind. Likewise, the probability that a variate from the distribution with mean m_1 is less than a variate from the distribution with mean m_2 by exactly r is

$$P_{-r}(m_1,m_2) = e^{-(m_1+m_2)}\left(\frac{m_2}{m_1}\right)^{\frac{r}{2}} I_{-r}(2\sqrt{m_1 m_2}). \quad (2)$$

The probability of a difference of $\pm r$ is then:

$$P_{\pm r} = P_r(m_1,m_2) + P_{-r}(m_1,m_2). \quad (3)$$

Given a difference of $\pm r$, the probability that the Poisson distribution described by mean m_1 exceeds the distribution described by m_2, is then:

$$W_r(m_1,m_2) = \frac{P_r(m_1,m_2)}{P_r(m_1,m_2) + P_{-r}(m_1,m_2)}. \quad (4)$$

Since I_r is equal to I_{-r}, when equations (1) and (2) are substituted into (3), the Bessel functions cancel. As well, the *exp* terms cancel, leaving:

$$W_r(m_1,m_2) = \frac{m_1^{\,r}}{m_1^{\,r} + m_2^{\,r}}. \quad (5)$$

The quantity $W_r(m_1, m_2)$ is the conditional probability a team that scores runs at a rate of m_1, and allows runs at a rate of m_2.wins, given that the final difference is r runs.

For $r = 2$, equation (5) becomes the familiar Pythagorean winning percentage formula, which can now be described as the probability that a team wins two-run games. The inherent assumption made in using this well-known formula is that runs are scored and allowed according to different Poisson processes, with means equal to the average runs scored and allowed per game. Given that run scoring likely follows a negative binomial distribution (see e.g. Reep et. al (1971), rather than a Poisson distribution, and that there is nothing obviously unique about a two-run margin, I find it difficult to use this description of the theoretical underpinnings of the Pythagorean formula to both understand why it performs as well as it does, and to assess the importance of a team's deviation from it.

It is straightforward to derive a formula analogous to the Pythagorean formula, but which takes into account all margins of victory/defeat. That is, Keller (1994) has shown that if runs are scored at a rate of m_1, and surrendered at a rate of m_2, then the probability of victory is:

$$P_v(m_1,m_2) = e^{-m_2}\int_0^{m_1} e^{-m_1} I_0(2\sqrt{m_1 m_2}\, dm_1. \quad (6)$$

Likewise, the probability of defeat is:

$$P_d(m_1,m_2) = e^{-m_1}\int_0^{m_2} e^{-m_2} I_0(2\sqrt{m_1 m_2}\, dm_2. \quad (7)$$

Equations (6) and (7) lead to a conditional probability of victory (conditional on the fact that ties are not permissible) of:

$$W(m_1,m_2) = \frac{P_v}{P_v + P_d}. \quad (8)$$

It is important to emphasize the difference between equations (5) and (8). The former estimates a probability of winning for a specified run differential, whereas the latter is the probability regardless of the final run margin.

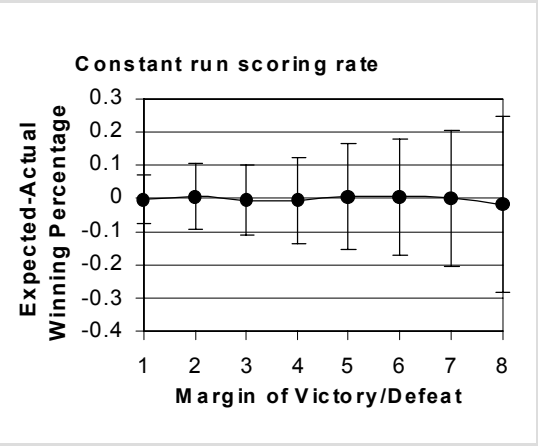
Testing the New Formulae

The validity of the underlying assumption, namely that scoring follows a Poisson distribution, has been studied in numerous publications (see e.g. Reep et. al (1971), Keller (1994) and references therein). In general these authors have concluded that the run distribution is typically overdispersed (i.e. the variance is greater than the mean). Nevertheless, the success of the Pythagorean formula is sufficient motivation to test the accuracy of these formulae. This is done by comparing computed and actual results.

Testing of equation (8) using recent results suggests it less accurately conforms to actual winning percentages than the does the conventional Pythagorean formula. For the 2002 American League, equation (8) estimates had a mean squared error of roughly six (as compared to an approximate mean-squared error of four for the Pythagorean formula). In light of this, it is remarkable that the Pythagorean formula can be used to accurately estimate a team's total number of wins, as it remains unclear as to why a team's probability of victory in two-run games is typically close to its actual winning percentage for all margins of victory/defeat.

On the other hand, preliminary testing of equation (5) suggests that it provides probabilities that are reasonably similar to actual winning percentages. However, before presenting these results, it is important to point out one computational detail. That is, in order to use equation (5), average runs scored and allowed are required. There are two ways in which these rates can be computed. The simplest method is to assume that these average-run rates are constant for each team in a particular season, and thus are unchanged regardless of the final run differential. Then, m_1 and m_2 are respectively the total runs scored and allowed divided by the number of games played. The second approach is to assume that run rates depend on the final run differential. In this second approach, unique values for m_1 and m_2 are computed for each final run differential for each team. For instance, $m_1(r)$ is the average of runs scored in games ending with a difference of r, and likewise for $m_2(r)$.

Aside from the results shown in this section, in which I assess the accuracy of equation (5), in the examples I present later, expected winning percentages for a single team in a single season are computed using a constant run scoring rate; however when analyzing composite results (i.e. the results of several teams and seasons combined) a variable run rate is used.



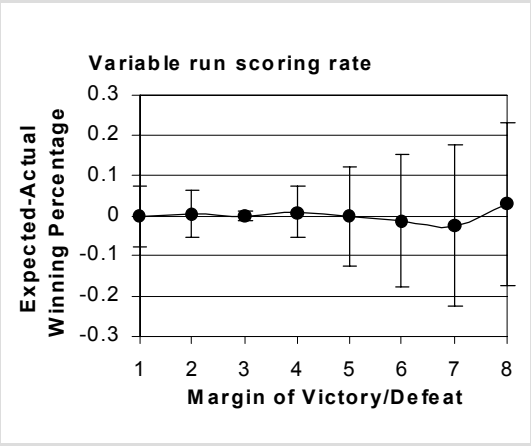


Figure 1. Average differences between the expected and actual winning percentages for teams in the American League for the years 1997-2002, as a function of the run differential. The left-hand panel is based on a constant run scoring rate, while the right-hand panel uses a run scoring rate that depends on the final run margin. The expected values were computed using equation (5). Error bars are the standard deviations.

Average differences between the quantity generated by equation (5) and actual winning percentage as a function of run differential, using scores from games played by American League teams for the years 1997-2002, are shown in Figure 1. These differences are fairly typical, as similar results are obtained for any single season for the past twenty years. These results suggest that equation (5) seems to be a reasonably reliable means for calculating expected winning percentage at a specified run differential.

Variation of expected winning percentage with the final margin of victory/defeat

If equation (5) describes how a team's expected winning percentage depends on the final run differential, then further analysis reveals an interesting property regarding the nature of this variation.

Differentiating $W_r(m_1, m_2)$ with respect to r, gives, after simplification:

$$\frac{\partial W_r(m_1,m_2)}{\partial r} = W_r(m_1,m_2)[1-W_r(m_1,m_2)]\ln\left(\frac{m_1}{m_2}\right) \qquad (9)$$

Equation (9) makes it clear that, provided m_1 and m_2 are constant, $W_r(m_1, m_2)$ has a positive gradient whenever $m_1 > m_2$, and a negative gradient whenever $m_1 < m_2$. The implication of this is that teams that on average outscore their opponents are expected to have their lowest winning percentages in close games. Their likelihood of winning increases as the run differential increases. On the other hand, teams that on average are outscored, have their maximum winning percentage in one-run games. Their likelihood of winning decreases as the difference in runs increases. Thus, it is important to note that good teams that have their lowest winning percentages in close games are performing as expected.

As an example, expected and actual winning percentages for the 1929 Yankees are shown in Figure 2. This suggests the Yankees performed better than expected in close games, but faltered somewhat in games decided by larger margins.

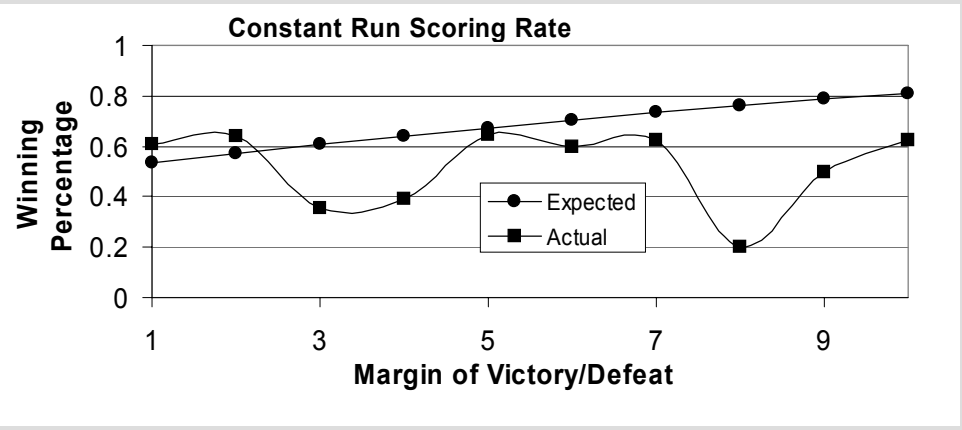


Figure 2. Expected and actual winning percentages for the New York Yankees in the 1929 season, as a function of the run differential.

Example: A method for evaluating managerial performance (perhaps)

The ability to calculate an expected winning percentage at a specified run differential presents several opportunities for analysis. In the following, I provide an example that shows that manger-of-the-year award winners have, in the past six seasons in the American League, been charged with teams that, on average win more than their share of close games. There are two conclusions that can be drawn from such an observation:

> Better than expected performance in close games is a reflection of managerial skill, and teams managed by exceptional all-around managers (i.e. by a manager-of-the-year award winner) have performed accordingly.

Or, equally tenable:

> Better than expected performance is an important criterion for the manager-of-the-year award, and teams that win more than their share of close games will overachieve in general (overachieving in close games is a particularly important prerequisite of overachievement, as close games are far more frequent than games with large differences in run totals).

Determining which of these is true is a matter of establishing cause and effect, and I am unsure as to how this might be done.

Figure 3 shows results for American League teams for the seasons 1997-2002. I have done several additional tests, and these results are typical for games played by American League teams in the past twenty seasons. On Figure 3, the following expected characteristics are apparent:

1. Average actual winning percentages that significantly exceed expected winning percentages in close games for teams managed by manager-of-year award winners.
2. Insignificant average differences between actual and expected winning percentages in close games for the league as a whole.
3. Insignificant average differences between actual and expected winning percentages, for games that end with large run differentials, for the league as a whole.

A rather surprising result is:

Significant average differences between actual and expected winning percentages for games that end with large run differentials, for teams managed by manager-of-year award winners.

This latter observation may be an artifact attributable to the fact that there are larger deviations between actual and expected percentages at higher run differentials (see Figure 1). Nevertheless, even if this is a true representation, suffering lower winning percentages at higher margins does not nullify the value of producing higher percentages at low margins, as close games occur far more frequently than do games with large margins of victory/defeat.

Finally, it is interesting to note that large differences between expected and actual winning percentages in one-run games are at least partially correlated with the voting for the manager-of-the year award. This was particularly true in the 2002 season. Table 1 shows the comparison between the teams with the maximum differences between expected and actual winning percentages in one-run games, and the voting results.

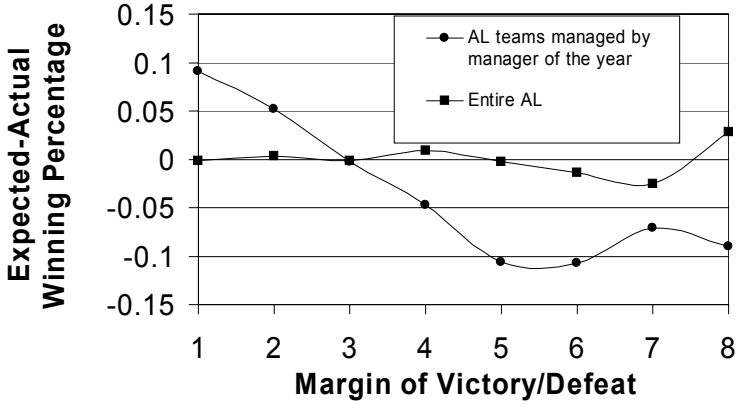


Figure 3: Comparison of average differences between expected and actual winning percentages for teams in the American League for the years 1997-2002. Better than expected performances are indicated by positive differences. Average run rates are variable, and depend on the final margin of victory/defeat.

Summary

I have shown that the well-known Pythagorean winning percentage formula is actually a team's probability of winning two-run games, assuming run scoring is governed by Poisson processes. It remains unclear as to why this probability is consistently close, over many different teams in many different seasons, to actual overall winning percentages. This is particularly true in light of the fact that I have presented a formula that does not depend on the run differential, but is less accurate than this famous formula. Perhaps this somewhat cursory study will serve as motivation for a more thorough investigation that addresses this rather interesting phenomenon.

Team	Difference between expected and actual winning percentage in one-run games	Team	Points earned in vote for manager of the year
Oakland	0.169	Anaheim	116
Minnesota	0.127	Oakland	74
Anaheim	0.074	Minnesota	59
Toronto	0.020	New York	3

Table 1. Comparison between voting results for manager-of-the-year award, and teams with the maximum differences between actual and expected winning percentages in one-run games. As before, better than expected performances are indicated by positive differences.

While this study fails in its original objective to provide a conceptual understanding of the Pythagorean formula, my efforts are not entirely fruitless, as a potentially useful formula for computing the conditional probability of victory, given a final run differential, has been derived. I refer to this quantity as the expected winning percentage as a function of run differential, as on average, albeit in a test of limited scope, expected and actual results are similar to one another. This equation makes it possible to calculate expected variations in winning percentage with run differentials for both good and bad teams. These expected variations suggest that good teams should not be expected to have their highest winning percentages in close games. Finally, a method is presented to demonstrate how the run-dependant winning percentage equation might be used to assess managerial performance. The results of this method suggest that recent manager-of-the-year award winners achieve better than expected results in close games.

References

Keller, J.B., 1994, *A Characterization of the Poisson Distribution and the Probability of Winning a Game,* The American Statistician, Vol. 48, No. 4, pp. 294-298.

Reep, C., Pollard, R. Benjamin, B., 1971, *Skill and Chance in Ball Games*, Journal of the Royal Statistical Society, Vol. 134, No. 4, pp. 623-629.

Skellam, J.G., 1946, *The Frequency Distribution of the Difference Between Two Poisson Variates Belonging to Different Populations,* Journal of the Royal Statistical Society, Vol. 109, Issue 3, p. 296.

Jeff Thurston, 3628 7A St SW, Calgary, AB, Canada, T2T 2Y5, jthursto@ucalgary.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
Jim Box	im.box@duke.edu	Statistics
Keith Carlson	kcarlson2@mindspring.com	General
Rob Fabrizzio	rfabrizzio@bigfoot.com	Statistics
Larry Grasso	l.grasso@juno.com	Statistics
Tom Hanrahan	HanrahanTJ@navair.navy.mil	Statistics
John Heer	jheer@walterhav.com	Proofreading
Dan Heisman	danheisman@comcast.net	General
Keith Karcher	karcherk@earthlink.net	Statistics
Chris Leach	chrisleach@yahoo.com	General
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
Joel Tscherne	Joel@tscherne.org	General
Dick Unruh	runruhjr@dtgnet.com	Proofreading
Steve Wang	scwang@fas.harvard.edu	Statistics

Toying with "The Favorite Toy"

Shane Holmes

Bill James developed "The Favorite Toy" as a method for estimating a player's chance of reaching a career goal (e.g., 3000 hits). Here, the author tests the performance of the Toy, by checking how well its predictions matched what eventually occurred.

The Favorite Toy (TFT) is a Bill James invention for estimating the likelihood that a player will reach a specific career milestone. In his annual abstracts, James used TFT to handicap a player's chances of reaching 3,000 hits.

The method, and an example, can be found in the boxes on this page and the next.

Testing the Toy

A few months ago, *Baseball Primer* contributor "Tangotiger" suggested testing TFT to see how accurate its predictions have been to date. I took the bait.

The Toy's test is based on the idea that over a large group of players, the total of their project probabilities of reaching a goal should equal the number of players who actually do reach it. For instance, if eight players each have a 25% chance of reaching 3,000 hits, and TFT is accurate, then two of the eight should actually make it.

I started by looking at all players (active and retired) who reached the halfway mark of a career milestone, and finding each player's age when he attained that point. For the 500 HR milestone, for example, I compiled a list of players who had hit 250 or more HR, and their ages when they hit number 250.[1] Then I computed every player's TFT chance of reaching the milestone (ignoring hitters whose chance falls below 0 and capping hitters at a 1.000 clip; we do this not because it improves the analysis - it doesn't - but because TFT is after all a toy). The sum of all of those players' chances should come close to equaling a count of the total number of players who succeed in attaining the milestone.

The Favorite Toy

TFT has four components and a result.

Need Hits

Need hits is the number of hits needed to reach the goal. Let's say the goal is 500 career home runs. If a player has accumulated 50 roundtrips in his career thus far, then his "need hits" equals 450.

Years Remaining

"Years remaining" is the number of seasons a player has left in his career. It's estimated by the formula 24 - 0.6 * (age). This formula gives 9.0 remaining seasons to a 25-year-old, 6.0 to a 30-year-old player, and 3.0 to a 35-year-old player. A caveat: A regular player is always estimated to have at least 1.5 remaining seasons.

Established Hit Level

The "established hit level" is a weighted formula designed to equal the number of hits the player achieves on a seasonal basis. For year 2003, the established hit level would be found this way: First, find the sum of the player's 2000 hits, 2 times his 2001 hits, and 3 times his 2002 hits. Then, divide the sum by 6. Another caveat: A player's performance level must exceed or be equal to 75% of his most recent season performance. This protects against certain intrusions in a great player's career, such as a season-long injury or a labor-related work stoppage.

Projected Remaining Hits

This projection is the product of the above two steps: Years Remaining multiplied by the Established Hit Level.

Result

Calculate

$$\frac{\mathrm{Pr}\,ojected\ \mathrm{Re}\,mainingHits}{NeedHits} - 0.5$$

This is the Favorite Toy's estimate of the probability of the player reaching his goal.

[1] Because I don't have the exact dates on which players reached 250 HR or 1,500 base hits, I had to estimate each player's age at the time he reached that halfway point. If a player in pursuit of 500 HR reached number 250 early in a season but his birthday fell late in a calendar year, he was considered to have been the younger age. In other words, I did not simply subtract every player's birth year from the relevant major league season year.

I tested TFT in this manner for two milestones: 3,000 hits, and 500 HR.

Test: 3,000 Hits

Going into 2003, 506 hitters had reached 1500+ hits. I looked at every player with 1,500 base hits and found his TFT-handicapped chances at 3,000. According to TFT, on the date they reached 1500, 5 of the 506 players had a 50+% chance at getting the remaining 1500 hits they needed to reach 3000.

Name	Chance
Cobb*	0.7768
Hornsby	0.5608
Sisler	0.5465
Keeler	0.5080
Aaron*	0.5036

* These players eventually accumulated 3000 or more base hits

When one adds those chances, the sum is nearly 2.9. This is close to the number of hitters from that output who succeeded (two). So far, so good.

The next chart does for every chance group what I have done in words to describe the 50+% chance group. (As you can see by following the second row across, the chart displays the figures about which I just wrote, 2.9, 5, and 2.)

TFT Example – Chipper Jones

Chipper Jones had 253 homeruns at the close of the 2002 season. What were his chances of reaching 500 by the end of his career?

His need HR was 247;

His years remaining was 6 (he was 30 years old by the time he slugged homerun #253);

His last three seasons of 26 HR (in 2002), 38 (in '01), and 36 (in '00) give him an established HR level of 31.7;

Multiplying 6 remaining seasons by 31.7 HR gives Chipper 190 HR projected remaining home runs from 2003 until the end of his career;

Dividing is 190 projected hits by his 247 need hits, then adding 0.5, gives a 26.9% chance that Chipper Jones will reach 500 career home runs.

Chance of achieving 3,000 hits	Number of players in sample	TFT estimate of number successful	Actual number successful
50% or better	5	2.9	2
25% to 49%	41	14.0	13
10% to 24%	65	10.6	5
0% to 9%	395	2.9	5
Total	506	30.4	25

In this study, most of the players have low chances because I included every player who reached the 1,500 standard. Manny Sanguillen, for example, finished his career right on 1,500. He obviously didn't have any chance at 3,000 when he made it to the halfway mark and retired. James designed TFT for great players, but he never went far in describing whose chances should not be handicapped. For my study, I set the bar at the halfway mark of the 3,000-hit milestone. That seemed fair if not generous. And, in fact, the toy's effectiveness is as evident in the 0-9% row as it was in the first row. In both cases, reality and the estimate are close enough for the casual use befitting a toy.

And the overall total of 30.4 estimated, 25 actual is also pretty good. The favorite toy acquits itself well.

However, it's important to note that included among these 506 hitters are several active players. The data for this study was treated as its own population. Because of the inclusion of actives, the Toy's estimates should exceed the count of actual real-life players who met the mark. That is, by the time all careers have ended, the number of successful players will likely grow from 25.

Age

To check if TFT is also accurate within age groups, I repeated the table after grouping the above players by age. Here are players under 30 years old:

Chance of achieving 3,000 hits	Number of players in sample	TFT estimate of number successful	Actual number successful
50% or better	5	2.9	2
25% to 49%	36	12.6	10
10% to 24%	17	3.1	2
0% to 9%	2	0.1	0

Again, TFT seems accurate.

The age 30-34 group:

Chance of achieving 3,000 hits	Number of players in sample	TFT estimate of number successful	Actual number successful
50% or better	0	0.0	0
25% to 49%	5	1.4	3
10% to 24%	48	7.5	3
0% to 9%	288[2]	2.8	5

Despite the odds given to them, several players who reached 1,500 hits while between ages 30-34 managed to hang around for another 1,500.

The bottom row, where the number of real-life hitters actually exceeds the estimate, is misleading. Paul Molitor and Rickey Henderson are part of the group, and both have odds greater than 8%. Dave Winfield, whose chance was 5.4%, would have also been around 8% had the work stoppage in 1981 not lowered his hit total. If these three players had been in the 10-24% group, their presence would have enabled that group to shadow TFT's estimate.

The two exceptional cases belong to Nap Lajoie (5.3%) and to Cap Anson (0.0%), and both carry reasonable explanations.

Lajoie reached 1,500 hits in 1905, a season in which he played just 65 games and racked up 82 hits. When James created TFT, he protected against the "established hit level" failing to reach 75% of the most-weighted season's total. He did not, however, protect against labor disputes, injuries, and other career intrusions that might have interrupted the most-weighted season. If one substitutes Lajoie's 1906 season total for 1905, Nap's chances soar to around 31%.

Cap Anson, meanwhile, played most of his career in the 19th Century, and we have come to expect statistical surprises in cases like his. The most games Anson had played in any single season up to 1885 were 112. If one adjusts for season length, Anson's hit totals (and thus his chance at 3,000) rise accordingly.

For completeness, here's the 35+ group:

Chance of achieving 3,000 hits	Number of players in sample	TFT estimate of number successful	Actual number successful
All 35+ players	105[3]	0.0	0

[2] Just 67 of the 288 players from the 0-9% chance group had a realistic shot; most in the group were 0%.
[3] Only about 5% of these players finished their careers with 2000 or more base hits.

Test: 500 Home Runs

Going into 2003, 151 hitters had reached 250+ HR. According to TFT, on the date they reached 250, 19 of the 151 players had a 50+% chance to reach 500.

Name	Chance
Foxx	0.9872
Griffey Jr.	0.9716
Rodriguez!	0.97
Killebrew*	0.868
Sosa	0.864
Kiner	0.8152
Gonzalez	0.7844
Aaron*	0.6492
Banks*	0.644
Ruth*	0.622
Snider	0.6132
Mantle*	0.5972
Mathews*	0.5864
Ramirez	0.5692
Gehrig	0.534
Schmidt*	0.52
Ott*	0.5192
Thome	0.516

* reached 500 HR
! capped

Once again, grouping TFT chance estimates, we find:

Chance of achieving 500 HR	Number of players in sample	TFT estimate of number successful	Actual number successful	Active at end of 2002 (excluding 500 HR hitters)[4]
50% or better	19	13.1	9	6
25% to 49%	26	9.2	7	9
10% to 24%	23	2.5	1	3
0% to 9%	83	0.0	0	11
Total	151	24.8	17	29

The favorite toy again does well, especially if you make an allowance for the number of active players who will eventually reach the goal.

Again, here are the breakdowns by age, starting with players aged 25-29:

Chance of achieving 500 HR	Number of players in sample	TFT estimate of number successful	Actual number successful	Active at end of 2002 (excluding 500 HR hitters)
50% or better	17	12.1	8	5
25% to 49%	10	3.0	4	2
Total	27	15.1	12	7

[4] Considering the furor surrounding the supposed dilution of the 500 HR Club, I've added this column to Part 2.

The age 30-34 group:

Chance of achieving 500 HR	Number of players in sample	TFT estimate of number successful	Actual number successful	Active at end of 2002 (excluding 500 HR hitters)[5]
50% or better	2	1.0	1	3
25% to 49%	16	5.3	3	7
10% to 24%	11	1.9	1	3
0% to 9%	58	0.6	0	8
Total	87	8.8	5	21

And the age 35+ group:

Chance of achieving 500 HR	Number of players in sample	TFT estimate of number successful	Actual number successful	Active at end of 2002 (excluding 500 HR hitters)
All	37	0.0	0	5

Again, if you take into account the active players still to come, the Favorite Toy seems quite accurate.

Active Players

Here are the top ten active players with a chance at 3,000 hits (calculations heading into 2003):

Name	Chance
RAlomar	0.8824
ARodriguez	0.3988
Guerrero	0.3649
Jeter	0.3616
Williams	0.2541
Tejada	0.2041
Erstad	0.1954
CJones	0.1926
Damon	0.1726
Pujols	0.1476

Other notables: MOrdonez 0.1415, Palmeiro 0.1407, Helton 0.1273, Renteria 0.1101, Sosa 0.1034, LCastillo 0.0982, Bagwell 0.0977, Vidro 0.0843, Ramirez 0.0838, Rolen 0.0835, Erstad 0.0826, Olerud 0.0657, Berkman 0.0538, Soriano 0.0418, Sweeney 0.0372, Ichiro 0, Kent 0, Vizquel 0, Grace 0, Piazza 0, Thomas 0, Bonds 0, Griffey Jr. 0, AJones 0, Sheffield 0, Garciaparra 0, Eckstein 0, Biggio 0, Martinez 0, Franco 0.

[5] Considering the furor surrounding the supposed dilution of the 500 HR Club, I've added this column to Part 2.

And the top ten active players with a chance at 500 HR (calculations heading into 2003):

Name	Chance
Sosa*	1.0000
Palmeiro*	1.0000
McGriff*	1.0000
Griffey Jr.*	1.0000
ARodriguez*	1.0000
Thome	0.9024
MRamirez	0.6526
Guerrero	0.6017
Bagwell	0.5900
AJones	0.4952

*- capped

Other notables: Green 0.4323, Delgado 0.4160, JGonzalez 0.3547, Piazza 0.3196, JaGiambi 0.3007, CJones 0.2692, Berkman 0.2544, Pujols 0.2385, Chavez 0.2291, BGiles 0.2147, Sheffield 0.2125, Thomas 0.1532, Soriano 0.1382, Burrell 0.1129, Vaughn 0.0229, Galarraga 0, MWilliams 0, Vaughn 0, Burks 0, Walker 0, Gant 0, TMartinez 0, Sierra 0, Ventura 0, Palmer 0, EMartinez 0, Karros 0, Salmon 0, Kent 0, LGonzalez 0.

Aaron

TFT wouldn't be much fun one didn't ask who if anyone will break Aaron's record.

Name	Chance
A. Rodriguez	0.3969
Sosa	0.2588
Bonds	0.0822

TFT underestimates Bonds' career length (he receives only 1.5 years left under TFT rules), but this is intentional because he is clearly an outlier.

Shane Holmes, *holmes@northwestern.edu* ♦