By the Numbers

Volume 13, Number 3 ** The Newsletter of the SABR Statistical Analysis Committee August, 2003

Summary

Academic Research: Batting Orders

Charlie Pavitt

The author reviews a recent paper on using Markov analysis to optimize batting orders.

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.

Joel S. Sokol, A Robust Heuristic for Batting Order Optimization Under Uncertainty, Journal of Heuristics, 2003, Volume 9, pages 353-370

In this issue

This is another attempt to use Markov process analysis of run potential from the various base-out situations to find optimal batting orders, and apparently it is a very good one. (For those of you unfamiliar with this topic, I strongly suggest reading Mark Pankin's piece in *The Best of By the Numbers*). Further, it makes a great deal of intuitive sense. Sokol's method begins by dividing each batter's offensive performance into two parts. He calls one part "potential value" and defines it as "how well the player creates good situations for the following batters" (p. 359); it is roughly the batter's ability to get on base. He calls the other part "realization value" and defines it as "how much the player takes advantage of the situation created by previous batters" (p. 359); this is roughly the batter's ability to hit for power. Using these, Sokol then classifies players as either "table-setters" (high potential value, low realization value), "all-around contributors" (high on both), "run producers" (high realization value, low potential value), or "weak hitters" (low on both). Not surprisingly, the optimal batting order uses that general sequence. Within those groups, one orders the table-setters according to increasing potential value, the all-around contributors according to decreasing potential value, the run producers according to decreasing potential value, and the weak hitters according to increasing potential value. As an example, Sokol's optimal batting order for the 1989 Giants started with table-setters Terry Kennedy (potential value = -.092) and Brett Butler (-.073), followed by all-around contributors Jack Clark (-.050) and Kevin Mitchell (-.071), then by run producers Robby Thompson (-.095), Candy Maldonado (-.098) and Matt Williams (-.140), and rounded out by weak hitters Pitcher (average -.149) and Jose Uribe (-.107).

Rather than running thousands of possible lineups and comparing each for run potential, Sokol's heuristic finds the best classification for a team's batters and then orders them as just described, with a few additional rules for dealing with exceptional cases. According to his data, its lineup selections are only about half a run a season worse than the best previous method, that of Bukiet, Harold and Palacios (Operations

** *A correction has been made to this issue since original publication. See page 8.*

Research, 1997, volume 45, pages 14-23), which I reviewed in BTN in August, 1999. Further, he claims his method to be a thousand times faster, quick enough for a manager to actually use for determining batting orders on game day. Unlike Bukiet et al., Sokol knows his literature; for example, he cites two of Mark Pankin's SABR presentations and uses some of Mark's baserunner advancement estimates in his work. Incidentally, I showed this paper to Mark, and he spent "an enjoyable and enlightening Sunday afternoon" playing with both the Sokol and Bukiet et al. models and concluding that both are about 7 runs better than his own.

Note: In my review of Jim Albert's Hitting with Runners in Scoring Position in the last issue of BTN, I noted a potential problem with Jim's run potential analysis of clutch hitting; the inclusion of circumstances in which base-out situations change during at bats due to stolen bases, caught stealing, wild pitches, passed balls, and balks. I invited a response from Jim, and I received one; he did remove those circumstances from the analysis. Thanks to Jim for the clarification.

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 study in which catcher influence on ERA is called into question; an old new stat from Yale researchers; an in-depth study of the Voros hypothesis; and a stock analyst's application of "Moneyball" to investing.

Can catchers save runs with their game calling?

In 1989, Craig Wright's essay "Catcher ERA" (in his excellent book *The Diamond Appraised*) argued that catchers can save their teams many runs with their ability to call the game or frame the pitch. A good catcher, in effect, makes his pitchers better, and since the statistical result goes in the pitcher's record rather than the catcher's, this aspect of defense goes underrated.

But in the 1999 *Baseball Prospectus*, Keith Woolner questions Wright's conclusions. "[W]ithout a comprehensive analysis … it's impossible to tell whether [Wright's] examples are selected because they serve to make [Catcher ERA] look good, or whether they are truly representative of a larger phenomenon."

Woolner sets out here to create that comprehensive analysis. He analyzes all seasons from 1981 to 1997, and compares a pitcher's LW/PA with a specific catcher to the same pitcher without that catcher.

If there were no catcher effect, the resulting distribution would result only from chance, and would follow a specific normal curve. But if the catcher does influence the pitcher's performance, the result would be wider than that normal curve, with more extreme (statistically significant) points.

Woolner's comprehensive analysis finds no evidence of catcher influence: the distribution of is exactly what would be expected if there were no such influence. Furthermore, Woolner says, there is almost no correlation between the catcher's influence this year and the catcher's influence next year – a result you would not expect if what is being measured is a skill of the catcher.

A confirmation of Woolner's results is found in a *Baseball Primer* essay by Chris Dial. Dial starts by arguing that Woolner's study should have used BB, K, and HR only, since Voros McCracken has shown that other plate appearances do not correlate with a pitcher's skill. He then analyzes pitcher/catcher pairs for each of those statistics only. His results confirm Woolner's.

So it seems catchers may not be as relevant as we thought. Which does seem strange: catchers do call pitches, so either it makes no difference what pitches they call, or they are so similar in what pitches they choose that no differences are evident. And what about pitch framing? Convincing an umpire to call a ball a strike saves .14 runs per pitch. Are catchers so equally talented at this that we can find no difference?

Keith Woolner, "Field General or Backstop?" Baseball Prospectus 1999, Brassey's, Inc., ISBN 1574881922
Website: http://www.baseballprospectus.com/news/20000110woolner.html

Chris Dial, "Game-Calling Revisited," http://www.baseballprimer.com/articles/cdial_2003-01-29_0.shtml

New Stat from Yale

According to a report in *Business Week* magazine (and a longer version available online), two Yale researchers, Benjamin Polak and Brian Lonergan, have come up with a "new" player evaluation statistic, which they are marketing to player agents.

Polak and Lonergan measure a team's chance of winning the game before a player's plate appearance, and after. The difference is credited (or debited) to the player. If (to use the Business Week example) the team has a 39% chance of winning before an at-bat, but only a 33% chance after, the batter is charged with negative .06 wins.

If this approach sounds familiar, it might be because it's been around since 1970. Several SABR-L posters pointed out that the Mills Brothers, Eldon and Harlan, came up with this idea over 30 years ago, and it was described in Thorn and Palmer's *The Hidden Game of Baseball* in 1984 (p. 47). The Mills' "win points" were apparently calculated differently from the Polak and Lonergan stat, but the approach is the same.

Another similar stat was introduced by Gary Skoog in the 1987 Abstract. That stat is calculated the same way as Mills/Lonergan, but based on runs instead of wins. Since a win is roughly ten runs, the Skoog number for a given player should be roughly ten times the Polak and Lonergan number. Any difference would be caused by either a different frequency of clutch opportunities, or different clutch performance.

According to the *Business Week* article, Polak and Lonergan are marketing their new stat to player agents. It quotes Jeff Moorad, agent for Manny Ramirez: "We'll use everything we can get our hands on." It also says Moorad expects at least half of MLB player agents to also be interested.

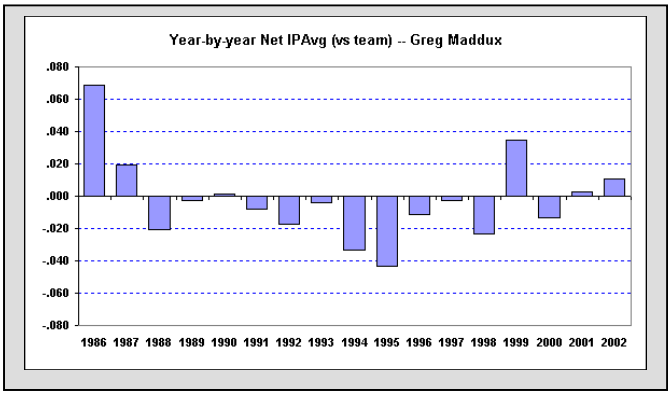
This is interesting, if true. While it might be expected that mainstream journalists would be unaware of existing sabermetric research, it is surprising to find agents interested in paying for a statistic that has been around for more than a generation, and that was written about in mainstream sabermetrics almost 20 years ago.

"Ballpark Figures to Bet On," Business Week, November 17, 2003, page 16
Website: http://www.businessweek.com/print/bwdaily/dnflash/nov2003/nf2003115_2313_db016.htm?db

An Important Voros Analysis

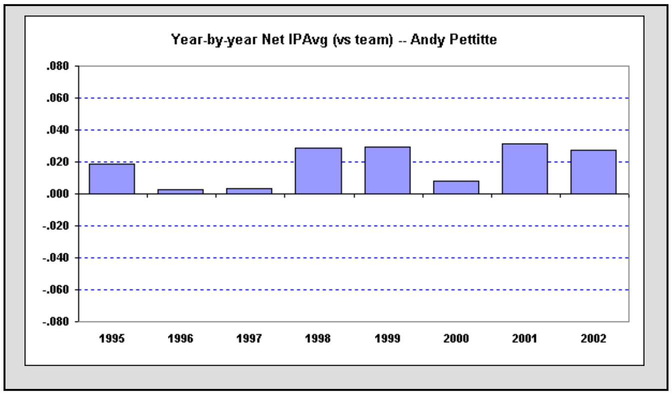
One of the fundamental things about of scientific research, as opposed to sportswriting, is that controversial findings are continuously revisited by other researchers, who either confirm or refute the original results.

This process continues with Voros McCracken's "balls in play" theory. McCracken suggested that if a ball is put in play, and stays in the park, whether or not it turns out to be a base hit is independent of the pitcher. That is, a ball hit of Pedro Martinez is no more likely to be converted into an out than a ball hit off any other pitcher.



Various studies of McCracken's theory (or "Voros' theory", as for some reason he is often referred to by his first name) have been published in the past few years, including several here in BTN. But a new study by Tom Tippett, published on the web in July 2003, is perhaps the most comprehensive of them all.

Tippett starts by showing that some pitchers consistently yield a better ball in play average (IPavg) than others. Greg Maddux, for instance, beat the other pitchers on his team in this stat almost every year, while Andy Pettitte's IPavg was worse every year (see graphs from Tippett's article; the bars represent IPavg versus the league average (lower is better)).



"It goes without saying," Tippett then says, "that one cannot prove or disprove [Voros'] idea by examining only ten or twelve careers." Running a regression between a pitcher's past season and his current season, he finds a correlation coefficient of .09 for IPavg relative to the team. With an r-squared of .0081, the pitcher explains less than 1% of the variance in IPavg. That's small, but not zero.

Finally, Tippett notes that more than 12% of pitchers had IPavg marks that surpassed the 99% significance level – 12 times as many as you would expect by chance.

His conclusions? While skill in IPavg is not obvious, and is often drowned out by noise in the data, it does seem to exist. And since balls-in-play are very common, a small difference in IPavg can mean a big difference to a player's career. Tippett estimates that Charlie Hough, the all-time best in IPavg, saved more than 150 runs above an average pitcher, and "owes much of his career to his ability to excel in this respect."

Tippett's study is discussed in a (long) series of postings on the Baseball Primer website. As is usual for such things, many postings are interesting and insightful, while others are best ignored. But one point made seems especially relevant – the difference between flyball pitchers and ground-ball pitchers. If it's easier to get a hit on a ground ball, or vice-versa, that would make the pitcher type a very large influence on IPavg, independent of the skill of the pitcher. How much of that 1% of variation would a ground-ball/flyball factor explain?

Tom Tippett, "Can Pitchers Prevent Hits on Balls in Play?" *http://www.diamond-mind.com/articles/ipavg2.htm*
Baseball Primer, "Clutch Hits," *http://www.baseballprimer.com/clutch/archives/00008086.shtml#200*

Moneyball and the Stock Market

In another example of the ideas in "Moneyball" making it into common discussion, brokerage firm Morgan Stanley placed ads in various finance magazines entitled "Searching for the Financial Equivalent of a Walk."

In it, and in an expanded version available on the Morgan Stanley website, analyst Steve Galbraith notes that the Oakland A's success in fielding great teams with low salaries comes from understanding that walks are underrated, and stolen bases overrated. Applying the idea to finance, the authors ask: what is the financial equivalent of a walk or a steal? What factors, ignored or overemphasized by the market, can lead to greater success in stockpicking?

The stolen base, they argue, is growth. Investors choose the glamor stocks that are growing most quickly, just as scouts draft the "failed beefcake boys" with great bodies but minimal baseball skills.

Walks, on the other hand, are price metrics: price/earnings, price/sales, and price/book ratios – "boring old valuation." Instead of using "everything from ouija boards to price-to-eyeball valuation metrics to pick stocks," investors should use the price ratios. Backing up their assertion with data, they say that if an investor had bought stocks with the best ratios, while shorting (that is, betting against) the stocks with the worst ratios, "one might enjoy an early retirement."

While their analogy is entertaining, the specifics don't really hold up: as the authors say, you can choose stocks simply by P/E ratio – but if you try to choose players only on walks, you won't necessarily have a very good team. Runs Created is a better match for P/E – it summarizes a player's offensive contribution, just as P/E summarizes a stock's return on its stock price.

Still, the article is worth reading, especially for financially-oriented sabermetricians.

Steve Galbraith (with Mary Viviano and Frances Lim), "Searching for the Financial Equivalent of a Walk,"
http://www.morganstanley.com/ourviews/articles/walks_bases.pdf

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 Accuracy of Preseason Forecasts

Frederic Reamer

Do baseball writers and researchers do a creditable job of forecasting which teams will be successful? Here, the author examines several pundits' worth of preseason predictions and introduces a statistic that can be used to evaluate them.

Each spring produces a new crop of preseason forecasts by baseball reporters and other sports pundits. About a week or two before the first official pitch is thrown, prognosticators survey rosters, injury reports, and trade rumors, and speculate about the order of finish in each American and National League division. Page One of the typical newspaper sports section presents side-by-side comparisons of columnists' predictions.

Rarely, however, does one find any serious end-of-season accounting of the accuracy of these forecasts. An exception was the ESPN.com column (October 30, 2002), which reprinted the forecasts made by ESPN.com's baseball staffers at the start of the 2002 season and compared them with the season's final standings. As a measuring rod, ESPN.com assigned one point for each place missed in the final standings.

After reading the ESPN.com column, I realized that an adaptation of a well known statistical procedure would provide a much more sensitive and robust indicator of the accuracy of preseason forecasts and facilitate comparisons among various staffers and publications. I think this approach – which I have dubbed the *Prediction Accuracy Index* – can be used broadly whenever sports columnists predict the final standings in any sport.

The statistic on which my approach is based is known as *Spearman's rho*. It is one of many statistics that researchers use to analyze quantitative data. This particular statistic is known in the trade as a nonparametric measure and provides a clear, direct indication of the extent to which there is a linear relationship, or correlation, between two sets of ranked (or ordinal) scores. Put more simply, Spearman's rho allows one to determine the extent to which two sets of ranks of the same phenomenon – in this case, the preseason and postseason ranks of each baseball division's teams – are similar. Did the final standings resemble the preseason forecast or not?

Based on a mathematical formula that is relatively easy to calculate, Spearman's rho is a number, or coefficient, that is very intuitive and easy to interpret. The final coefficient is a number between +1.0 and –1.0.[1] If the two sets of ranks (e.g., preseason and final standings) are identical, Spearman's rho would be +1.0 – that is, if the postseason order of finish is identical to the preseason forecast. If the two sets of ranks are completely reversed, the Spearman's rho would be –1.0: that is, if the team that the columnist thought would finish first actually finished last, the team the columnist thought would finish last actually finished first, and so on up and down the line. Spearman's rho will equal zero, or will be close to zero, if the results are very mixed, i.e., some predictions were on the mark, some were close, and some were way off the mark.

[1] The formula for Spearman's rho, which can be found in any standard statistics textbook, is:

$$1-\frac{6\sum D^2}{n(n^2-1)}$$

where D is the difference between each pair of ranks and n is the number of items (i.e., teams) ranked. In our case, each team in the preseason forecast would be ranked from first to last (rank=1, rank=2, and so on). These same numbers or ranks would then be listed based on the actual postseason finish and D is calculated by subtracting the postseason rank from the preseason rank. For example, if the predicted order of finish were: Yankees, Red Sox, Blue Jays, Orioles, and Devil Rays, the Yankees would be assigned rank #1, the Red Sox rank #2, the Orioles rank #3, and so on. If the Blue Jays finished in first place, D would equal +2, which is the result of 3 (the predicted place in the final standings) minus 1 (the actual place in the final standings). Some D scores will be positive (when teams do better than predicted) and some will be negative (when teams do worse than predicted). D will equal zero if the team's place in the final standings is exactly as predicted, no matter where in the standings the team falls. Each D score is squared (to avoid the problem of working with positive and negative numbers), and the sum of these squared D scores is multiplied by 6 (the mathematical reasons for this are complex). This result is divided by the mathematical product of the number of ranked teams (n) and the total of the number of ranked teams squared (n^2) minus 1 (again, the mathematical reasons for this are complicated and can be read in any traditional statistics textbook). This number is then subtracted from 1, and the result is the Spearman's rho coefficient, which ranges from +1.0 (a perfect direct correlation) to –1.0 (a perfect inverse correlation). For an overview of Spearman's rho, see R. P. Runyon and A. Haber, *Fundamentals of Behavioral Statistics*, 2nd ed. (Reading, Mass.: Addison-Wesley, 1971), pp. 102-104.

For illustrative purposes, I compared the accuracy of the preseason forecasts (2002 season) of five ESPN.com staffers: Jayson Stark, Rob Neyer, Jim Caple, Matt Szefc, and Sean McAdam. To compute the *Prediction Accuracy Index* for each staffer, I began by calculating the Spearman's rho coefficients for each staffer's predictions for each of the three divisions in the American and National leagues. Thus, I calculated six Spearman's rho coefficients for each staffer. Positive coefficients are better than negative; i.e., they indicate more accurate preseason forecasts.

For example, Matt Szefc's six Spearman's rho coefficients were:

AL East	+1.0	perfect prediction of the final standings
AL Central	+0.6	a pretty good result – what hurt Szefc most was that he predicted that the Twins, who finished in first place, would come in third
AL West	+0.4	a fair, but not great result – Szefc predicted that the Mariners would come in first, but they came in third
NL East	0.0	the results were quite mixed – Szefc accurately predicted the Braves' first-place finish, but he predicted that the Expos, who finished second, would come in last; he predicted the Mets would come in second, but they came in last
NL Central	+0.54	a good result – what hurt Szefc the most was that he predicted that the Cubs, who finished fifth, would come in second
NL West	+0.6	a good result – Szefc predicted that the Padres, who finished in last place, would come in third

The sum of these coefficients is +3.14 (out of a maximum total of +6.0).

An easy way to compare the accuracy of each staffer's predictions is to calculate a final score – the *Prediction Accuracy Index* – based on the sum of each division's Spearman's rho coefficients (adding up both positive and negative coefficients) divided by +6 (the largest possible total if all preseason forecasts perfectly match the final standings). A set of perfect predictions would generate a *Prediction Accuracy Index* of +1.00 (+6 divided by +6); a set of perfectly inaccurate predictions (the final results are the reverse of the predictions) would generate a *Prediction Accuracy Index* of –1.00 (-6 divided by +6). Positive *Index* scores (scores between 0.0 and +1.0) are better than negative *Index* scores (scores between 0.0 and -1.0); scores closer to +1.0 are the strongest and scores closer to –1.0 are the weakest. Scores around zero are truly mediocre.

One can also compare the staffers' *Prediction Accuracy Index* scores with a final score corresponding to what are commonly referred to as "naive predictions," that is, the extent to which the final standings at the end of one season (in this case 2001) accurately forecast the final standings at the end of the following season (2002). To compute the final score for naive predictions, we run through the same process, simply treating last year's final standings as this year's prediction.

In Szefc's case, the *Prediction Accuracy Index* is +0.52, indicating a good, but not outstanding set of predictions (+3.14 divided by +6.0). One can then compare this *Index* score with the *Index* scores of all other staffers who made preseason predictions; the higher the score the more accurate the prediction. For the ESPN.com staffers and naive predictions, the results for each league[2] and overall are:

	AL	NL	Combined
Jayson Stark	.70	.53	.62
Sean McAdam	.73	.45	.59
Rob Neyer	.63	.47	.54
Matt Szefc	.67	.38	.52
Jim Caple	.67	.35	.51
Group Average	.68	.44	.56
Naive Predictions	.67	.56	.62

The *Prediction Accuracy Index* facilitates easy comparisons among staffers and the naive predictions. Jayson Stark was the most accurate prognosticator and Jim Caple was the least.[3]

[2] The American and National League scores were calculated by adding the three Spearman's rho coefficients for each league (one coefficient for each division) and dividing by 3 (the largest possible total if the preseason forecasts match the final standings).

[3] The results produced by this statistical procedure are different than the results produced by ESPN.com's analysis, which was based only on how far each *individual* team's actual finish was from the predicted finish (as reported in the ESPN.com column, each place missed in the final standings was worth one point). The ESPN.com approach does not take into consideration the overall *pattern* among the ranked scores when the two sets of ranks are compared. The advantage of my approach is that it takes into consideration the overall *pattern* of ranks and compares the *total* preseason forecast with the *total* postseason results.

Overall, the group's preseason forecast was good, but not outstanding (+.56). Also, the group's preseason forecasts for the American League (+.68) were much more accurate than for the National League (+.44). The naive predictions for the American League (+.67) were virtually identical to the staffers' group average (+.68), but the naive predictions for the National League (+.56) were considerably better than the staffers' group average (+.44). The overall final score for the naive predictions (+.62) was somewhat higher than the summary *Index* score for the group of staffers (+.56). None of the staffers had an *Index* score that was higher than the overall score based on the naive predictions.

The *Prediction Accuracy Index* provides a straightforward, precise, and intuitive measure of the accuracy of preseason predictions in any sport and at any level.

Frederic Reamer, Ph.D., School of Social Work, Rhode Island College, Providence, RI, 02908, 401-456-8248, freamer@ric.edu ♦

Corrections To This Issue

Due to an editing error, the formula in the footnote on page 6 was not correct in the original printing of this issue. It is corrected in this current printing.

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

Fair-Weather Fans

Darren Glass

Rob Neyer once wrote that attendance relates more to the success of the team than whether they play in a "good baseball city." Here, the author examines attendance and performance data to see if that's really the case.

In Rob Neyer's chapter on San Francisco in his *Big Book of Baseball Lineups*, he speculates that there aren't really good baseball cities, and that attendance is more closely correlated with winning percentage than with any other factor. He also suggests that a statistically-minded person look at this. I took the challenge and have been playing with a lot of data.

Methodology

I looked at all seasons from 1973 until present. In particular, I looked at the correlation coefficients between the following variables:

- Average home attendance per game (ATT)
- Home attendance per game divided by Average Home attendance over all teams (to normalize for nation-wide trends) (ATT/AVE)
- Final place in divisional standings (PLACE)
- Winning Percentage. (WIN)

All data came from www.baseball-reference.com.

Correlation with Winning Percentage

To begin with, let us look at the most naive study: the correlation between winning percentage and home attendance. Over the 30 years between 1973 and 2002, the baseball-wide CC was .464.

The following teams can be described as having fair-weather fans – their correlation between winning and attendance is more than .2 greater than the baseball-wide average:

Atlanta	0.884
Seattle	0.815
New York N	0.786
Cleveland	0.755
Montreal	0.753
Chicago A	0.752
San Francisco	0.673

On the other side of the spectrum are those teams that have correlation coefficients significantly lower than the baseball-wide average. An optimistic interpretation of this would be that the fans stick with the team no matter how badly they are doing (the case of the Red Sox and the Cubs), while a pessimistic interpretation might be that the fans refuse to support the team no matter how good they are.

The following cities have correlation coefficients between ATT and WIN more than .1 below the MLB average:

St Louis	0.345
Chicago N	0.321
Texas	0.304
Tampa Bay	0.266
Milwaukee	0.234
Arizona	0.142
Pittsburgh	0.131
Los Angeles	0.117
Boston	0.004
Colorado	-0.087
Florida	-0.118
Baltimore	-0.246

We note the presence of all four of the 90's expansion teams on this list. This makes sense, as the already small sample size is severely distorted by the first few years in which the city is likely to be more excited and the teams are likely to be not very good.

The most interesting datapoint on this list to the author is the Orioles, where the fans of Baltimore actually supported the team significantly more the worse they have been over the past 30 years. This is likely due in large part to the draw of the "new" ballpark at Camden Yards, and the fact that it has been successful in bringing in fans despite the fact that the Orioles have had losing records in 6 of the 11 years since it opened.

A slightly less naive study would try to normalize for the effects on attendance of baseball as a whole. The average attendance at baseball games has nearly doubled over the last 30 years, and all of baseball took a hit in 1995 when the average attendance dropped nearly 6000 fans per game. Thus, I also computed the CC's between ATT/AVE, a given team's average home attendance divided by the average attendance of baseball games league-wide, and winning percentage. The data did not qualitatively change significantly. The league-wide CC went up to .55. Table 1 shows this CC for all teams.

Table 1: Correlation Coefficients between ATT/AVE and WIN for all 30 teams

Atlanta	0.925
Cleveland	0.832
Seattle	0.786
Philadelphia	0.753
New York N	0.752
Cincinnati	0.724
San Francisco	0.713
Oakland	0.692
Detroit	0.691
Kansas City	0.677
Minnesota	0.667
New York A	0.598
Tampa Bay	0.596
San Diego	0.573
Los Angeles	0.563
Montreal	0.557
Chicago A	0.541
Pittsburgh	0.539
Boston	0.532
Chicago N	0.520
Houston	0.505
Texas	0.489
St Louis	0.485
Toronto	0.478
Milwaukee	0.433
Anaheim	0.387
Colorado	0.303
Arizona	0.079
Florida	-0.035
Baltimore	-0.092

Statisticians say that a Correlation Coefficient is statistically significant if it is greater than the value of a certain T-test. While I will not go into the details of this calculation, I will point out that for our sample size of 802 team-seasons, any CC over .116 is statistically significant with probability 99.9%. In particular, our league-wide CC of .55 is extremely significant.

For the individual teams, our sample sizes are much smaller. In particular, for the non-expansion teams we have 30 data points, and thus a CC over .570 will be statistically significant 99.9% of the time, a CC over .463 is significant 99% of the time, and a CC over .361 is significant 95% of the time. When we include the expansion teams with even smaller sample sizes, we see that the CC's are significant at the 99% level for every team except Milwaukee, Anaheim, Baltimore, Toronto, Tampa Bay, Arizona, Colorado, and Florida.

Of course, the CC is not enough to capture what we are interested in. In particular if a city's ATT/AVE and WIN were strongly correlated to a line with slope zero we would view it as much less of a "Fair Weather Fan" city than a city where you had a weaker correlation to a line with a very large slope. Thus, I also computed the slope of the line given by various linear regressions Baseball-wide, if we run a linear regression on ATT/AVE and WIN we get that ATT/AVE = 2.7525 * (WIN) -.3769. We note that while ATT/AVE is a more statistically meaningful statistic then it is also harder to get a feel for. For this reason we will note that the linear regression between ATT and WIN gives ATT = 63,476*WIN – 7740. In other words, by increasing your winning percentage by .100 (an improvement of roughly 16 wins in a season), a team could be expected to boost its home attendance by an average of 6347 fans per game.

See Table 2 for the slope calculation for the individual teams.

A natural question to ask, and one that more than a few people are looking at due to its various political implications, is how new stadiums affect attendance. While I did not investigate this phenomenon in any depth, I will note that if you remove all datapoints in the dataset corresponding to the first two years that a team is in a new city or a new stadium then the baseball-wide CC actually rises by .05.

Correlation with Standings Position

It is also natural to wonder if it is not the winning percentage that brings in the fans, but being in the hunt of a pennant race. I decided to test this hypothesis by calculating the correlation coefficients between our attendance variables and the place in which a team finished within their division, as well as how many games back they finished. Because the nature of both of these variables changed significantly with the realignment in 1994, I ran the study first looking only at the data from the years 1973-1993. In particular, it was not clear how to best handle the situation with the wild card, and teams that might be in the hunt for the wild card despite being many games out of the division lead (see 2003 Phillies and Marlins, for example). It came as a surprise to the author that including the last decade did not significantly change the results, as seen by the following charts:

1973 - 1993

Regression	Correlation	Slope
ATT/AVE vs. PLACE	-.559	- 0.105
ATT vs. PLACE	-.4632	-2136.5
ATT/AVE vs. GB	-.53	- 0.0164
ATT vs. GB	-.4535	- 343.129

1973-2002

Regression	Correlation	Slope
ATT/AVE vs. PLACE	-.559	- 0.0978
ATT vs. PLACE	-.5016	-2491.01
ATT/AVE vs. GB	-.4906	- 0.0145
ATT vs. GB	-.4131	- 334.6898

Note that in all of these examples, CC is negative. This is what we would expect as the “higher” your value of PLACE and GB the less attendance we might expect to see.

I have not included the team-by-team data, but it is qualitatively very similar to the above team-by-team data, with the teams falling in roughly the same order and with the same significance results. Anyone who is interested in the full data should feel encouraged to e-mail me.

Correlation with Past Performance

Another question that comes up is how correlated attendance is with past performance -- in particular, looking at the correlation between winning percentage (or standings) in year x and attendance in year (x+1). The idea would be that the rush of winning the World Series creates new fans (and season ticket holders) no matter how badly the team performs the following year.

However, when one runs the numbers they are not particularly illuminating. In fact, the CC's one gets from comparing last year’s winning percentage and this years ATT/AVE is .492, slightly less than when you compare this year's record with this year's attendance, .551. (See table below). Furthermore, the only teams for which there is a substantial difference in the CC's when you run the study the two ways are Colorado (which can be partially explained by the fact that you had a small data set to begin with and are reducing it even further), Minnesota, Montreal, Pittsburgh, and St Louis. Furthermore, in each of these cases there is a weaker correlation. So while my instincts agreed with what many of you suggested might be an interesting effect, the numbers don't seem to bear it out.

Table 2: Slopes calculated from Linear Regression between ATT/AVE and WIN

Cleveland	4.543672
Philadelphia	4.290944
Atlanta	3.850382
Cincinnati	3.735552
Los Angeles	3.431718
Seattle	3.328853
San Francisco	3.206009
New York N	3.134074
Kansas City	3.067628
Minnesota	2.862508
Montreal	2.772461
Oakland	2.403002
Chicago A	2.214931
New York A	2.202218
Detroit	2.186652
Houston	2.157452
Toronto	2.114608
Boston	1.920404
Anaheim	1.917665
Colorado	1.88844
San Diego	1.858157
Texas	1.775284
St Louis	1.746337
Chicago N	1.634861
Tampa Bay	1.578699
Pittsburgh	1.374932
Milwaukee	1.304664
Florida	-0.1523
Baltimore	-0.37538
Arizona	-0.99382

	Win	Prev Win	Place	Prev Place
Att/ave	0.5505	0.4926	-0.5016	-0.4651
Att	0.464	0.4293	-0.4669	-0.4329

One problem in trying to do such a study is that there is a relatively strong correlation between how a team does in year X and how it does in year x+1 (CC = .5 for my data set). Isolating that factor would be difficult but not impossible.

Conclusions

Every one of the tests which I ran seems to indicate that Rob Neyer's hypothesis is correct: attendance at ballgames is highly correlated with the winning percentage of the home team. This is certainly true baseball-wide, and is also true for almost every team individually. The exceptions by and large are the expansion teams of the nineties and the Baltimore Orioles. Furthermore, in almost every permutation of the data, it seems that the fans of Cleveland, Atlanta, and Seattle are especially prone to support their teams more the better they do. We do note, however, that all three of these teams got new stadiums while the teams were doing especially well – and in the case of the Braves and the Indians this was also at a time when baseball was seeing a drop in attendance nationwide – which likely skews the data somewhat.

Further Explorations

I think it would be very interesting to look at attendance in smaller units than seasons. This could take away some of this effect by looking at when in (for example) the 1991 seasons the fans stopped punishing the Braves and Twins for previous mediocrity and rewarding them for being good.

However, to do this one would have to control for factors such as weekend games (which generally have higher attendance) or superstar players coming through town (which certainly boosts attendance) or the like, factors which one can ignore over the course of a season but which could significantly affect the data when looking at units of individual games or weeks or even months.

Another thing that I would like to do is to try to adjust for ballpark size. The only way I could think of to do this would be to use "prcentage of seats filled" as my attendance variable, but this seems to pose more problems than it solves. I certainly like the idea of "rewarding" the Cubs and Red Sox and other teams which could sell more seats if they had the capacity, but I'm not sure if it makes sense to "punish" cities for having large stadia in this way. For example, if Stadium One holds 50,000 people and Stadium Two holds 60,000, I do not think that it makes sense to treat the fact that they both draw 30,000 fans differently. It also seems like a bit of opening Pandora's Box as we really don't know how many fans the Red Sox would average if they had an infinitely big stadium. It could be that their attendance would stay the same or it could be that it would quadruple -- we have no real way of knowing.

Darren Glass, 601 West 113th Street Apt. 6K, New York NY 10025; glass@math.columbia.edu ♦

Clutch Hitting and Statistical Tests

Dan Levitt

In another sabermetric attempt to find elusive evidence in favor of the hypothesis that some players have specific clutch-hitting skill, the author subjects a seven-year set of data to several statistical tests.

Introduction

Although nearly all observers in the baseball mainstream believe that some players are “clutch hitters”, baseball analysts who have examined the data are essentially unanimous in that they can find no evidence of a unique ability to hit in the clutch. Obviously the best hitters to have up in clutch situations are the best hitters overall. The question is whether some players can elevate their ability while others falter in clutch situations. In other words, do some players have a particular clutch hitting ability (or lack thereof), or is the appearance of clutch ability simply a random fluctuation such that over the long haul all players will hit equal to their overall ability in the clutch?

This paper takes another look at the issue of clutch hitting ability by subjecting the data to three statistical tests. The data I am using to measure clutch hitting ability is batting average in "Close and Late" (C&L) situations as defined and published by STATS Inc. for the years 1992 through 1998. In this analysis I am using their definition of Close and Late situations as a proxy for clutch situations.[1]

Batting average may not be the best statistic to use (as opposed to OPS or some other more all-inclusive value measure), but it has the advantage of being easier to analyze. That an at-bat can only have two mutually exclusive outcomes, hit or out, allows for several relatively simple statistical tests. Additionally, because we are evaluating the difference between C&L and all other (NonC&L) situations, as opposed to overall ability, the limitation of batting average is reduced. It seems to reasonable to assume that if clutch hitting ability exists, for the most part it affects batting average similarly to the other elements of batting.

I looked at data for the seven years 1992 through 1998. During that period 40 players had at least 350 at bats (just over half of a full seasons worth) in each season. I evaluated these 280 player seasons several ways. For background, note that from 1992 through 1998, the 40 player average in C&L situations is less than that overall:

	AB	H	BA
C&L	22357	6180	.276
NonC&L	123105	35872	.291
All	145462	42052	.289

There are probably several reasons for this difference, such as the appearance of the best relief pitchers in these situations, and the opposing manager using his relievers to get a platoon advantage.

A number of statistical tests exist which can be used to tease meaning out of this numerical information. I organized the data for three statistical tests: a year-to-year comparison for a binomial test, a look at each player's data over the seven year time frame for a Chi-square test, and a check of the correlation among the players over time. A secondary purpose of this paper is to highlight several statistical tests as a way to both bring them to the attention of other researchers and stimulate discussion on which might be the most applicable for this and other types of baseball data.[2]

Test 1 – Binomial Organization of the Data

The binomial distribution can be used to examine data that consists of separate trials with two possible outcomes, such as flipping a coin. I defined each trial to be whether a player had a higher average in C&L situations or NonC&L situations on an annual basis. I then counted how many times each player had a higher batting average in C&L situations over the seven seasons. The task then is to determine whether this set of “trials” is significantly different from what one would expect simply by chance.

[1] STATS Inc. defines "Close and Late" as occurring when (a) the game is in the seventh inning or later and (b) the batting team is either leading by one run, tied, or has the potential tying run on base, at bat, or on deck.

[2] Thanks to Michael Schell for reviewing this paper, and to Jim Box for reviewing an earlier draft.

For this analysis, the initial step is to estimate the probability that a player exceeds his NonC&L average in C&L situations, taking into account the fact that, overall, players have lower batting averages in C&L situations.

The average number of at bats in C&L situations was 80, the overall batting average for the 40 players in NonC&L situations was .291, and the overall batting average for the 40 players in C&L situations was .276. Therefore, one needs to estimate the probability a player gets more than 23.3 hits (.291 * 80) in C&L situations, given an overall average in C&L situations of .276 ([P(x>=23.3), p=.2764] in probability nomenclature). I looked at three ways to calculate this probability, the first two based on the Binomial distribution.

Method 1.1

Using MS Excel's BINOMDIST function, one can estimate the probability of 23 or more hits (for this calculation one needs to round the 23.3 to an integer). The easiest way to calculate this probability is to first calculate the probability of less than or equal to 22 hits, and then subtract the result from one.

Following this procedure gives a probability of .427.

As a check of the expected versus actual, note that in .39 of the player seasons (110 of 280) the C&L batting average exceeded the NonC&L batting average. This is close to our theoretical figure of .427.

Method 1.2

Because the data is fairly continuous and therefore one can use the more precise 23.3 hits as opposed to the rounded 23.0, I also looked at the Normal distribution approximation of the Binomial distribution.

First, we calculate [Pr (x<=23.3)], then subtract the result from 1.

```
Mean = u = np = 80*0.276 = 22.07
Standard Deviation = sqrt(npq)= sqrt(80*.276*.724) = 4.00
```

Using the MS Excel NORMDIST function, the probability of a value less than or equal to 23.3 given a mean of 22.07 and a standard deviation of 4.00 is .618.

Thus, the probability we seek is 1 minus .618, which equals .382.

Again, our observed figure of .39 is very close to the .382 theoretical probability.

Method 1.3

The next step of a binomial analysis involves using the probability just estimated to analyze the data. I assumed the probability that a player has a higher batting average in C&L situations is .382 as determined above. Because this method used a hit level of 23.3, rounding, and thus in effect altering the overall batting average, was avoided. This value is also fairly similar to that generated from the actual results.

Each player had seven seasons of data. Suppose we look at each season separately. Then, for any given player , he could have had his C&L average exceed his nonC&L average in zero of the seven seasons. Or, his C&L could have been higher in one of the seven seasons. Or two of the seven, or three, or four, or five, or six, or all seven.

For each of the 40 players, I counted how many times, out of seven, his C&L average exceeded his nonC&L average. The results are below:

Number of Seasons C&L beat nonC&L (out of 7)	Theoretical Probability of this happening	Theoretical number of times out of 40	Actual number of times out of 40
0	.034	1.4	1
1	.149	6.0	7
2	.276	11.0	12
3	.285	11.4	9
4	.176	7.0	6
5	.065	2.6	3
6	.013	0.5	1
7	.001	0.048	1

We can use statistical tests to draw conclusions from the comparisons in the table.

First, I evaluated the data using a Goodness of Fit/Chi-square test. This test allows one to compare the actual number of seasons in each row with the expected number to see if the difference is significant. Because several rows had an expected number of less than 5 (which I understand is the minimum for a valid Chi-square test), I combined the 0 and 1 rows to get an expected number of 7.3, and the 4 to 7 rows to get an expected number of 10.2.

To get the Chi-Square statistic, we sum

$$\frac{(Actual - Expected)^2}{Expected}$$

for each row, and then sum all the rows. This table summarizes the calculation of the Chi-square statistic:

Rows	Actual	Expected	(Actual-Expected)^2/Expected
0-1	8	7.3	.06
2	12	11.0	.08
3	9	11.4	.50
4-7	11	10.2	.06
Chi-Squared (sum of above)			0.70

For this data to be significant at the .05 level (that is, only a 1 in 20 likelihood that it was generated by chance) the Chi-square statistic would have to exceed 7.81. The resulting Chi-square value of 0.70 is well below this, and thus, based on this test, one cannot conclude that unique clutch ability exists.

When looking at the last row of the first table, however, at least the possibility of specific C&L ability emerges. As the third column shows, the probability of finding at least one player in the “7” row – a player who had a C&L average greater than their NonC&L average in each of the seven years -- is only 4.6%. Thus, the fact that one was actually found, suggests the possibility of statistical significance for that player's C&L average at the .05 level. That player was Tony Gwynn.

Test 2 – Chi-Square Organization of the Data

Another way to analyze the data is to calculate the seven year totals for each player and compare how they did in C&L situations to how they hit in NonC&L situations over the entire time period. The Chi-Square test is the statistical test used to determine if the difference between the two batting averages are meaningful. The method essentially evaluates the difference between the expected C&L average (calculated using the NonC&L average) and the actual C&L average.

As noted above, for the entire 280 seasons of data (145,462 at bats), players hit worse overall in C&L situations (.276 versus .291). The Chi-Square test confirms that this difference in batting average is statistically significant. In other words, overall, one can say that batters hit worse in C&L situations.

Consequently, to analyze the data using the chi-square test, the expected values need to be adjusted to incorporate the fact that on average players hit .013 less in C&L situations. What I did, then, was to simply estimate the player's "expected" C&L average to be 13 points less than his overall average. For instance, Mark Grace hit .317 overall in the period of the study. We then subtract 13 points to expect him to hit .304 in C&L situations. Since he had 609 C&L at-bats, we would expect 185 hits in C&L situations.

In fact, Grace had 213 hits in those situations. The Chi-squared probability of observing that discrepancy is .0075.

A similar test for Marquis Grissom gives a p-value of 0.36. Here is a summary of these two hitters:

Mark Grace	Actual, C&L	Actual, nonC&L	Expected, C&L	Expected, nonC&L	Chi-Squared probability
Hits	213	1007	185	1035	
Outs	396	2233	424	2205	
					0.0075

Marquis Grissom	Actual, C&L	Actual, nonC&L	Expected, C&L	Expected, nonC&L	Chi-Squared probability
Hits	177	970	168	979	
Outs	448	2485	457	2476	
					0.36

At first glance, it appears that Mark Grace's value is significant, while Marquis Grissom's is not. But over a large group of players, P-value significance is not quite that simple. For example, in a sample of 40 players, one might randomly expect to find two players with a P-value less than .05. Thus if there are several players for whom the data appears statistically significant, one needs to further check the probability of finding such P-values in the sample of players.

Only two players out of the 40 had P-values less than .05. The probability of finding 2 or more players with P-values below .05 is .60, well above a .05 threshold of statistical significance for the group of players.

The two players, however -- Mark Grace shown above at .0075, and Tony Gwynn at .0030 -- had extremely low P-values (well below .05). The probability of finding two or more players with P-values below .0075 in a sample of 40 is .037, a result that is significant at the .05 level. In other words, at least some indication exists that certain players may perform outside their norm in Close and Late situations.

Test 3 – Analysis through Correlation

If clutch hitting is a skill in the same way as hitting for power or speed then one ought to see some consistency over time for this skill just as one would expect to see from a power hitter. In this analysis, I calculate the difference between each hitter's batting average in C&L and NonC&L situations for two time periods, 1992 – 1995 and 1996 – 1998. If clutch hitting is indeed a skill held by particular players then one would expect, in general, the players to rank similarly over the two time periods. One way to measure the similarity between the two time periods is with a correlation coefficient.

The correlation coefficient for the 40 players is .13, a value close enough to 0 to lead one to believe that no true correlation exists.

A test exists to determine if this correlation is significantly (in a statistical sense) larger than zero, and thus indicative of a correlation in clutch ability over time. A t statistic can be calculated which can then be compared to Student's distribution to test for significance.

For the correlation coefficient r, and a sample size of N,

```
t = r*sqrt(N-2)/sqrt(1-r^2)= .13*sqrt(40-2)/sqrt(1-.13^2)
  = .81
```

On the basis of a one-tailed test at the .05 significance level, the calculated t-value would have to be larger than 1.69 to conclude that the correlation was significantly different from zero.

Conclusion

Based on the three tests here, one would have a hard time claiming that how a player hits in the clutch is a skill like hitting home runs or for a high batting average. When looking at the data distribution over all the players—-such as in Methods 1.1 and 3—-no evidence of any clutch hitting ability shows up. It should noted, however, that a couple of players exhibited a batting average in C&L situations that shows up as significant when looked at individually as in Methods 1.2 and 2. I would not necessarily conclude that Gwynn or Grace had the skill to materially increase their ability in the clutch, but they at least suggest further research may prove fruitful.

Part of the problem in looking for clutch hitting is the small sample sizes for each player. I have no doubt that players vary, relative to their baseline performance, in the clutch, but this variation is extremely small when compared to all the other player differences. The problem is that this skill is a minute part of all the elements going into the batter/pitcher matchup.

For example, assume a hitter's "true" batting average is .300, and that he can elevate this to .306 in the clutch due to an extra ability to focus. The complication is that one could never tease this skill out of the data. Assuming 80 C&L at bats a season, .006 in batting average is one extra C&L hit every two years. There is no way to segregate this clutch hitting ability from all the other factors, including luck.

The evaluation of clutch ability is further complicated by the fact that no true definition of a clutch situation exists. C&L is probably a pretty good approximation, but it is far from perfect. Furthermore, one could argue that clutch situations should be defined along a sliding scale that would weight many different factors such as the importance of the game in the pennant race, the number of runners on base, the number of outs, etc.

The secondary purpose of this essay was to stimulate some discussion as to which statistical methods to apply to various baseball questions. Are there other tests I should have run on the data set? I am always surprised at multiple ways data can be analyzed, and I hope this paper causes further discussion of some of these as they relate to baseball.

Dan Levitt, danrl@attglobal.net ♦

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.

An MVP Voting Model (Part I)

Tom Hanrahan

What do sportswriters consider when voting for Most Valuable Player? The author examines voting results and player data to try to find out, building on a previous study of this topic from 1999.

Introduction

I want to answer the question "What have sportswriters considered when casting MVP votes?" This is the same question asked by Rob Wood in his article "What drives MVP voting?" (*BTN*, February, 1999). My approach will build on his work.

This will enable us to look at many issues, such as:

1. Which players were perceived by the BBWAA to be more valuable than their statistics suggested?
2. How does "value" as measured by an MVP voting model correlate with "value" in terms of sabermetric measurements?
3. How have voting patterns changed over time?
4. Who might win the awards this year?

In this article, aptly named Part I, I will look primarily at how the model was created and fine-tuned, and take a peek at the 2003 season. Part II will focus on the other questions asked above.

An obvious approach to the problem is to collect as much data as practical for all of the seasons of MVP voting, and see what factors correlate with winning the award. Having begun this process, I ran into many questions, as it became apparent that there is no "best" way, only many possible good ideas.

Building the Model

How far back to go?

MVP voting has occurred regularly since 1931. However, the current voting structure has only been in place since 1938. I chose this later year as the cutoff date, and used all years through 2002.

Static or changing patterns?

More data are better, but it also possible that voters' logic has changed over the years, so after drawing initial conclusions with respect to the whole data set, it would be wise to re-inspect the conclusions to see if they hold over differing periods. The desire to notice trends changing over time must be balanced with the problem of drawing from too small of a sample. In the end, I broke the 65 years into 3 almost-equal time periods.

It is possible that AL and NL voters could use different criteria. I did not test for this.

Pitchers vs. Hitters?

Few pitchers win MVP awards. Attempting to compare stats of the few pitchers with all of the hitters' stats is, in my view, just plain nuts. I have chosen to ignore pitchers. When any hurlers finished among the top MVP vote-getters, I used the next lower players.

How to best fit the data?

One basic tool would be a regression analysis. Ideally, a model would usually predict the winner, and be roughly accurate in the placement of the other top finishers as well. So, how many of the top finishers each year should be used? If I use only the top 2 per year, the analysis will focus on how often the actual winner finished ahead of the actual runner-up according to the model. This may miss a player who, according to the model should have won, but since he finished 3rd or lower was not included in the data. Adding in additional players will give more robustness to the data, but has the drawback of trying to make the regression fit the top N finishers equally, when in reality I am more interested in the top of the ballot than what happened lower. By some logical reasoning and guesswork, I wound up using only the top 3 (non-pitching) finishers on each ballot.

I adjusted the actual MVP points received by each player so they were normalized to a 12-team league, with 2 voters per team. So, a perfect score became 336 points (14 for a first place vote, times 24 voters).

At this point, let me stop and say that in an exercise such as this, there is a balance that must be chosen between simplicity and accuracy. The simplest MVP model might be, to quote Bill James in one of his early *Abstract*s, "the big RBI man on a pennant winner." This in fact would predict the winner reasonably often. A bit less crude might be "AVG + HR + RBI + some # of bonus pts for winning." The other extreme is to take 100 pieces of data for every MLB player, and create a formula using all of these. But then, since some factors work only in combination, you would have to test every multi-factored combination of these (which would be 100*99/2 = 4950 combinations), and come up with a formula that has way more coefficients than your Excel spreadsheet has columns. And that's only linear regression; we haven't touched polynomial, logarithmic, etc. Also, it is easy to fine-tune the formula to make it work better in reverse. If the formula underestimates by a small amount on players who played third base in the 1970s and 80s, hey, add a small coefficient for third basemen during those 20 years, and it's "Better!" I have chosen, rather, to use the outputs of the regression analysis of the top three finishers and not tweak the results just so I can produce more "winners" from the formula.

What Data to Use?

The model will only use statistics accumulated at year's end. Surely there are many factors besides year-end stats that are in voters' minds when creating their ballots. However, these are not readily available, and the amount of "other" data and how it could be combined is staggering. For example, year-end stats cannot capture Willie Stargell's "fam-i-lee" leadership in the late summer of 1979. Rather, one of the main points of building and using the model is to see which players (if any) under- or over-performed their year-end stats; and thus, in voters' eyes, were more or less valuable than their "raw" statistics that are etched in history for all to easily inspect.

Obviously, I should test virtually every statistic that has received attention as being "valuable." For example, players' stolen base totals are often shown, but caught stealing data is not. I made the decision not to enter CS as a variable. Stats that had little attention (GIDP) or little value because of their infrequency (HBP) were not used.

Some statistics lend themselves to analysis by a sliding scale, more or less in linear fashion. Surely all voters in the past 65 years have noted that a player's batting average is important, and that hitting .350 is more valuable than .290, which is better than .230. I believe the same holds true for the other Triple Crown stats that have been with us since, well, since someone thought of the Triple Crown. So, for Home Runs, RBI, and Average, I used the players' totals. I also used stolen bases this way, since this also is a visible stat. In order to account for different eras (big hitting versus pitching-dominated years), I actually modified each player's stats by the 3rd place finisher in each category in each league. So, if a player hit .322 and the 3rd place finisher in batting average that year hit .318, he gets a +4 for that year. If I did not do this, the model could not possibly predict that Bill Freehan in 1968 (a pitchers' year) would do better than Mark McGwire in 1998. This method allowed me to compute an r^2 value for the regression equations as I compared players in different league/seasons with each other, since it normalized much of the data.

Other stats might be important because they are league-leading figures. I cannot tell you how many triples any player hit last year, but I at least seem to remember that Willie Wilson used to lead the AL quite often, so employing some of these types of stats as binary "dummy variables" in the regression seemed like a good idea. If a man's name pops up on the top of the leader board in categories that are often mentioned, this certainly might be important in voters' minds. Another reason I looked at league-leaders was that in Rob Wood's previous study, he correlated the categories a player led the league in with MVP voting. He found that leading the league in some metrics (such as RBI and slugging percentage) was a good predictor of MVP voting success. While this study goes into many areas besides the ones looked at by Mr. Wood, the groundwork he laid was a solid foundation on which to build. I hope to compare his results with this model's in part II of this article. The list of league-leading dummy variables I tested included the following: AVG, OBA, SLG, D, T, HR, hits, TB, runs, RBI, SB, and BB.

If leading the league in a statistic is important, then it also may be valuable in voters' eyes if a player finishes near the lead in that category. Of course, we would expect the effect of finishing second to be much smaller than finishing first. So, only for those league-leading statistics that tested to being highly relevant, I chose to also look at the 2nd and 3rd place finishers in that statistic also.

It turns out that for AVG and HR, there was very little evidence for the importance of ordinal or rank finishes; rather, voters seemed to rely on the difference between the values. I was a bit surprised by this, but after considering players such as Dave Kingman and Tony Armas who at times excelled in some areas but faired poorly in others, but received little MVP support, I can see how this is so.

In between a variable being linear or only a league-leading mark, there is the notion that passing a certain mark known as a "good season" indicator also might be important. Everyone knows hitting .300 sounds lots better than hitting .299. The difference between a season total of 102 and 98 RBI might be perceived as much bigger than 85 to 81 RBI, even though both are the same 4 RBI apart. So, additional dummy variables were set up for all seasons where players' stats passed specific milestones: .300 AVG, 100 R, 100 RBI, 30 HR. I believe these standards, even though they have been much more difficult to reach in some seasons than others, have over long periods of time been looked at as benchmarks.

What else besides raw stats might be factors? Possibly a player's defensive position might be important. So, a variable was added for their primary position played in that season. 1B, RF, LF and DH were lumped together. It is true that statements have been made by BBWAA members who didn't want to "vote for a designated hitter", but there were so few DHs in this sample that I elected not test for this. Also, the player's defensive ability could be viewed as important, which could best be captured by the gold glove awards (as imperfect as they are, they *do* represent perceived value, and it is perceived value we wish to measure). I decided the best way to capture this was to record if the player won a gold glove award in the season in question, or in the previous two seasons. Even though the gold glove awards are not officially given until after MVP voting is conducted, there is enough evidence of the seasonal play (both in the stats and in the press) that writers could easily perceive if a player was likely to win a gold glove. I also gave credit for an award won in the previous two years, as a way of capturing whether the player was judged to be a gold-glove quality fielder at the current time. These data are only available from 1958 to the present, so all players pre-1958 are credited with winning a small portion (2/9) of a gold glove.

Also, the team's success is very important. Initially, this was the most difficult factor to get a grip on. It is obvious that players on a winner get a boost. It is also obvious that a poor player on a winner gets no votes. Initial attempts at correlating pennant winning with MVP votes were not very successful. Then, I happened upon a solution: by combining 2 factors, often the truth was seen. By far the strongest example of this is shortstops. Given two players on the same team with identical hitting stats, one a shortstop and one a first baseman; if the team finishes poorly, they tend to receive the same amount of MVP support. On the other hand, if they win the division, the shortstop gets far more credit. Apparently there has existed in voters' minds a tendency to reward shortstops with extra credit if (and only if) they play on winners. So, I set up dummy variables which were "position played * {team in playoffs}", so that for all non-playoff teams, the same score is rendered as if the player were a catcher or a corner OFer. This turned out to be a HUGE thing for shortstops (and catchers and second basemen) on pennant winners; easily the single largest factor of all of the dummy variables. This caused me concern in that since I only profiled the top 3 finishers; maybe the model was only picking out the shortstops who DID do very well in the voting, and assigned a high positive coefficient, while maybe there were others who didn't receive nearly as big a boost, but because of their ballot absence the model did not take this into account. So, before using the model as predictive, I went back and checked to see how key defensive position players on other winning teams fared. As it turns out, there was no evidence that the model has missed other SS/2B/C who played on winners…except in the past 5 years. More on that later.

Others of these "combination variables" (winning + something else) were tested as well. Remarkably (to me), many others were found to be critical factors. Stealing bases is nice…but stealing bases for a winner is much more valuable in voters' minds (apparently, if they won, you sparked your team to success, but if they lost, your running was to no avail). The other variable that proved *very* useful when combining with team success was leading the league in RBI. In fact, this was seen as such a crucial factor, that I used dummy variables for players who finished 2nd or 3rd in RBI when playing for winners as well.

The definition of "winning" has changed over time. From 1938-1968, there was only one pennant winner per league. From 1969 to 1993, there were two division winners. From 1995 on, there have been 4 playoff teams in each league. I found no substantial evidence that winning a division of 6 or 7 teams was any different than winning a pennant of 8 or 10 teams. So, a "winner" in all of these years simply meant winning your division or league. The data set is limited in the last eight seasons, but I found a better fit when dividing the importance of "winning" by two than by keeping it the same. Thus, when I combined data across all of the years, I only counted playoff teams in the modern wild-card era as half of a winner. I did not attempt to codify the value of a team finishing in a first place tie and losing a playoff; there aren't too many instances of this.

I did not run variables for teams coming in a close 2nd, or finishing in a tie and losing in a playoff. There were not enough players on "tied" teams to make a definite determination, although it is obvious that there was consideration given to some players in these circumstances (Maury Wills, 1962) while others were ignored (Ted Williams, 1948). It was difficult to codify when a team was "in the race until the end" with only knowing the final standings, and finding a single variable that would suffice, so no bonus for finishing 2nd or less than X games behind was given.

Other factors considered: team win totals, playoff success in previous season (do voters favor those whose team was newly successful over having won previously?), seniority (is there a bias toward veterans?), previous MVP awards won (do voters tire of naming the same man, as Bill James and others have suggested?). Lastly, there seemed in my adult lifetime to be unusual publicity surrounding a player who was new to the team when they had (unexpected) success; he often got credit for the team's fortune. I tried multiple combinations of criteria, such as "players who were new to a team in that year, whose team made the playoffs, after not having made the playoffs the previous year." I was aiming to account for players like Terry Pendleton in 1991 out-pointing Bonds and denying him 4 consecutive MVP awards, and also like Ichiro in 2001. Alas, I was unable to codify the phenomenon that I was sure exists.

I considered ("I considered" typically means "some wise person suggested to me") putting in a variable for race (or skin color). However, I was unable to obtain data to make a definite determination of "dark or light" skin color.

I also considered creating a dummy variable for New York bias (which one of my friends swears has been in effect), or big-market bias. "Big market" seems difficult to define, so I did not pursue this angle. As for the NY factor, it will be easy to see at the end if players on some teams in general, if any, have been over- or under-rated.

Lastly, I thought about the problem of two or more players one on team potentially "splitting the ballot." Despite my best efforts, I was unable to create a workable variable that tested for significance using the data. In hindsight, if I had used more than the top three balloters, this probably would have worked. There were not many times when two or more players on the same team appeared in the top three in one year; possibly just because of this effect! Additionally, sometimes the presence of a pitcher on the ballot could also affect the votes given to a teammate, which also would not show up here.

How did I decide on which variables to keep?

First, the variable ought to be something reasonable in a baseball sense. Just as it is absurd to conclude that some hitters do better in night games on Tuesdays simply because some statistical test is passed, so we ought not to conclude that some criterion is important to MVP voters if there is no common sense evidence for it. Said another way, when there are pages of press written about how many runs a batter drove in, it is reasonable to conclude that this is a relevant piece of data; but if the data showed that players who hit at least 37 doubles in a year fared better in the voting, we should question the use of this information, since I know of no newspaper articles extant that extol the value of a player's 37th two-bagger in a season. Second, the variable should pass a reasonable statistical test for significance. Traditionally, 5% or less chance of a relationship being due to mere randomness is used, so I chose to use that here, although sometimes I attempted to balance common sense with the hard math, particularly when using smaller amounts of data (fewer years). What this means is, if a variable proved to be relevant over a large sample of years, I attempted to keep it when moving to a subset, even if the statistical significance was becoming borderline. Lastly, the relationship also had to be relevant in an absolute sense. It does not matter if a variable is statistically significant if it only adds a miniscule amount of accuracy; therefore a variable which purports to confidently add fewer than 5 points to a player's MVP vote total will not be used.

Overall Results

A quick outline - the most important factors are, in rough order:

Very Important	Great Triple Crown stats, leading the league in RBI for a playoff team, and playing SS for a playoff team;
Important	Being a new guy on a surprise winner, playing C or 2B for a playoff team, leading in SLG and/or OBA, hitting over .300, winning a gold glove, and finishing 2nd or 3rd in RBI for a playoff team;
Somewhat Important	Stealing lots of bases for a playoff team, and leading the league with lots of steals;
Helpful	Playing for a team that won many games (and didn't win the previous year), scoring and/or driving in at least 100 runs, and being a veteran.

The Formula

For the entire data set of years, 1938-2002, the model would say that the following procedure leads to the best fit of MVP voting:

Start

Start with 114, which is the intercept in the regression formula.

Normalized Triple Crown stats

Subtract the league 3rd-place finisher's total for each of these 3 categories (producing a negative ## for anyone who did not exceed the 3rd-place finished in a given category.

Then add: normalized AVG * .34 (after removing the decimal point in average)
Add normalized HR * .62
Add normalized RBI * .57

General Stuff

Add .35 * number of team wins
Subtract 9 if team won last year
Add 2.2 * number of years played in MLB
Add 20 for being a gold glover

Thresholds

Add 17 for hitting .300+
Add 8 for 100+ Runs scored
Add 8 for 100+ RBI

League leading bonuses

Led league in Steals: add .50 * SB
Led league in SLG: add 20
Led league in OBA: add 24

If the player was on a playoff team:

Add 50 for leading league in RBI
Add 25 for finishing 2nd in RBI, 13 for 3rd
Add .82 * SB
Add 25 if player is a 2B or C
Add 50 if player is a SS
Add 31 if he this was his first year with the team, and the team did NOT make the playoffs the previous season

An Example: DiMaggio/Williams

Let's illustrate the system with the great controversy of 1941, DiMaggio vs. Williams.

Joe Dimaggio

Start with 114.

Joe had Triple-crown stats of 357-30-125. The third place numbers were 337-31-122. Normalized, then, DiMaggio is 20/-1/3.

*Add 20 * .34 for batting averate.*
*Add –1 * .62 for home runs.*
*Add 3 * .57 for RBI.*

His total is now 121.8.

The Yankees won 106 games (prorated to a 162-game schedule).

*Add 106 * .35.*

Joe was in his 6th year in the major leagues:

*Add 6 * 2.2.*

DiMaggio's total is now 172.1.

Joe hit over .300, scored 100 runs, and drove in over 100 runs.

Add 17 + 8 + 8.

He gets credit for 2/9 of a gold glove, or 4.4 points.

Add 4.4.

The Yankees won the pennant, so Joe gets credit for leading the league in RBI (50 points), and for his 4 SB (.82 points each).

*Add 50 + (4 * .82).*

His final total: 262.9 points.

Ted Williams

Start with 114.

Williams' Triple-crown stats were 406-37-120. The third place numbers were 337-31-122. Normalized, then, Ted is 69/6/-2.

*Add 69 * .34 for batting averate.*
*Add 6 * .62 for home runs.*
*Subtract 2 * .57 for RBI.*

His total is now 140.0.

The Red Sox won 89 games (prorated to a 162-game schedule).

*Add 89 * .35.*

Williams was in his 3rd year in the major leagues:

*Add 3 * 2.2.*

Ted now sits at 177.8 points.

He hit over .300, scored 100 runs, and drove in over 100 runs.

Add 17 + 8 + 8.

He gets credit for 2/9 of a gold glove, or 4.4 points.

Add 4.4.

Finally, Williams led the league in SLG and OBA.

Add 33 + 44.

His final total: 259.2 points.

The model believed Joltin' Joe would barely nudge out The Splendid Splinter, by a score of 263 to 259. His actual margin of victory was slightly larger, 291 to 254.

More Results

This model predicts the correct winner in 82 of the 130 races, for a 63% score. The r^2 value for correlating MVP points received to those predicted is .420. That may not sound very high, but considering that it measures votes across seven decades in many different circumstances, in my opinion that isn't bad. The r^2 score can be higher if a smaller sample of years is used. The RMS error between MVP points predicted and received was 49. Stated another way, the model is accurate to within 49 points for the top 3 vote-getters, two-thirds of the time.

Comments to the above data:

1. HR and RBI are more closely correlated than any other pieces of data. Therefore, it is difficult to precisely measure their effects separately.

2. Throughout most of the testing, the bonuses for scoring and driving in at least 100 runs seemed to be about equal, so I tied them together.

3. The use of leading the league in OBA is a bit problematic; it borders on violating one of the principles I set up in the beginning, in that for many years OBA was not a well-published statistic. However, it was clearly a strong predictor, and statistically significant at p=.011 (almost 99% significance level). Since it is possible that sportswriters were in a general sense aware of some hitters who reached base via the walk quite often, I chose to keep the variable. Besides, I wanted to test if it became more significant in differing eras.

4. I first attempted to find 3 separate coefficients for finishing 1st, 2nd, and 3rd in RBI for a winner, but there is much noise in the data and a limited sample. Once I had a general feel for the numbers, I settled on this common-sense approach that closely approximated the data: finishing 1st was twice as important and finishing 2nd, which was twice the value of being 3rd.

In the aforementioned study by Rob Wood, he found that leading the league in RBI (as compared to other league-leading categories) was the single most important predictor of MVP award voting. My finding here matches well with Rob's study.

5. The position bonuses were calculated by first determining the relationship between the different positions: playing SS on a winner was typically about twice as important as playing 2B, which was about equal to catcher, and I found no advantage for other positions. I chose to use integer relationships and simply assign SS a value of 2 and the others as 1, and then put them together as position code, keeping the proportion the same as letting the regression determine the joint values. In a similar fashion to how I approached the RBI leadership issue, there probably there would be slightly different numbers if I went back and coded them all separately, but this creates many more variables. I will re-address this when looking at each time period.

6. Stolen bases only appeared to be beneficial in voters' opinions if they were swiped for a team that won, or if the player led the league. In other words, whether a player stole 20 or 0 bases for a loser is irrelevant, but stealing them for a winner apparently meant that you provided the "spark." I first attempted to use a simple dummy variable for the stolen base leader, but leading the league with 90 steals was worth more in the voters' perception than leading with 35.

Data found NOT to be significant

1. Leading the league in batting average, hits, doubles, triples, total bases, walks, or runs scored. In some cases, there was significance found in stats which shadowed these (slugging instead of total bases, for example). Others of them simply may not be prevalent in many voters' eyes.

2. Playing 3B or CF (which tested to be the same as playing 1B/RF/LF). Bill James has claimed that centerfielders received a voting boost around the 1950s decade, so I will address this when looking at each time period.

3. Hitting 30 home runs.

4. Leading in RBI for a non-winner. Now, this really surprised me. When I originally set up the model, this looked to be a big factor. But, after I combined the RBI leader with winning the division/pennant, leading in RBI by itself lost its significance. There was some positive correlation for leading-in-RBI-for-a-non-winner, but the noise in the data was too much to call it a definite factor. It may be that amassing a large quantity of RBI, plus the peripheral stats that typically accompany this, gives enough weight without adding in the dummy factor for being first in the league.

5. Winning previous MVP awards. There was a small negative effect noted, but it was dwarfed by the noise in the data. However, a cursory look at the data shows that this effect probably was true for part of the voting history, so I will discuss it when breaking down the timelines, and when looking at Mays and Mantle.

The player for whom the most points were predicted (the most worthy MVP in any one year) was Joe Morgan in 1976. The model predicted him to collect 366 points (which is actually higher than the maximum of 336, given a 12 team leagues; this is one problem with a linear model. This phenomenon occurred one other time, with Frank Robinson in 1966). That season, Morgan hit .320, with 27 HR, 121 RBI (2nd in the league), and 50 stolen bases for the pennant-winning Big Red machine. He also won a gold glove at 2B, and led the league in both SLG and OBA. However, Joe was not the unanimous MVP in actual voting that year (his actual point total was 311), as his teammate George Foster, who led the league in RBI, and Mike Schmidt, who had a fine year for the other division winner, garnered some support as well.

The largest margins, or most lopsided MVP votes, were predicted to be the two triple crown winners who played for pennant winners: Carl Yastrzemski in 1967, and Frank Robinson in 1966. I suppose this is not exactly an earth-shattering revelation. In this case, the model didn't even need to know about Boston's rise from the basement the year before or Yaz's September heroics, or the fact that F. Robby came over in a big trade and was perceived to be the main reason the team won.

The largest positive error in any season, the player who gained more MVP recognition than the model predicted occurred recently. The Braves' Chipper Jones in 1999 garnered 324 (adjusted) points, compared to a predicted 184. The model saw him coming in 3rd, a long way behind the RBI leader from the NL West champ Diamondbacks, Matt Williams. The model does not know of Chipper's single-handed weekend destruction of the rival Mets in September of that season.

The largest negative error, the player who in one year got no respect from the voters, was Del Ennis in 1950. Del hit .311 with 31 HR, and led the league in RBI for the winning Phillies. The model saw him finishing with 262 points, ahead of Musial and Stanky. The model did NOT take into account that Jim Konstanty, the Phil's ace reliever, would be given the lion's share of the credit for his team's performance, and so Ennis finished 3rd among the eligible hitters with 104 points. There are a few other instances like these when the presence of pitchers on the ballot created havoc with the voting.

The 4 other largest misses:

		pred	actual		
Player	year	pts	pts	diff	comments / explanation
Slaughter	1946	274	144	-130	Led league in RBI, but mate Musial's .365 avg won voters
F Robinson	1969	224	102	-122	Teammate Boog Powell's #s were slightly better
E Howard	1963	177	297	+120	Yankee catcher given much credit for team success
Rose	1973	159	274	+115	Writers loved Rose's .338 AVG and hustle

Some of these errors, and over- and under-predictions, may look different when the model is tweaked to reflect values in different time periods.

To put some perspective on these numbers, the average point total for MVP winners in the years 1969 to 1992 was 286, pro-rated for a 12-team league. That would be 334 in a 14-team league, and 381 for today's 16-team NL. (This is discarding winners who were pitchers, as I have done in this study.)

The overall model had a good streak in the period from 1965 to 1983, when it correctly predicted the winner 32 times out of 38 (84%). This might mean that the voters of this period closely approximated the overall average; or that they stuck more with year-end statistics in this time; or that it's just one of those things.

What about Park Effects?

I did not test explicitly for park effects. Firstly, because I did not see any obvious indications that writers discounted the stats of players in good-hitting stadiums. Secondly, because it is difficult to codify specific park-effect numbers. The most obvious example would be all of the hitters who had the good fortune of playing half of their games in Fenway, long known (correctly) as a fine place to hit. However, with the singular exception of Ted Williams (more about him in Part II), there is little evidence that the Red Sox players received less consideration from the voters because of the environment in which they plied their trade. In other words, it is safe to conclude that for most of voting history, hitting .300 in Wrigley is just as good in voters' eyes as hitting .300 in Chavez Ravine.

However, when the Rockies brought us mile-high baseball in 1993, and all of the stats in Colorado home games were so obviously distorted, it seems that writers have indeed noticed this, and have applied a discount to the numbers but up by Rockies hitters. For example, Andres Galarraga hit .370 his first year in Colorado. He finished 10th in the MVP voting. In 1996, he clubbed 47 home runs and drove in 150. He finished tied for 5th. I don't have enough data to give a confident estimate of just how much the typical Rockie hitter is "penalized"; but it was enough to give Barry Larkin his MVP award in 1995 over Dante Bichette, yet not enough to deny Larry Walker his award in 1997 over Mike Piazza.

2003

So, what did the model say for this season (dang, I *knew* you were going to ask that question)? Actually, after all of this work, I am convinced that using a model is not the most accurate way of predicting future MVP award voting. Why? Simply because so much of what actually occurs is beyond the year-end statistics. The media begins attaching themselves to various candidates in mid-season, and often anoints their favorites as the teams come through the home stretch. A last-minute charge or swoon by a team and a key player may change voters' minds in some seasons. And when it is obvious who the MVP will be, the statistics correlate very well with the amount of press a for-certain player is receiving, so who really needs a model to tell us that Barry Bonds' 73 home runs in 2001 was going to net him another trophy? In addition to all of this, it should be obvious that a model which uses data from voting patterns in 1938 may not be accurate in 2003; I will look at this in part II of the article.

Well, gee whiz, Tom, what's the point of creating a model if it has less predictive power than that of someone who merely finds what the media is thinking by reading a few September daily newspapers? Well, again I will say: what I see as the main point of this exercise is NOT to predict what might happen in the future, but instead to discern what occurred in the past, as we in the 21st century look back and question why some players were more well-thought-of than others, according to the statistics that we view now, without having "been there" to recall what happened then and there. In other words, I am much more fascinated by what the model says about Wade Boggs' lack of any MVP support throughout his career (answer: voters are not impressed with batting leaders who also draw lots of walks, but who don't get RBI or steal bases; so his lifetime lack of MVP support is more a function of the voters' preferences than any particular devaluing of Wade on their part) than if the model will be "right" this year or not.

But hey, let's play along anyway. Here's the National League prediction:

		Tm	Win	Nth	Gold	Hit	100	100	SB		Total
Player	AVG-HR-RBI	wins	yr-1	yr	Glv	.300	runs	RBI	ldr	Winner bonuses	Predicted
Sheffield	329-39-131	101	Yes	16	No	Yes	Yes	Yes	No	RBI 2nd, 19 SB	322
Bonds	339-45- 89	100	Yes	18	No	Yes	Yes	No	No	SLG, OBA, 7 SB	307
Lopez	328-43-109	101	Yes	12	No	Yes	No	Yes	No	Catcher	266
Helton	358-33-117	74	No	7	No	Yes	Yes	Yes	No		239
Pujols	358-43-124	81	Yes	3	No	Yes	Yes	Yes	No		234
Pierre	305- 1- 41	88	No	4	No	Yes	Yes	No	Yes	64 SB	231
Thome	268-47-131	86	No	13	No	No	Yes	Yes	No		221

And the NL results:

		Actual Pts	Predicted Pts	Difference
1	Bonds	426	307	+119
2	Pujols	303	234	+ 69
3	Sheffield	247	322	- 75
4	Thome	203	221	- 18
5	Lopez	159	266	-107

Todd Helton finished a distant 7th (6th among hitters).

Note on vote totals: since I only attempted to match the top 3 on each ballot in the model, there should be little expectation for predicted vs. actual vote totals to match as you go further down in the ballot. The actual vote totals in most years fall off much more quickly than this linear model suggests. If anyone wishes to build a model for a more encompassing top 6 or top 10 look each year, the approach would need to change.

Note on gold gloves: The 2003 Gold Glove awards are not known prior to casting of MVP ballots. So, the system I used in hindsight (current year or any of 2 previous years) is slightly different than when attempting to use the model for predictions.

Contrary to what actually occurred, the model saw extremely fractured voting in the NL. The model saw that Sheffield, with his (tie for) second place in RBI for the winning Braves, would squeak by Bonds and the others. The model also figured that the Cardinals' and Phillies' failures to capture playoff spots hurt the chances of Thome and Pujols. The model saw that with only 89 RBI, Bonds would not walk away with the award; however, the media seemed to have taken to heart the fact that it's difficult to drive in runs when being intentionally walked whenever runners are in scoring position, and have granted Barry essentially an exemption for not driving in so many runs.

For most of the summer, Pujols was the media MVP darling; he came reasonably close to a Triple Crown, and his team came close to making the playoffs, two factors for which the model gave him no credit. In contrast, there was no bandwagon in the press regarding Mr. Sheffield's candidacy; he posted some impressive numbers about as quietly as anyone ever has. Javy Lopez received much credit for the Braves' fine year, which probably hurt Sheffield. Lastly, since voters in the past decade have discounted Rockies hitters' stats somewhat, it was not a surprise to see Helton finishing much lower than 4th.

Here's the American League:

Player	AVG-HR-RBI	Tm Wins	Win Yr-1	Nth Yr	Gold Glv	Hit .300	100 Runs	100 RBI	SB Ldr	Winner Bonuses	Total Predicted
Garciaparra	301-28-105	95	No	8	No	Yes	Yes	Yes	No	SS, 19 SB	281
Ramirez	325-37-104	95	No	11	No	Yes	Yes	Yes	No	OBA, 3 SB	259
Rodriguez	297-47-118	67	No	10	Yes	No	Yes	Yes	No	SLG	246
Delgado	301-41-141	86	No	10	No	Yes	Yes	Yes	No		242
Posada	281-30-101	101	Yes	8	No	No	No	Yes	No	C, 2 SB	192

Posada was predicted by the model to finish 9th, behind Tejada, Boone, Soriano, and Jeter

Here are the actual results:

		Actual Pts	Predicted Pts	Difference
1	Rodriguez	242	246	- 4
2	Delgado	210	242	- 32
3	Posada	194	192	+ 2
4	Stewart	140		
5	Ortiz	130		
6	Ramirez	103	259	-156
7	Garciaparra	99	281	-182

The good news: the model was extraordinarily accurate in predicting how many points the actual 1st through 3rd place finishers received. The bad news: the predicted winner finished 7th.

It was a real mess to pick a winner in the American League this year. The guys with the gaudiest stats played for losers. Most of the winning teams were led by more than one great player, which split credit for their accomplishments. So, the model foresaw that the voters will have mixed opinions, with Nomar coming out on top, being a fine-hitting shortstop for a playoff team.

But wait a minute. Although I haven't yet done a complete analysis of changes to the model over time, it should be very obvious to anyone who looks at recent history that there are currently an unusual number of big bats that play shortstop (A-Rod, Garciaparra, Jeter, Tejada). And, at the same time, the large amount of extra credit that writers have historically given to these key defenders has magically shrunk. Coincidence? I don't think so. The model erroneously saw Alex Rodriguez garnering the 2000 AL award. He in fact finished a distant 3rd behind Jason Giambi and Frank Thomas that year. While there may have been other factors, I believe that had A-Rod been the *only* shortstop putting up amazing numbers in the past five years, that he would have at least one more trophy on his shelf. Other instances of shortstops who have fared far worse in voting in recent seasons than the model showed were Edgar Renteria in 2002 (predicted finish 3rd, actual finish way back), and both Jeter and Garciaparra in 1999 (predicted to finish a close 4th and 5th, actual finish further behind). This effect may be the primary reason that Nomar did *not* get the model-expected benefit of playing shortstop for a winner. Just stop for a second, and think about Nomar's year: if a shortstop for a playoff team hits .300, clubs 28 home runs, scores 100 runs, drives in 100 runs, steals 19 bases, and plays well in the field, doesn't that often scream "MVP"? Not so in the past few years. Tejada and Jeter also finished well below the model's prediction, as did the two slugging second basemen, Boone and Soriano. Instead, Posada received most of the credit for being the key player for the Yankees; he was the all-star catcher.

Complicating matters even further was that with no clear-cut favorite based on the numbers, the media latched on to pet candidates who in their eyes rose above the numbers to become MVP candidates. David Ortiz of Boston had many key hits down the stretch. The extra votes he got came at the expense of his teammates, especially Manny, who was by traditional stats the "big bat on the winning team." Shannon Stewart has received a ton of credit for the Twins strong 2nd half; Jayson Stark of ESPN was hawking his exploits in the press, and obviously others followed his logic. This is another reason why a model using only year-end stats will never be a highly accurate predictor of voting. Thirty years from now you can attempt to explain to your grandchildren how or why Mr. Stewart finished 4th in the 2003 voting, because it sure will look odd to anyone who merely looks up the numbers.

In one sense, you could say that both Ramirez and Garciaparra missed their vote total predictions by historic amounts; over 150 points each! So, do we have a new record for "model misses by biggest amount"? Well, not really, and again here is a lesson learned. When attempting a best fit by regression, by only using the top three finishers, it is possible for the model to miss a player who was predicted to win (or finish 2nd or 3rd), but actually finished 4th through 10th, since that data point is not recorded. This adds credence to the argument that a more accurate model might be achieved by utilizing more than three of the actual top finishers. Back when I began this research, I performed a cursory look at many of the 4th through 6th place finishers to see if the model was missing some key information by only looking at the top three. I did not find much, which led me to only use three, but obviously this season's results, if we make use of Nomar's poor finish, scream loudly that the underlying assumptions of the model did not hold.

Summary

Often, the league MVPs are defined by great Triple Crown stats, with large edges going to players on playoff teams who drive in lots of runs or play a key defensive position.

A linear regression model using about 20 year-end stats predicts (in hindsight!) the correct MVP winner 63% of the time, with an RMS error in (actual minus predicted) point totals of 49.

Many things happen during a baseball season that will not be accounted for by only using year-end statistics.

By using one model for 65 seasons, the model may miss trends in smaller groups of years. The de-emphasis on shortstops in recent years may be one of these trends.

That's all for now. In part II, I will analyze how things have changed over time. Is there more emphasis on some statistics now than in 1940 (uh, that would be "what is Runs Batted In", Alex)? I'll also look at some of the commonly proclaimed all-time greats of the past 65 years, and see how their MVP vote totals square with their numbers. And maybe there will be some other tidbits that are uncovered along the way.

Tom Hanrahan, Han60Man@aol.com ♦

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.