By the Numbers

Volume 10, Number 3 The Newsletter of the SABR Statistical Analysis Committee August, 2000

Committee News

Neal Traven, Committee Co-Chair

Greetings, fellow SAC members! I have quite a bit to report to you this time around. I'll offer a short wrap-up of the 30th annual SABR National Convention, recently-received information about the future SABR publication I touched on briefly in the last issue of BTN, a request for assistance from another SABR committee, and more.

Convention report

This year's meeting in West Palm Beach was very poorly attended. The final count barely crept over 300, far below the norm of 450 or so. Maybe it was anticipation/fear of hot and humid weather (yes, it was hot, but not beastly) or the significant distance to the nearest major league ballpark (Pro Player Stadium wasn't great, but wasn't awful … the same can be said for the Marlins). Perhaps it was the indifferent preparation by the local organizers or uncertainty over the recent resignation of the Executive Director. In any case, those of us who braved the Florida sun still had a pretty good time of it.

In this issue

In the convention program, I count 42 oral presentations. Among those presentations, nine (not necessarily statistical) were given by SAC members. In alphabetical order: **Tony Blengino**—*Pedro Martinez, 1999: The Greatest Pitching Season Ever?*, **John Jarvis**—*Hitting in IBB Situations*, **Mark Kanter**—*Eight Days in September and October*, **Steve Krevisky**--*1950: The Year The Tigers Roared*, **Mark Pankin**--*What Is Speed Worth?*, **Doug Pappas**—*111 Years of Major League Ejections*, **David Raglin**—*Go To The Co'*, **David Smith**—*From Exile to Specialist: The Evolution of the Relief Pitcher*, and **Dick Thompson**—*Joe Pinder: Baseball's Greatest Hero*.

In addition, I'd consider six presentations by non-members of the SAC to be statistical in nature, or at least in intent. They include: **Ronald Cox & Daniel Skidmore-Hess**—*Baseball Competitiveness in the Free Agent Era*, **Jonathan Dunkle**—*The Closer: The Impotance of Grooming*, **Stephen Grimble**—*Setting the Record Straight: Baseball's Greatest Batters*, **Ron Seltser**—*Baker Bowl in the 1930s*, **Stan Slater**—*Hits: A Misleading Statistic*, **Ted Turocy**—*A Strategic Analysis of Stealing Bases: Game Theory at the Ballpark.*

Approximately half a dozen posters were displayed during the convention. Unfortunately, neither titles nor presenters were listed in the convention program. I picked up the handout from **Clem Comly**'s poster of *ARM: Average Run-equivalent Method (see page ? of this issue)*, and I recall that **Tony Blengino** displayed this year's version of his standard report on minor leaguers. To me, however, the most interesting poster was by Western Carolina University professor Angela Grube, displaying her doctoral dissertation *Team Cohesion and Its Relationship to Performance Success in the Cape Cod Baseball League*. She hypothesized that teams establishing more cohesiveness among the newly-met teammates would perform better. In other words, she was studying "chemistry." The results, however, showed no association whatsoever between cohesiveness and winning. She plans to continue this research by studying baseball players in the entire Southern Conference over several years.

Along with **John Matthew IV, Tom Ruane**, and about half a dozen others, I was a judge for the Baseball Weekly Award, given for the best presentation during the meeting. Unfortunately, I was assigned to grade another talk when **Doug Pappas** gave his award-winning talk. **Dave Smith** was a close second in the voting, as he has been on at least two other occasions; that's why I've dubbed him the Dave Stewart of SABR presenters.

Newly-hired Executive Director George Case made the rounds at the convention, hitting the Committee Chairs meeting, the Regional Leaders meeting, and most of the individual committee meetings. Others have taken issue with the process by which Case ended up with the job, about which I know nothing aside from what I read in Nic Antoine's email newsletter. My own discomfiture with Case arises from his very first comments to the committee chairs. While generally waving the flag about the good work of SABR and its research, he went out of his way to mention dislike, even disdain, for sabermetrics and statistical analysis. It was, in my opinion, an extremely undiplomatic thing to say, in that or any other public setting. It would have been perfectly OK to express unfamiliarity or even disinterest in our work, but Mr. Case went well beyond that. One wonders what sort of response would arise in the general populace of SABR if someone in his position were to say something similar about, say, the minor leagues or umpires.

Future "SABR Special" publication

In the last BTN, I briefly mentioned that SABR President Jim Riley had asked Publications Director Mark Alvarez to explore whether research committees and regional chapters were interested in contributing articles for a future publication. We discussed the concept – what kind of article to submit, how to choose one, etc. -- at the SAC meeting in West Palm Beach, even though at that time it was not known whether the publication would even get off the ground.

That last question, it seems, has been answered. On August 21, Alvarez sent an email to the committee chairs saying that the project has gotten the go-ahead from the SABR Board. Committee submissions, camera-ready including any and all graphics, are expected to be in the Publications Director's hands by May 1, 2001. A number of members at the meeting offered to participate on an ad hoc subcommittee to work on coming up with a manageable number of candidates for this publication. It seems logical that the best place to look for the candidates is in the pages of *By The Numbers*, though other sources would also be of interest. Another avenue to examine might be to put together a "point-counterpoint" article, combining the work of several contributors with differing views on a specific topic.

It's important for this article to demonstrate to the SABR community-at-large that statistical analysis can be insightful, challenging, and valuable. Beyond that, though, the article needs to be accessible, understandable, interesting, catchy, and entertaining. Therefore, the author of the piece might be asked to update the analysis with more recent data, or to reword or tighten or clarify the text. To reiterate, this article could be a valuable selling point for the Statistical Analysis Committee within SABR, so it needs to be as good as it can get.

To be sure, it's a tall order. If we find the right sort of material, this is a chance to attract the rest of the organization to the excitement that we find in our analyses, as well as a chance to build our constituency in SABR. Who knows, maybe George Case will want to read it!

What I ask of all of you is to look back through your back issues of BTN for articles or a sets of articles that strike you as possible condidates for this type of general SABR publication. If something jumps out at you, let me and/or Phil know about it. Please provide us with titles, authors, publication dates. Also, let us know *why* you suggested using this material as our contribution to the publication.

Sabermetric definitions

While we're in the uncharted waters of what might be construed as a "committee project", I'd like to bring up another such item. In West Palm Beach, Skip McAfee asked for assistance with identifying and defining statistical terminology to be included in the next edition of Paul Dickson's *Baseball Dictionary* (Skip is the book's editor). Soon after the convention, he emailed me a list of several dozen terms gleaned from SABR-L. Those familiar with the Dickson dictionary know that it contains not merely definitions, but also such characteristics as etymology and first usage.

Dickson and McAfee want to separate the terminological wheat from the chaff, ascertain that the definitions of terms are correct, and describe the origins of the terms they include in their tome. I plan to participate in their endeavors as best I can, and I'm sure there are others out there fascinated by both baseball statistical analysis and the nature or character of language. If you'd like to get involved in strengthening the sabermetrics (defined in the 1999 edition as "the study and mathematical analysis of baseball statistics and records") content of the Dickson dictionary, please drop me a line.

Neal Traven, 500 Market St. #11L, Portsmouth, NH, 03801; 603-430-8411; beisbol@mediaone.net ♦

Pitchers' Equal Hit Ratios: Evidence, Please

Tom Hanrahan

Last issue, Clifford Blau referenced a study suggesting that the likelihood of a ball-in-play falling in for a hit is independent of the pitcher. In this comment, the author finds the result counterintuitive, and suggests another study to provide more evidence.

In the May 2000 *BTN*, Clifford Blau prepared a review of a website article entitled "Defense Independent Pitching Stats (DIPS)". Based on thearticle's findings and Cliff's review, I would propose a research project for someone to undertake:

Find a large set of pitchers with long careers, and see if any of them have significantly different ratios of hits (excluding home runs) allowed to balls in play. The project may want to focus on pitchers of the same park to minimize those possible effects, but of course this reduces the potential sample of pitchers. One method which could be used to mitigate park effects woud be to find a team average (use the team stats over a number of seasons) and either back out any park-induced biases, or only use pitchers who worked in stadiums that apparently allow a "neutral" ratio of hits to balls-in-play.

The article's conclusion (all pitchers have essentially the same ratios and therefore this is not a "skill") seems counterintuitive to me. After all, locaton of pitches is a choice, and when a pithcer chooses to aim for the heart of the plate or 2 inches off the corner, it seems he is choosing bewteen the potential evils of a walk versus an increased chance of a hit. If hits allowed to balls-in-play is a constant, why not always throw sinkers down the middle? However, if this is what the data show, then I won't argue with the facts. I'm just not ready to accept something yet that appears on the surface to be different than the reality I perceive. But go ahead, show me.

Tom Hanrahan, 21700 Galatea St., Lexington Park, MD, 20653, HanrahanTJ@navair.navy.mil. ♦

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).

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

"Baseball Dynasties" Crisp and Readable

Gabe Costa

Sabermetrically based, this well-written book is an informative and enjoyable evaluation of baseball's best teams ever.

This book is very readable, the style being reminiscent of sabermetrician Bill James. And much like James' writings, this book can be opened at random, with the reader immediately getting into the authors' line of thought. While it probably would not be considered a "source book", there is a lot of baseball information in this publication.

The format of the book is somewhat similar to *The Bill James Historical Baseball Abstract*. Fifteen 20th century teams are investigated (ranging from the 1906 Chicago Cubs through the 1998 New York Yankees). Interspersed are chapters dealing with such topics as the greatest teames of the nineteenth century, the worst teams in major-league history and great teams of the Negro Leagues.

For much of the book there are "sidebars" by both authors, often on each page. Their writing is crisp, sometimes peppered with "salty" language. At times, the reader will find himself/herself laughing out loud.

The book is fairly sabermetrical in nature. A glossary is provided at the beginning defining terms such as runs created, real offensive value and the Pythagorean method. And to compare teams across the years, the authors define a measure they call the standard deviation score (SD). While there are enough statistics to support the authors' positions, one is not overwhelmed with numbers, symbols, etc.

Baseball Dynasties: The Greatest Teams of All Time

By Rob Neyer and Eddie Epstein

Norton, paperback, 384 pages, $17.95
ISBN 0-393-32008-1

Regarding the above, the SD is the authors' key statistic. They compute SD by simply adding an offensive component to a defensive component. For example, in 1927 the Yankees scored 975 runs, while the league average was 762 with a standard deviation of 115. Hence, the offensive part of their SD becomes (975-762)/115 = 1.85. Similarly, Yankee pitchers allowed 599 runs, which was 163 fewer than the league average. Because the league's standard deviation in runs allowed was 88.5, the Yankees' defensive component of SD is 1.84. Hence, their total SD = 1.85 + 1.84 = 3.69.

Regarding this, the authors assert that an SD of 3.00 or higher is very good; in fact, through 1998, the figure has been surpassed only thirty-seven times.

Ten SD charts are given at the end of the book; these cover both the best and worst totals ever, ranging from for one year through five years.

In addition to the sabermetrical measures, the authors investigate Hall of Famers from each team (who should be in the HOF; who shouldn't), the home field, the worst regular, various books about the team, how the teams came to be, etc. These are extremely helpful in a qualitative sense. For example, I vividly remember the 1961 Yankees (I was 13 years old); I did not realize how poorly Bobby Richardson and Tony Kubek fared as table-setters for Roger Maris and Mickey Mantle.

At the end, the authors independently rank their fifteen teams (and actually agree on the best team ever!). This dialogue approach moves along briskly and is quite refreshing. It also brings in the "dynastic" aspect of not just limiting a great team to one year.

Again, I enjoyed the book. It was very readable. I recommend it.

(Father) Gabriel B. Costa (Seton Hall University and USMA), Seton Hall University, South Orange, NJ, 07079, costagab@shu.edu. ♦

Academic Research: The Equity Effect

Charlie Pavitt

Management Science uses a concept called "equity theory," which hypothesizes that employees who feel they are underpaid will expend less work effort. Here, the author summarizes an academic study that evaluates this effect in a baseball context.

This is the one of what I foresee as occasional 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 www.udel.edu/johnc/faculty/pavitt.html, use it for your research, and let me know what I'm missing.

Steve Werner and Neal P. Mero, Fair or Foul?: The Effects of External, Internal, and Employee Equity on Changes in Performance of Major League Baseball Players, Human Relations,Volume 52, 1999, pages 1291-1311.

Back in 1962, a business administration researcher named J. Stacy Adams proposed "equity theory." This proposal is relevant to the circumstance in which a person judges whether their job compensation is fair. People who consider themselves undercompensated are hypothesized to feel angry and to respond by expending less effort on the job. People who consider themselves overcompensated are hypothesized to feel guilty and to respond by expending more effort on the job. Equity theory was consistent with mainstream social scientific theorizing at the time of its proposal, and soon became quite influential not only in the field of business administration but also in social psychology (for example, it is directly relevant to people's reactions to feeling that they get more or less than they deserve in a marriage or friendship). Research has tended to support the hypothesis concerning undercompensation, but in the case of overcompensation, it seems that people tend to feel that they are deserving of it and do not increase their effort further.

Equity theory was used as the backdrop for two early studies of undercompensation and baseball player performance. Lord and Hohenfeld (Journal of Applied Psychology, 1979, vol. 64, pages 19-26) found that 13 batters "playing out their option" in 1976 (the year after the reserve clause had been overruled) suffered declines in their batting averages, home runs, and RBI during that year as contrasted with 1975 and 1977, whereas those who signed for 1976 did not. However, Duchon and Jago (Journal of Applied Psychology, 1981, vol. 66, pages 728-732), with a larger sample of 30 over the years 1976 to 1978, were unsuccessful at replicating this finding. The presumption in these studies is that those playing out their option were dissatisfied with their compensation, whereas those signing were not. Werner and Mero have taken on the same issue, but with greater sophistication. First, they used actual salary data for comparisons, allowing for a much larger sample size (205). Second, they considered three types of over- and undercompensation: "external," in which salary for batters from one team as a whole was compared to batters from other teams; "internal," in which the ratio of 1991 salaries for batters from one team when contrasted to pitchers for that same team was compared to the same ratio for other teams; and "employee," in which salary for batters for one team was compared to salary for batters for the same team. The presumption in this case is that those who were "objectively" underpaid relative to other players would "subjectively" feel undercompensated. As a nice test of the presumption, the authors found that those filing for arbitration for 1992 were more likely to suffer external and employee underpayment and less likely to enjoy external and employee overpayment than those who did not file.

Using regression techniques, the authors calculated what each batter "should" have been paid given their 1990 performance and their longevity, and defined each as over- or underpaid in 1991 compared to that calculation. The results showed that the change in runs created per at bat between 1990 and 1991 was positively, although weakly, associated with external and employee overpayment and negatively, and again weakly, associated with external and employee underpayment as predicted. In other words, an owner will get somewhat better performance from batters who are relatively well paid in comparison with other teams and their own team. Although an owner can overpay their team compared to other teams, there is not much they can do about employee underpayment, as by definition half of a team's players have to be paid below the team mean. Note that internal over- and underpayment never had any impact, and I for one would have been surprised if batters cared about how their pay scale relates to that of their team's pitchers. This is one of the rare equity theory studies in which overcompensation had the predicted results; the authors' explanation is that the public visibility of the overpaid player produces the guilt and resulting motivation to play well not found in the "average."

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

Who Was The Most Awesome Peak Offensive Player?

Tom Hanrahan

There are many possible ways to attempt an answer to the question, "Who was the best of all time?" In this article, the author concentrates on determining which players were the best, relative to their league, at the peak of their career.

The title question has been asked and answered many times before in various forms. This particular study is designed to answer the specific question – "What player had the highest established level of dominance in creating runs for his team, compared to others in his league, at the peak of his career?"

How do we measure this? There are two questions that must be addressed in order to give an answer. First, how to measure offensive dominance, and second, what is meant by "peak."

Offensive Dominance

I have used Bill James's Runs Created (RC, the technical version explained in the original *Historical Baseball Abstract*) as the measure of offensive production. It has been shown to very closely approximate team runs scored, and it was available on my FanPark Baseball Encyclopedia. I took park effects into account in the later part of the study when trying to find the most dominant hitter, but for initial findings of who was the best in each set of years, the amount of data crunching to incorporate park effects would have been prohibitive, so the initial results are based solely on the statistics (RC).

In measuring dominance, I have chosen *to compare the hitter to other great hitters of his day in his league*. The reasoning for this and the exact methods used are explained later in the section called "The Best Peak Hitter - Methods."

I chose to measure only within a league because there have been times when the leagues diverged by offense/pitching-and-defense tendencies, particularly in the 1930s and since the DH rule in 1973.

Peak

What is "peak"? A player's best season? Best X number of seasons throughout a career? What if a player's best seasons are separated by many years with a few not-so-good seasons in between? I chose to answer this by attempting to think like a GM in determining whom I would trade for. How long does it take for someone to establish his current value? Is Sammy Sosa an established 60 home run per year hitter? Are McGwire's back and foot problems behind him, enough that we would project a full season for him? Has Vladimir Guerrero had enough years yet for us to say he is a .330 hitter? My conclusion is this: *A player's peak is his best set of (at least) 5 consecutive seasons*. If it turns out that his numbers are actually better if we

Table 1
Year-by-Year NL RC5 Leaders

Player	Years	Player	Years	Player	Years
Wagner	02-10	Ott	33-37	B Williams	70, 71
Doyle	11, 13	Mize	38-41	Bobby Bonds	72
Magee	12	Nicholson	42, 43	Morgan	73-76
Cravath	14, 15	Musial	44-54	Parker	77
Burns	16, 17	Snider	55	Schmidt	78-83
Hornsby	18-27	Mays	56-60, 62-64	Murphy	84
Wilson	28	Aaron	61, 65, 67, 69	Raines	85-87
Waner	29	Santo	66	Clark	88, 89
Klein	30-32	McCovey	68	Barry Bonds	90-95

use 6 (or 7 or 8) consecutive seasons, then that is what will be used, which happens in some cases. This set of 5 consecutive seasons will establish a player's peak offensive value as of the middle year of the 5.

Yearly Leaders

Starting with the National League, I found the Runs Created leaders for the first 5 year period, 1900-1904. I will call this the "RC5" leader for 1902. Honus Wagner created the most runs for his team in this time. He also led in RC5 for the next 8 consecutive periods, giving him 9 seasons where he would have been judged, according to this method, to be at the time the most valuable offensive player in the NL. Table 1 lists the year-by-year NL RC5 leaders, through the period centered on 1995 (93 thru 97).

Stan Musial had the most total years and the most consecutive years leading in RC5, with eleven. The most interesting line to me is Aaron's, who had to compete with a variety of hitting stars: Mays, Santo, McCovey and Billy Williams. Hammerin' Hank led in RC5 four different times, but never consecutively.

Table 2
Year-by-Year AL RC5 Leaders

Player	Years	Player	Years	Player	Years
Lajoie	03, 06	Williams	40-42, 46-50	Murcer	72
Flick	04, 05	Cullenbine	43, 44, 45	Jackson	73
Crawford	07	Doby	51	Carew	74-76
Cobb	08-18	Rosen	52	Rice	77-80
Ruth	19-30	Mantle	53-62	Brett	81
Gehrig	31-33, 35, 36	Colavito	63	Murray	82-84
Foxx	34, 37, 38	Killebrew	64	Boggs	85-90
DiMaggio	39	Yastrzemski	65-71	Molitor	91
				Thomas	92-95

Table 2 lists the year-by-year American League RC5 leaders. From the table, we can see that Cobb and Ruth owned the first 3 decades (Tris Speaker never gets a mention), Gehrig and Foxx fought a great back-and-forth duel in the thirties (shutting out Hank Greenberg), Joe D. got stuck behind Williams, and there sure are a lot of Red Sox on this list (we'll get to park effects later). Frank Thomas, like Barry Bonds in the NL, was on a roll when I ran out of data after 1997.

Table 3 presents the players who have led in RC5 five or more times in their career:

Table 3
Players who led in RC5 the most often

American League		National League	
Player	Years	Player	Years
Ruth	12	Musial	11
Cobb	11	Hornsby	10
Mantle	10	Wagner	9
Williams	8	Mays	8
Yastrzemski	7	Bonds	6
Boggs	6	Schmidt	6
Thomas	5	Ott	5
Gehrig	5		

I was fascinated by the fact that some of the long-time best hitters played at the same time, across from each other: Ruth vs. Hornsby in the 20s (the Rajah may never have made the list he played in the other league!), and Williams/Musial in the late 40s.

The Best Peak Hitter – Methods

The previous tables showed who the best hitters were for each 5 year period, and who stayed the best for the longest. Now, who had the single best RC5 of them all?

How to measure Offensive Dominance? Pete Palmer derived methods to compare batters to the league average (Linear Weights). Others have attempted to measure a player's value above "replacement level". In measuring dominance, I have chosen to compare the hitter to other great hitters of his day in his league. There are two reasons for this: first, I believe it helps correct for some (not all) of the changes in the game over time. Bill James has noted in his measurements of Offensive Winning Percentage that the top hitters are able to run up better numbers in comparison to league averages during high scoring eras. Craig Wright pointed out in his book *The Diamond Appraised* how in the early days of baseball the league leaders tended to tower over their contemporaries far more than those of today. Also, many historians agree that in the earlier days of play when scouting was not as thorough as today, that many decent players may have been excluded from the majors, thus lowering the "average" playing ability and allowing the best ones to dominate the pack.

So, I have chosen to measure dominance by comparing the BEST hitter to what I will call the "90th percentile hitter"; in a league of 8 teams, this will be about the 8th best hitter (about 10 hitters per team may typically get significant playing time). If we assume that the Sam Crawfords and Paul Waners would have found their way into the majors no matter when they were born, this method will measure how much a truly great hitter is above the "good" ones. I actually use the average of the 7th thru 9th best hitters instead of the 8th, to lessen reliance on one data point. I didn't want to compare the best to only the 2nd best hitter, because that would hurt a player unfairly whenever 2 great hitters played at the same time (Aaron/Mays, Gehrig/Foxx, Cobb/Speaker). The Most Dominant Hitter will be the one whose RC5 is farthest ahead of his competition, after adjusting for his home park. As the leagues expand over time, I change the comparative point to match (Barry Bonds is compared to the 14th best hitter). This helps to control the effects of the growing available population for major league players over the past half-century.

Park effects were measured by calculating the team runs scored and allowed in home and road games for the 5 year periods. If a team scored and allowed 10% less runs at home, then their park effect ratio is 1.10. The RC5 for such a hitter would be adjusted by the formula

```
Park Adjusted RC5 = RC5 * (1 + (Park Effect Ratio – 1) * (N / 2 - 1) / (N - 1) )
```

Where N is the number of teams in the league.

This formula takes into account that a team plays half of its games in its own park, whereas other teams visit park only 1/(2N)th of the time. This is the same formula used in The Baseball Encyclopedia (Reichler, 7th edition, 1988), derived from Pete Palmer's work, much of which can be found in his book *The Hidden Game of Baseball*.

The Best Peak Hitter – The Data

For an example, I will use Mickey Mantle. For the years 1955-59, Mick created 778 runs, for an average of 156 RC per year. The next highest spots in the AL belonged to A. Kaline, T. Williams, J. Jensen, M. Minoso, R. Sievers, H. Kuenn (96), N. Fox (94), and Y. Berra (92). The average of the 7th through 9th hitters was 94 RC. So, Mantle was 62 runs better than the 90th percentile hitter, or 66% better. Another way to say this was that Mantle was 62 runs better than the typical American League All-Star hitter of the late 50s. However, Mantle was also 62 runs above the 8th best hitter in the years 56-60, and 65 runs better in 57-61. If you take the 7 years combined, he has an average RC7 of 154, as compared with an 8th best average RC7 of 84. His totals (70 extra RC) are actually better by extending his peak period from 5 to 7 years, so this is what I will use. This happens because his very best years were those in the mid-fifties and again in 60 and 61, a period which encompasses 7 years. Other truly great batters for whom this phenomenon occurs are Hornsby, Williams, and Bonds (6 years each), and the Babe, who stretched out his peak dominance over 8 seasons.

I computed each RC5 leader's ratio for every year, and recorded the ones that exceeded 1.70. Then I took into account park effects, computing the adjusted RC5, and found the park adjusted RC ratio, as well as the park adjusted extra RC a player accounted for. Table 4 shows the results of the players with the highest adjusted RC ratios.

The BEST Peak Hitter – So Who Was It?

Table 4 lists the players with the highest RC ratios, in chronological order. All-time leading figures are in **bold**. A player is only listed once, in the years he established his own peak, regardless of how many times his RC ratio was over 1.70.

The highest RC ratio (park adjusted or not) belongs to Ty Cobb from 1907-1911, who towered over his dead-ball contemporaries.

The most Extra Runs Created belongs to Ted Williams, in the 6 non-war years from 1941-1949[1].

The highest park adjusted RC belongs to Ruth, who created a monstrous 187 runs a year for 8 seasons. Over 8 years, the Babe was worth 90 more runs a year to his team than 90% of the other hitters in the league!

[1] This was done because Ted (and many others) did not play in 43-45, and so I chose to count the years in which he was eligible to play as consecutive seasons, comparing him to other AL players of 41, 42, and 46 thru 49. Some would argue that this is being overly generous to Mr. Williams, but others would counter that putting up these numbers while taking off 3 years to fight a war makes them even more phenomenal; the reader can decide. His highest "real" RC5 occurred in 1945-49 (even though he played only 4 of the 5 years). The RC ratio was 2.00, park adjusted to 1.90, which still ain't too shabby.

The only player in the past 35 years to make the list? The name is Bonds. Barry Bonds. If I had data for the 1998 season, he may have improved his numbers even further. The park adjustments for him are listed in *italics*, as I could only find park data for 3 of the 6 years. Also, all of the RC data for Bonds are adjusted downward by 5% to fairly compare to a 154-game schedule.

Did I adjust for difference in quality of play over time?

Well, this question creeps up in any study of all-time greats, doesn't it now?

More hitters from the early part of the century show up on this table, which seems to indicate that it was easier to dominate the competition back then, even though I had hoped to alleviate this phenomenon by the method I chose. We could attempt to account for the growing available population of potential Major Leaguers by adjusting the comparison point (8th best hitter in 1910, to maybe the 12th best in 1948, etc.); this could be appropriate, particularly after blacks were allowed to play. The problem with this is we can debate forever exactly what adjustment to make, given population, integration, scouting, foreign players, availability of other sports, amount of outdoor hours playing baseball when they were kids, and who knows what else. But just for grins, here are some "what-if" comparisons:

Table 4
Highest RC5 Ratios Ever (chronological order)

				Raw (NOT Park Adjusted)					Park Adjusted		
Player (my ranking)	League	Years	# of teams	RC5	90% RC5	RC Ratio	Extra RC	Park effects	RC5	RC Ratio	Extra RC
Wagner	NL	04-08	8	138	77	1.79	61	1.01	138	1.80	61
Cobb (1)	AL	07-11	8	153	66	**2.34**	88	0.90	146	**2.23**	81
Hornsby	NL	20-25	8	167	97	1.73	70	1.01	168	1.74	71
Ruth (3)	AL	20-27	8	184	97	1.90	87	1.03	**187**	1.92	90
Williams (2)	AL	*41-49*	8	**188**	86	2.18	**102**	0.89	178	2.07	**92**
Musial	NL	48-52	8	165	96	1.71	69	0.97	162	1.69	66
Mantle	AL	55-61	8	154	84	1.82	70	1.11	162	1.91	77
Bonds	NL	92-97	14	147	85	1.72	62	*1.11*	*154*	*1.81*	*69*

Note: Due to rounding, not all subtractions or divisions may appear to be correct. Trust me, it's rounding.

Ted Williams, compared with the 12th best hitter in his peak years, would pass Cobb (park adjusted RC Ratio of 2.29).

Babe Ruth, compared with the 15th best hitter, would pass Cobb (2.28 ratio).

Mickey Mantle, compared with the 15th best hitter, would pass Cobb (2.27 ratio).

Barry Bonds, compared with the 28th best hitter, would pass Cobb (2.24 ratio).

Is it likely that conditions had changed enough between 1910 and 1946 that there were indeed 50% more good hitters in the majors, which would indicate that Teddy Ballgame deserves to be compared with a different reference point (12th versus 8th) than Tyrus Raymond? If so, how much more do we account for the current game, with full representation of blacks and a huge (23%!) presence of foreign-born players? Your guess is as good as mine.

Observations and Conclusions

Why have I used the *RC ratio* instead of the *extra RC* to rank the hitters? Because in the high scoring eras, it takes more runs to make a win. In Cobb's peak years, the league scored 16% less runs than the AL 1940s of Ted Williams, so in terms of wins, Cobb's marks truly were more valuable.

I was surprised that the answer to the question of best peak hitter was not a resounding "Ruth". I believe there are 2 reasons for this. First, he never had his most awesome seasons together without interruption for injuries (he missed significant time and didn't play as well in 1922 and 1925). Second, even though he towered over the average player of his day and was for a time the ONLY player with awesome power, enough others in his league were able to adjust and drive the live ball for great batting averages, and so he did not dominate the "next best hitters" quite to the extent I had expected. His peak park adjusted RC ratio is "only" the third highest ever, but it does extend over 8 full years. However, if we extend Cobb's peak to his best 8 seasons, his park adjusted RC ratio is 2.04, still ahead of the Babe's.

One perspective on Williams' domination in the 1940s: if you draw up a list of highest Runs Created in a season for the whole decade among AL hitters, Williams collects *all* of the top spots with his 6 year run. No other hitter had any year as good as Ted's *worst* year between '41 and '49! (7th place was Greenberg's 1940, and next was DiMaggio's 1941).

Table 5 compares Cobb, Ruth and Williams' average yearly batting lines with the 8th best hitter in each case.

Table 5
The Most Dominant Hitters compared with other greats of their day

	AB	AVG	OBA	SLG	SB	R	RBI
Cobb 07-11 avg	571	.370	.420	.527	62	111	110
Hal Chase 07-11 avg	486	.288	.316	.364	32	66	60
Ruth 20-27 avg	481	.361	.498	.750	10	135	129
Ed Collins 20-27 avg	477	.347	.437	.445	23	85	64
Williams 41-49** avg	516	.359	.505	.657	2	136	130
M. Vernon 41-49** avg	580	.287	.346	.417	14	77	80

Cobb's raw numbers don't look so hot compared with Ruth or Williams (he looks to me like Tony Gwynn), but you can see how big of a gap exists between him and Hal Chase, who was 5th in both runs scored and RBI in the AL during this period. Think about that – the 5th best hitter in the league back then was scoring and driving in a paltry 63 runs a year.

Honus Wagner's presence in Table 4 confirms that as a good defensive shortstop and one of the best peak hitters, he in his prime was certainly one of the greatest to ever play the game.

In his *Historical Abstract*, Bill James rated Mickey Mantle as the 3rd best peak player ever (behind Ruth and Wagner). The reason Mantle does not rate as high here is that James apparently gave Mantle credit for all of his best seasons, which occurred years apart, whereas by the methods I used Mantle's "good" years were averaged in with his "outstanding" ones. If Mantle had had his best seasons consecutively, he may have had a better prime than any other hitter.

Summary

1. Ty Cobb dominated his league for 5 years like no one else ever has.
2. The Splendid Splinter – his run in the 40s was spectacular. Much of Cobb's value came from his legs as well as his bat, so maybe in terms of peak HITTER, Ted really was the best that ever lived.
3. Ruth was almost as fearsome, for 8 full seasons. He also had great post-season stats in these years, unlike the other top two listed here, whose teams didn't win a World Series in their peak periods. These 3 stand head and shoulders above the rest, and any of them could arguably be #1.
4. Barry Bonds in the mid-nineties reached a level of offensive performance that has not been seen in well over a generation.

Tom Hanrahan, 21700 Galatea St., Lexington Park, MD, 20653, HanrahanTJ@navair.navy.mil. ♦

ARM – Average Run Equivalent Method

Clem Comly

While it is difficult to accurately evaluate outfielder arms through the use of traditional statistics, the availability of play-by-play data now makes such evaluation possible. Here, the author examines the data, introduces a method to compute how many runs were saved by the throwing prowess of a given outfielder, and presents lists of bests and worsts.

I would like to present my findings on a method to analyze more completely the ability of outfielders to shut down the running game. In the past OF assists and anecdotal evidence were used. More recently, the STATS Scoreboard has presented the percentage of runners who have taken the extra base when the opportunity was provided. Let's call this new method ARM (Average Run equivalent Method). ARM takes into account assists, extra bases taken by either the batter or the runner, and errors made by the outfielder whenever a single is fielded with runners on first and/or second (regardless of the third base runner situation). ARM, which requires play-by-play data, was used for most of the major league games from 1959 to 1987 (thanks to Retrosheet and the Baseball Workshop). Data for different OF positions are kept separate. Among other questions ARM could let us answer is what is the difference between having Greg Luzinski in LF rather than Carl Yastrzemski for a full season (and the answer is not 43 pounds despite what TB says).

First, we use play-by-play to identify situations where we know a single was hit to a specific outfielder. ARM notes his name, the outs and the runner positions before the hit, and the result. If there was no runner on first nor on second, ARM discards the result. If there is also a runner on 3B, 99% of the time we throw that runner away but keep the result of the other runners. The less than 1% of the time where the OF throws out the runner from 3B at home, ARM treats as if the runner on 3B had been on second. Any subsequent infielder error or pick-off of a runner is not recorded as the actual result, but instead a best guess of the result without that extraneous play. The OF gets no credit for the out on the bases unless he gets an assist, but he does get credit for an out when he gets an assist but the runner was actually safe when an infielder dropped the throw for an error.

So in effect there are 9 starting states: 3 out possibilities multiplied by three runner situations: man on first (called “1” below), man on second (“2”), and men on first and second (“12”). For each event from the play-by-play, ARM records the resulting runner positions, any out that was made by the OF’s throw (+ below), and any runs scoring (- below for one run scoring). When ARM has finished for a particular OF for a particular position, ARM uses Pete Palmer’s expected runs (see *The Hidden Game of Baseball*, p.153) for the resulting out and base situation and adds 1 for each run scored. Each such expected run result is recorded, and the league average for that OF and that number of outs is subtracted. The sum of all these differences is the ARM total – the number of runs that outfielder saved (or allowed) over the league average. A negative number means the outfielder saved runs (since fewer runs than average scored), and a positive number means that the outfielder cost his team more runs than average.

In the period 1959-1987, for instance, two center fielders with similar raw numbers are shown below:

Name	S	MS	Singles w/ 1/2/12	Assists on those singles
Rick Miller	1313	116	539	16
Jose Cardenal	1342	164	561	19

To interpret the headings, in that period the play-by-play showed 1313 singles hit to CF while Miller was playing there. Some singles in the play-by-play data are anonymous in terms of who fielded them. While Miller was in CF, there were 116 that on a pro rata basis were hit to him which I will call MS (missing singles). (These 116 are strictly to show the level of accuracy of the play-by-play and are not included in the 1313.) ARM looked at the 539 singles of the 1313 that happened with runners on first and/or second regardless of the runner on 3B. Miller garnered 16 assists after fielding those 539 singles. Looking at Cardenal, both his assist total and his singles fielded are a little higher. Through these traditional stats, the two outfielders’ arms look roughly equal.

But the ARMs are significantly different. Calculating the ARM for Miller on those 539 hits gives –9.0, so he saved his team 9 runs for his career from 1959-1987 compared to the baseline CF. Cardenal’s ARM works out to +5.2, meaning his arm allowed 5.2 more runs to score than the average centerfielder’s arm would have allowed.

So even though the two players’ traditional numbers look roughly the same, Miller’s ARM is much better, by over 14 runs. What happened?

We can find out by first breaking down the three base runner starting positions which, allowing for rounding, add up to the ARM:

	Net	=Net	+Net	+Net
	Runs	1	2	12
Rick Miller	-9.0	-2.7	-2.7	-3.5
Jose Cardenal	+5.2	+9.6	-3.6	-0.7

Cardenal's problem, for the most part, is due to his performance with a runner on first without a runner on second. What was the problem? Let's add together the 0 out, 1 out, and 2 out events and compare them to get an idea.

Here's the breakdown, by finishing situation after the single. (The headings are final baserunner positions – so "13" means runners on first and third – while the "+" means a runner was thrown out, and a "-" means a run scored.)

		Result situations								
Name	Opp	1+	2+	3+	12	13	23	1-	2-	3-
Rick Miller	292	3	0	1	194	81	4	0	8	1
Jose Cardenal	288	5	1	3	127	126	16	5	4	1

Cardenal threw out more runners (9 (5+1+3) to 4 (3+0+1)) and was only slightly worse in allowing the runner to score from first (10 (5+4+1) to 9 (0+8+1). Cardenal's problem was the runner was going to third almost half the time (49%, (126+16)/288) while Miller was at 30% ((81+4)/292). ARM balances these factors and shows Miller is more valuable.

Let's look at a gold glove versus a hitter, Yaz versus Luzinski:

Yastrzemski

Season	GS	S	MS	TS	ARM	A	
1961	146	238	16	111	0	6	
1962	160	192	89	87	-1	5	
1963	150	197	71	74	-2	6	
1964	16	30	3	11	-1	1	primarily CF
1965	115	111	100	46	-3	4	
1966	151	160	91	63	-8	8	
1967	157	194	30	82	-7	8	
1968	152	212	44	88	-3	5	
1969	138	248	10	108	-7	8	
1970	64	126	7	52	-1	0	primarily 1B
1971	144	272	9	127	-9	11	
1972	82	139	26	71	-1	4	
1973	15	36	5	13	1	0	primarily 1B
1974	62	82	12	31	0	0	primarily 1B
1975	8	4	10	1	0	0	primarily 1B
1976	51	113	10	44	-2	2	primarily 1B
1977	138	209	57	87	-7	9	
1978	63	102	17	47	-1	2	
1979	34	56	6	24	-3	1	primarily DH
1980	32	34	6	11	1	0	primarily DH
1983	1	2	0	2	-0	0	primarily DH
ALL	1879	2757	630	1010	-56	79	

Luzinski

Season	GS	S	MS	TS	ARM	A
1972	145	145	32	47	2	2
1973	157	199	36	89	-1	5
1974	81	128	3	51	-1	4
1975	159	255	2	111	1	6
1976	144	206	1	71	2	3
1977	148	200	0	86	2	3
1978	154	195	0	70	-3	2
1979	124	185	0	74	3	1
1980	105	148	1	61	2	2
ALL	1217	1661	82	660	9	28

Key to Tables

GS: Games Started in LF
S: singles fielded
MS: prorated unidentified singles
TS: singles fielded with a runner on 1B and/or 2B regardless of 3B (excluding MS)
ARM: equivalent runs minus base line LF
A: assists on singles fielded with man on 1B and/or 2B regardless of 3B

Both were starting in LF at age 21 (below are seasons primarily at LF):

												Ave
Yaz ARM	0	-1	-2*	-3	-8	-7	-3	-7	-9	-1	*	-4.1
Yaz A	6	5	6	4	8	8	5	8	11	4		6.5
Luz ARM	2	-1	-1	1	2	2	-3	3	2			1.3
Luz A	2	5	4	6	3	3	2	1	2			4.0

* five years later Yaz will go back to LF

ARM suggests that Yaz was a better throwing LF than Luzinski to the tune of over 5 runs per season.

Now, let's look at the best and worst single-season ARMs for each position:

Best and worst single season ARMs

LF			CF			RF		
1978	S.Hendersn	-10.7	1976	Beniquez	-12.4	1963	Callison	-10.9
1978	Cromartie	-10.6	1980	O.Moreno	-10.7	1974	G.Gross	-10.6
1985	J.Leonard	-10.5	1983	E.Milner	-10.6	1978	E.Valentn	-10.3
1971	Yaz	-9.5	1978	Dawson	-9.2	1985	Barfield	-10.3
1983	J.Leonard	-8.8	1982	Dw.Murphy	-9.2	1986	vanSlyke	-9.7
1982	Lon.Smith	-8.5	1974	Geronimo	-9.1	1987	Barfield	-8.9
1973	Stargell	-8.3	1972	Unser	-9.0	1977	J.Clark	-8.3
1980	LeFlore	-8.0	1968	Berry	-8.2	1973	K.Singltn	-8.2
1974	Rose	-7.6	1980	Dw.Murphy	-8.2	1986	G.Wilson	-8.0
1966	Yaz	-7.7	1973	Cedeno	-7.7	1987	G.Wilson	-7.7
...			...			...		
1971	F.Howard	4.4	1961	K.Hunt	4.0	1984	C.Washngtn	4.8
1961	Minoso	4.4	1967	Pepitone	4.0	1980	G.Matthews	4.8
1967	J.Alou	4.5	1966	CleonJones	4.1	1964	Christopher	4.8
1978	Page	4.7	1968	Reg.Smith	4.3	1975	Burroughs	5.4
1982	Winfield	4.9	1969	Reg.Smith	4.6	1980	Griffey	5.6
1965	F.Howard	5.2	1965	Flood	4.9	1960	J.Cunnghm	6.3
1968	F.Howard	5.1	1962	Bruton	5.0	1969	T.Coniglro	7.1
1975	Kingman	5.1	1968	T.Gonzalez	5.0	1969	K.Harrelson	7.2
1963	L.Wagner	6.0	1970	Cardenal	5.2	1960	Allison	7.3
1964	L.Wagner	6.1	1959	Ashburn	5.8	1977	Burroughs	7.5
1968	R.Allen	6.3	1964	Cowan	5.9	1967	Swoboda	7.8
			1983	G.Thomas	6.9			

And here are the career bests and worsts:

Career Best and Worst ARMs by Position

LF ARM career 1959-1987

Name	S	MS	ARM
Yaz	2757	630	-56
J. Rice	2006	958	-26
Wil.Wilson	1103	48	-24
Stargell	1507	314	-21
J.Leonard	643	490	-21
Lon.Smith	740	418	-20
Raines	891	346	-19
R.Henderson	1050	270	-18
Cromartie	649	142	-17
George Bell	472	365	-15
Oglivie	1783	468	-15
Page	386	119	7
H.Lopez	686	36	8
Hinton	628	39	8
Brock	3366	308	8
Luzinski	1661	82	9
Al.Johnson	1448	64	9
R.White	2613	181	10
Baylor	994	111	10
Covington	710	60	12
L.Wagner	1228	273	21
F.Howard	1290	19	22

CF ARM career 1959-1987

Name	S	MS	ARM
Dw.Murphy	1461	616	-38
Cedeno	2165	632	-35
Geronimo	1557	133	-29
Dawson	1652	272	-29
G.Maddox	2609	375	-28
Blair	2472	343	-27
O.Moreno	1966	240	-26
W. Mays	3206	618	-26
Dal.Murphy	1207	703	-25
Willi.Davis	3740	201	-24
Del Unser	1995	94	-23
Mota	306	54	5
K.Gibson	234	48	5
Hisle	911	90	5
Cardenal	1342	164	5
J.Briggs	476	13	5
Cowan	363	19	6
Landis	1556	292	6
Pepitone	579	10	7
Lenny Green	922	115	7
T.Gonzalez	1503	78	8
Ashburn	608	39	12

RF ARM career 1959-1987

Name	S	MS	ARM
Callison	2189	55	-39
Barfield	854	437	-38
Clemente	2257	478	-34
J.Clark	1324	148	-28
E.Valentine	1058	174	-26
Dw.Evans	2163	912	-25
O.Brown	1030	148	-23
Glen.Wilson	606	368	-20
Hank Aaron	1775	352	-18
Parker	2157	511	-18
Winfield	1784	611	-17
Al Cowens	1627	483	-17
G. Gross	497	141	-16
M.Hershberger	647	296	-16
Pinson	648	60	11
K.Harrelson	333	32	11
Singleton	1571	285	12
J.Cunningham	401	11	12
Murcer	1161	78	13
Fr.Robinson	1355	226	15
C.Washington	1164	435	15
Allison	822	20	16
Burroughs	1108	83	20

Conclusions

The key point of this study is we now have an idea of how much outfield throwing talent can be worth. The difference from absolute best season to worst is about two victories (around 20 runs). This is a combination of:

- the limited number of opportunities in a season (usually 50-100 singles with a runner on 1B and/or 2B),
- the talent is one of degree not kind, and
- the refusal of managers to keep putting a real rag arm in the field.

From season to season, even the best don't average saving their teams even one victory compared to an average OF. Of course, the average RF in 1964 was 10% Callison and 10% Clemente. Speaking of Clemente, a Pirate pitcher was quoted on SABR-L as saying that Clemente liked to show off his arm by throwing to third base where there wasn't a play, allowing the batter to get to second and thus hurting the Pirates. Clemente's ARM reflects batters taking second, so overall he didn't hurt the Pirates (as his Gold Gloves also attest to).

As Retrosheet holdings expand towards the present and further into the past, ARM can be calculated for more outfielders in baseball history. This methodology can be used for other studies; the one that most immediately springs to mind is baserunner evaluation. Obviously, the average result of a single to LF with a runner on 1B is the result for the average left fielder is in large part the result of the average runner on 1B.

Disclaimers

I purposely chose singles only because extra base hits are much more a function of the ball park. Also, the sample size is low for singles, for doubles and triples... ARM is limited to the accuracy of the play-by-play data files, which no CPA would sign. More than 95%, but not 100%, of the games were available for the period 1959-1987. There may also be a slight bias that anonymous singles will tend to be hits on which no assist or error occurred. The baseline RF, CF, and LF were the averages of 3 ML seasons, 1961, 1966, 1968, and the NL of 1962 and 1969 (which were chosen because of availability at the beginning of the project). For most seasons, the sum of all outfielders is a little better than zero. If that bothers you, consider the baseline a replacement level player. I call this "ARM", but it also measures the judgment of the OF and his ability to get into position to throw. Obviously, there may have been some singles that Yaz left his feet on that Luzinski let get by him for extra bases. ARM punishes Yaz for letting the runner go to third on those singles. Defensive average or range factor, which ARM supplements and does not replace, should reward Yaz for that play.

Clem Comly, 308 Colonial Drive, Wallingford, PA, 19086-6004; ccomly@erols.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@mediaone.net. Or write to him at 500 Market St. #11L, Portsmouth, NH, 03801.

Run Statistics Don’t Work For Games

Phil Birnbaum

Run estimator statistics, such as Linear Weights and Runs Created, have been used to predict runs scored for players and for seasons. It is also assumed that they are unbiased estimators for individual games. But this is not so. Here, the author explains why any statistic that is unbiased for seasons will not be equally accurate for games.

A few issues ago, I ran a study where I simulated a couple of thousand 162-game baseball seasons, and checked which of the big three predictor statistics – Runs Created, Linear Weights, and Estimated Runs Produced – was the most accurate. After that study appeared, a couple of people asked me why I chose to use a simulation. Instead, they suggested, why didn’t I just use actual major league games? Why not just check all the real-world games where teams hit .250 and slugged .400, for instance, to see which statistic most accurately predicted the number of runs that scored in those games?

The reason I didn’t do that is that it wouldn’t have been a fair test.

Suppose that over the course of a season, a team (the Tigers, say) has 1,458 hits, or an average of 9 a game. How many runs will they score?

Well, that depends on how the hits are bunched. If the hits come exactly one an inning, the Tigers will leave a lot of runners stranded, and score maybe 1 or 2 runs per game. At the other extreme, if all 1,458 hits come in the same inning, Detroit will score about 1,455 runs, or about 9 per game. Run scoring depends on how the hits are bunched into innings.

The same is true for bunching by game. The most runs score if all Detroit’s hits come in one game – if that happens, they’ll score over 1400 runs. The fewest runs score if their hits are divided exactly equally among the 162 games – exactly 9 hits each and every game.

Of course, the Tigers won’t be at either extreme -- they’ll be somewhere in the middle. Some games, they’ll get one or two hits, and some games, they’ll get 15 or more. They’ll score fewer runs than if their hits were bunched up into one game, and more runs than if their hits were spread evenly over 162 games.

When you “ask” runs created to give you a season estimate, you’re asking, “how many runs would this team score if they average 9 hits a game, but often get more and often get less?” But when you ask for an estimate for individual games, you’re asking “how many runs would the team score if they get exactly 9 hits each and every game?”

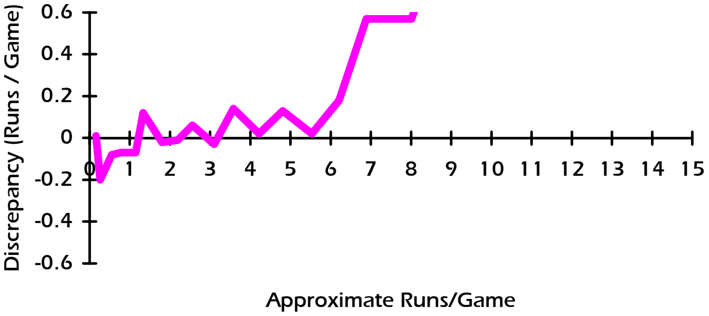
As we have seen, the answer to the second question is smaller than to the first question, and so there’s no way the same statistic can answer both. If a statistic is accurate when applied to seasons, there’s absolutely no way it can be accurate applied to individual games.

Because individual games have no bunching, any statistic that predicts accurately for seasons will predict too high for games. (The estimate will be the same, but the actual runs will be smaller for the games.)

But how big is the effect? Is it one run per game, a half-run, a tenth of a run?

Runs Created

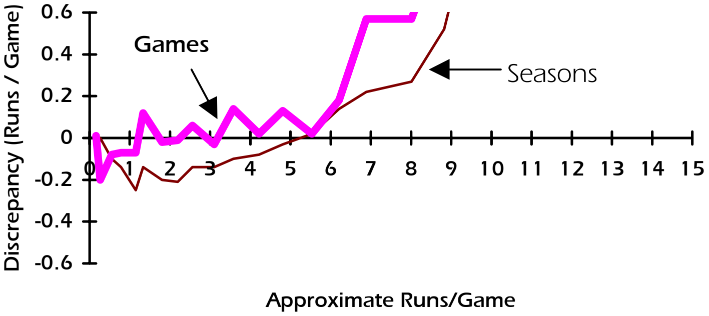
Using Retrosheet and Project Scoresheet data for the 1983 and 1988 American League, I calculated Runs Created estimates for each individual game, and compared them to actual runs scored. Then, I broke up the games into high-offense games and low-offense games, by on-base percentage plus slugging percentage (OPS). Here’s a graph of the results:



To make things easier to read, I used approximate runs on the X-axis instead of OPS. (For instance, 7 runs per game actually corresponds to an OPS of about .900.) The Y axis is predicted runs minus actual runs. For games in which 1 to 6 runs were expected to score, Runs Created was extremely accurate. But for games in which 7 or 8 runs were expected, Runs Created predicted over half a run too high.

So what's going on here? If predicting for games is supposed to yield estimates that are too high, why aren't they too high? Why do the estimates for 1-6 RPG appear so accurate?

The answer: because Runs Created normally predicts too *low* for seasons in the 1-6 range. Here's the same graph again, but superimposed on Runs Created for seasons (taken from my other study):



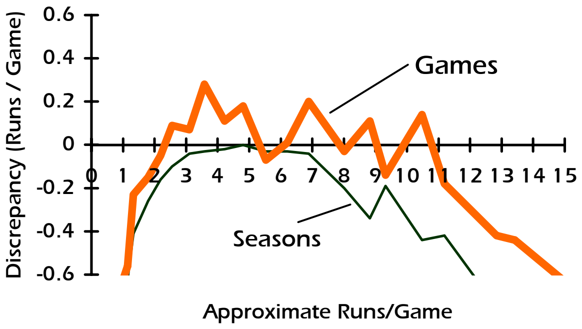
In almost every case, the discrepancy for games is more on the positive side than the discrepancy for seasons. (Remember, the RC estimates are the same, but the *actual runs* are smaller.) The difference seems to be in the 0.1 to 0.2 range for low-offense games. For high-offense games, Runs Created gives season estimates that are far too high, and game estimates that are even farther too high.[1]

[1] Bill James, in describing Runs Created in the 1982 *Abstract*, likely did not realize there is a difference in accuracy between seasons and games, and implied that RC is equally accurate for game predictions. "You can take the box score of a game … and project the scores for each team. The team which should win according to these projections … will win over 80% of the time. … if a team should score 5 [runs], they will actually score 4, 5 or 6 about 70% of the time." (p. 7)

James' statement is still true, mostly because the discrepancy (about a fifth of a run per game) is small compared to the number of runs that score.

Linear Weights

Here's the same graph, but for Linear Weights:

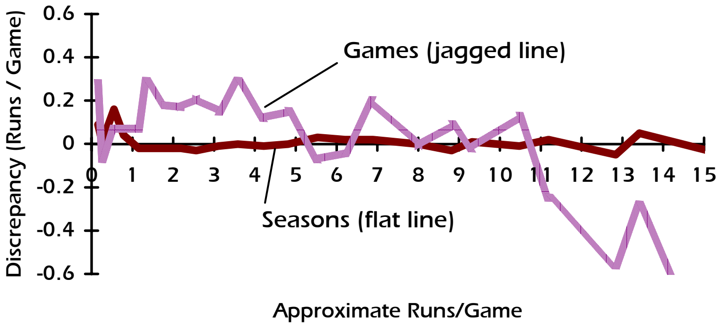


Here, the effect is almost exactly what we expected – at every point on the offensive spectrum, Linear Weights for games is more optimistic than for seasons. Again, the effect seems to be about a fifth of a run per game.

More specifically, where Linear Weights is almost exactly accurate for seasons – between 3 and 7 runs per game – it's consistently too high for games.[2]

Ugly Weights

Ugly Weights was my attempt to use regression to come up with a stat that's accurate over a wider range of offenses. Here's the graph for Ugly Weights:



While Ugly Weights is almost perfectly accurate for seasons, the estimates for games are a little too high. The effect seems to reverse itself for high-offense games. This might happen if at the high end, the real-life games are different from the simulated games – that is, the simulation might create high-power games, while real life includes more high-average games.

[2] Estimated Runs Produced, the Paul Johnson statistic, would likely show very similar results, because it's almost the same stat. This similarity was discussed in my previous study.

Summary

Run estimator statistics that work for seasons work because they assume a typical distribution of high-offense and low-offense games. But for any offense as a whole, the more variation there is in per-game offense, the more runs score. And, therefore, using any run statistic for individual games of a given offensive level – where the between-game variation is zero – will lead to skewed results.

And so any statistic that accurately predicts runs for a season will not accurately predict runs for a game.

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

Book Reviews Wanted

Every year, a number of books and magazines are published with a Sabermetric slant. Many of our members have never heard of them. Our committee members would like very much to hear when this kind of stuff comes out.

If you own a copy of any baseball book of interest, we'd welcome a summary or a full-length review. The only restriction, please: the book should have, or claim to have, some Sabermetric content.

For a sample of what we're looking for, check out Gabe Costa's review of "Baseball Dynasties" elsewhere in this issue.

Send reviews to the usual place (see "Submissions" elsewhere in this issue). Drop me a line if you want to make sure no other member is reviewing the same publication, although multiple reviews of the same book are welcome, particularly for major works. Let me know which book you're doing, so I don't assign the same book twice.

And if you're an author, and you'd like to offer a review copy, let me know – I'll find you a willing reviewer.

The Recipe For A Stolen Base

Sig Mejdal

There are many factors that go into the outcome of a stolen base attempt – the runner's speed, the catcher's arm, the pitcher's delivery, and even the umpire and the stadium surface. In this study, the author examines how much explanatory power these factors have on the outcome, and which are the most important.

The score was 0-0 in game three of the World Series as Steve Finley stepped off first base, carefully eyeing the pitcher David Cone, as he inched his way towards second. Meanwhile behind the plate, catcher Joe Girardi readied himself for the impending throw to second. Cone delivered as Finley broke towards second. Girardi caught, turned, and hurled the ball to second base…the tag was applied…and alas; the safe sign was given by umpire Tim Tschida.

This seemingly routine steal of second made me wonder exactly what factors led to that safe call by Tim Tschida. In other words, "What influences the outcome of a stolen base?"

We know that Finley is a very good base stealer. Over the last three years, he has been caught only 81 times in nearly 330 attempts – a 75% clip. Certainly Finley's base stealing ability had something to do with the outcome. Behind the plate is Girardi. Attached to Girardi's shoulder is a below average throwing arm (runners stole on him at an 80% rate). It's a safe bet that he had something to do with it. Now, on the mound is David Cone. He must have had something to do with the outcome. After all, he was holding the runner on, maybe occasionally making a pickoff attempt, and it was the leg kick of his delivery that acted like the starter's pistol in this race. As if that wasn't enough, the replay revealed that Finley was actually tagged before he reached the haven of second base. Albeit a very close one, Tschida blew the call. Could there be a tendency for some umpires to consistently slant their safe/out calls? Maybe Tschida "safe/out tendencies" should be thrown into the soup as well. And while we are at it, it was hard not to notice that Finley ran the whole way to second on dirt -- certainly he could have got there more quickly if he had the benefit of running on artificial turf. Let's include the surface as well.

So our ingredients are baserunner, catcher, pitcher, umpire, and surface. Our goal is to find out how important each one is to the results of a stolen base attempt. The stolen base data is out there. We have (or can get) stolen base percentages for baserunners, catchers, pitchers, umpires and surfaces. Clearly the biggest hurdle in our way to determining the correct recipe is the confounding effects between pitchers and catchers on the same team. That is, pitchers are paired with the same backstop the majority of the time and it is difficult to separate their individual contributions. The last few Piazza years in LA is a good example. The Dodger pitchers were almost always paired up with Piazza. So whether Piazza is exceptional or poor at throwing out baserunners, the Dodger hurlers stolen base data is going to be influenced by him. Similarly, and more confounding, is the simultaneous benefit that Piazza may be getting from the staff if they are, on average, good at halting the stolen base.

It is not intuitively easy to figure out how to solve this problem. The pitcher is influenced by the catcher, but then the catcher is also influenced by the remaining staff. The goal, however, is obvious – a measure for the pitcher and a measure for the catcher that are both accurate and independent of each other's skills. As we have it now, the pitchers SB rate is really a measure of the pitcher's ability to hold the runners on, and the arms of the catchers that he has been paired up with.

One step towards a solution to this would be to look at a larger sample – say 3 years. During this time, catchers are teamed up with so many different pitchers that it is very unlikely that they have a significantly "unbalanced" sample of good or bad pitchers. Indeed, a correction to adjust the catchers percentage by the average pitchers SB% that they have teamed with showed no significant change to the catchers numbers. In general, over a 3 year sample, catchers team up with so many different pitchers that their stolen base percentage is an "unconfounded" measure of their skill. Indeed, it is more accurate to say that the pitcher's data is contaminated/confounded by the catchers and not vice-versa.

Of course, to get an "independent" measure of the pitchers skill, we must account for the catchers influence by subtracting the catchers "worth" from the pitchers results. In other words, if Cone had been paired up with catchers that averaged a 60% stolen base success rate, then we would expect him to give up stolen bases at a 60% rate IF he were completely average. Now if he outdoes this estimate (e.g. runners only steal 40% off of him), then this difference (the 20%) can be attributed to him. I'll call it the pitcher's stolen base value.

Thanks to the Retrosheet play-by-play data available, I was able to collect and analyze every steal of second base attempted (more than 10,000) in the 3-year sample. Information on the stolen base percentages of the baserunner, catcher, and umpire, as well as the type of surface and the pitcher's stolen base value, were put into a computer. (Of course, so as not to be influenced by the results of the particular SB attempt in question, the numbers were calculated for all the other attempts in which they participated in during the 3 year sample.) Got it? For instance, if Tim Raines has just attempted to steal, I have in the database his percentage for all the other attempts he has had over the 3 years. Similarly, the pitcher's, catcher's, and umpire's percentages for all their other attempts are also included.

If you could imagine a very long table (or spreadsheet) of numbers divided into six columns, you would see what the computer had to work with. The first four columns contain the SB percentages for the baserunner, catcher, pitcher, and umpire involved in the particular attempt. The 5th column contains information on the surface (turf or grass), and the 6th column has the result of the SB attempt (1 for safe, 0 for out). Each row in this table corresponds to a particular SB attempt in the 3 year sample. Hence a table with a whopping 10,000 rows. I politely asked the computer to go through each and every row and determine what the best way to weigh the measures of the runner, catcher, pitcher, umpire and the surface were, in order to predict the outcome of the stolen base attempt as accurately as possible. By analyzing all the data, not only can the computer determine which measures have nothing to do with the outcome, but it can compute just how much each measure contributes to the outcome of the SB attempt. That is, we can look at the partial correlation constants and determine just what effect the baserunner, catcher, pitcher, umpire and the surface have upon the stolen base outcome.

I am sure that most of you have been exposed to a *coefficient of determination* (r^2) and a *correlation coefficient* (r). Briefly, if all of the variation of a dependent variable (i.e. the variable that we are trying to explain) is explained by the independent variable (the variable that we believe may have an effect on the dependent variable) then a correlation coefficient (and a coefficient of determination) of 1.00 exists. If none of the variation is explained, then a correlation coefficient of 0.00 is calculated. For instance, we may suppose that team salary has something to do with team wins. We can tabulate and analyze the data in order to determine how much of the variability between team wins can be explained by the variability from team salary. For instance, we may calculate a coefficient of determination (r^2) of .50 which suggests that half of the variability between individual team's success is explained by team salary. The correlation coefficient is actually the square root of the coefficient of determination, so we can also calculate a correlation coefficient of .71 (the square root of .50).

As in our case with the stolen base attempt result, we have multiple independent variables (baserunner, catcher, and the pitchers skill, along with the umpires tendency and the surface) that may influence the dependent variable (outcome of the stolen base attempt) instead of just one variable. So now, we will look at a *partial correlation coefficient* instead of a correlation coefficient. The partial correlation coefficient is a measure between the dependent variable (stolen base result) and one particular independent variable when all other variables involved are kept constant; that is, when the effects of all other variables are removed (often indicated by the phrase "other things being equal"). It is the unique contribution of the particular independent variable to the prediction of the dependent variable. In other words, it is the effect that each of our "ingredients" have upon the stolen base.

The resulting listing gives the partial correlation coefficients[1] for each of our independent variables.

```
Partial correlation coefficients

Baserunners SB%      .16
Catchers SB%         .08
Pitchers  SB value  .15
Umpires SB%          .00
Surface              .05
```

Another way to think of these coefficients is as relative weights of contributions to the outcome. Of the variation that can be explained by this model, the table below gives a percentage breakdown for each factor:

```
Percent of explainable variation

Baserunners SB%      36% of the explainable influence
Catchers SB%         19%
Pitchers  SB value  34%
Umpires SB%           0%
Surface              11%
```

[1] You will note that I did not include the actual regression formula. This was done on purpose for a couple of reasons. First of all, the magnitude of the numbers associated with each of the "ingredients" is not the best indicator of the contributing effects of each. That is, the magnitude of the regression coefficient is influenced by the variability of the measurements themselves. I wanted to discover the relative importance of each factor, not necessarily generate a predictive formula. Secondly, I conducted a linear multiple regression, while a logistic multiple regression would have been more appropriate since we have maximum and minimum allowable values (i.e. 100% and 0% predictive range for the stolen base success rate). The relative values of the partial correlation coefficients will be nearly the same; however, the linear regression equation would generate nonsense predictions (110% success rate) for some relatively extreme combinations of players. For those reasons, I left it off for now.

Not surprisingly, the Baserunner has the biggest influence on the outcome. But, believe it or not, just a hair behind him is the pitcher's influence. This means that the two most important bits of information to have if we want to discover the likely outcome of a stolen base attempt are measures of the baserunner's skill and the pitcher's skill – not the catcher's skill. It is quite enlightening to discover that the catcher's arm is just slightly more than half as important as either of these effects. In fact, the surface alone is about half as influential as the catcher's arm. As you can see, the data reveal that the umpires have nothing to do with it. Sure, some umps may have made 75% safe calls while other made 65%, but these differences mean nothing – just normal deviations that come from sampling.

So next time you see Finley running on Cone, with Girardi behind the plate, and Tschida umpiring second in San Diego, remember that the matchup you are really witnessing is the Baserunner Finley vs. the Pitcher Cone…with a little help from the Catcher Girardi and a tiny bit of influence from San Diego's playing surface.

Sig Mejdal, smejdal@montereytechnologies.com. ♦

Receive BTN by E-mail

You can help save SABR some money, and me some time, by receiving your copy of *By the Numbers* by e-mail. BTN is sent in Microsoft Word 97 format; if you don't have Word 97, a free viewer is available at the Microsoft web site (http://support.microsoft.com/support/kb/articles/Q165/9/08.ASP).

To get on the electronic subscription list, send me (Phil Birnbaum) an e-mail at birnbaum@sympatico.ca. If you're not sure if you can read Word 97 format, just let me know and I'll send you this issue so you can try

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

General Relativity Using POP: Comparing Apples and Grapefruit

Mike Sluss

In a previous article in BTN, the author presented POP, which evaluates a batter by the likelihood of a league-average player duplicating the batter's performance. Subsequently, a Rob Wood article suggested that using a random player, rather than a league-average player, produces a more meaningful result. Here, the author responds to that criticism, arguing that there are both intuitive and technical reasons for prefering the original POP.

Introduction

The Probability of Performance (POP) method of analyzing baseball players' performances has been presented previously[2][3] and has been critiqued by Rob Wood.[4] I appreciate Rob's review, in which he has made some interesting observations and suggested a revision for POP. This paper will review the basic concepts of POP, comment on Rob's observations and proposed revision (RW-POP), and show the advantages of using POP rather than the traditional means of making relative comparisons, even across eras that have widely differing (apples and grapefruit) levels of performances.

Overview

The basic premise is that a less probable performance is a more outstanding performance. POP specifically asks how likely a given player's performance would be achieved by a league's average hitter, defined as the league average. POP is calculated using the binomial probability function, which is well suited to determine the probability of events such as hits per at bats. This calculation considers the amount of success (hits) and the rate of success (batting average), as well as the league batting average. POP is also well suited for cross era comparisons of other performances, such as walks, home run hitting, on base performances, base stealing, and slugging averages.

Calculation of POP is a two step process. By using the binomial probability function, the probability that an average player (with an average league batting average of L=(league hits - H)/(league at bats - A)) would get at least H hits in A at bats (for a personal batting average of H/A) is P, where

$$P = \sum_{h=H}^{A} \frac{A!}{\left(h!(A-h)!\right)} L^{h} (1-L)^{A-h}$$

and

$$\text{POP} = -\log_{10}\text{P}$$

Rob Wood has suggested that when comparing a player's performance to the rest of the league, more of the league distribution should be used than just the mean. He proposes these equations for RW-POP

$$P_{RW} = \sum_{j} \left(\frac{Aj}{AL} \left(\sum_{h=H}^{A} \frac{A!}{h!(A-h)!} Avgj^{h} \left(1 - Avgj\right)^{A-h} \right) \right)$$

and

$$\text{RW-POP} = -\log_{10}\text{P}_{\text{RW}}$$

[2] Mike Sluss, "POP Analysis of Career Walks", *By the Numbers*, November 1999, pp. 5-6.
[3] Mike Sluss, "Probability of Performance: General Relativity for the Baseball Fan", *Baseball Research Journal*, 1999, pp. 124-129.
[4] Rob Wood, "Probability of Performance - A Comment", *By the Numbers*, May 2000, pp. 7-11.

where j = the index for all league hitters other than the hitter for whom RW-POP is being calculated, Aj = j's at bats, AL = league at bats, and Avgj = j's batting average.

To Bayes or not to Bayes, is that the question?

I disagree with Rob's assertion that POP presumes that "all players in the league have equal abilities." Neither POP nor RW-POP is a Bayesian method. League average is not to be considered a "prior probability". POP should not be stuffed into the mold of Bayesian statistics and then be criticized that it doesn't fit the model well.

Far from "believing" that Bill Terry was a league average hitter, POP calculates the exact probability that a hitter, with a .310 chance (the league's batting average) of getting a hit each time at bat, will get at least 254 hits in 633 at bats (Terry's 1930 totals). According to POP, this probability that an-average-hitter-would-randomly-do-at-least-as-well-as-Terry is slightly less than 1 chance in a million.[5] POP proves that Terry was not an average hitter. POP provides a measure of how much above average he is.

Contrast the above probability with RW-POP's calculation of about 1 chance in 50 as "the likelihood of a player chosen at random [from the 1930 NL] would achieve the league leader's [Terry's] batting average". To me this seems to be an unreasonably high probability which barely exceeds a p =.02 threshold for believing that Terry's performance was above average.

Hypothetically Speaking

Enough of dueling intuitions.[6] Let's look at Rob's hypothetical leagues, League 1 and League 2. The data in each league is the same as Rob's examples. I have added POP and RW-POP for each player.

League 1

player	H	AB	BA	LA	POP	RW-POP
A.	155.	500.	0.310	0.254	2.56	2.54
B.	128.	500.	0.256	0.272	0.10	0.20
C.	127.	500.	0.254	0.273	0.08	0.17
D.	126.	500.	0.252	0.273	0.06	0.14

League 2

player	H	AB	BA	LA	POP	RW-POP
F.	154.	500.	0.308	0.240	3.50	0.92
G.	150.	500.	0.300	0.243	2.70	0.65
H.	120.	500.	0.240	0.263	0.05	0.18
I.	90.	500.	0.180	0.283	0.00	0.00

Basically, Rob argued that A's achievement should be considered greater than F's, but that POP shortchanges A because POP "ignores the spread and all the rest of the players." Actually it is not obvious to me that F did more poorly. It can be argued that F's 3.50 BA POP is higher than A's 2.56 BA POP because F achieved nearly identical hits and BA (154, .308) as did A (155, .310), but F achieved this in a league in which it was more difficult for others to hit, on the average (2.40 league average for F, 2.54 league average for A).

The flaw in making generalizations when looking at Rob's hypothetical League's 1 and 2, as though they were analogous to baseball leagues, is that when a hypothetical league has so few players, calculating POP (or RW-POP) inordinately changes the league average (by subtracting the individual hitter's data). Much of the effect that Rob ascribes to distribution is actually due to widely varying league averages for players in the same league, caused by the smallness of the league.

[5] This translates into a POP of 6.09. In this paper (but not in my two pervious papers) the hitting data of players who were primarily pitchers are excluded from the league batting average. Discrepancies in POP between this paper and 1999 BRJ paper are due to this exclusion. The slight differences in POP and RW-POP values in this paper compared to Rob's paper appear to be due to the use of different data base's pitchers' hitting data. For this paper, hits-at bats data were extracted from *www.baseball-reference.com*, July 2000. Also, in the calculations for this paper, RW-POP may have minor inaccuracies because I did not consolidate hitting records of any players traded within the league during the year.

[6] No one famous once said, "one man's intuition is another man's fallacy."

Hypothetically Expanded

Keeping with the spirit of baseball expansion, I have created 5 leagues, each with 100 players having 500 at bats each. These hypothetical leagues (Leagues 2-7) more realistically keep the league average relatively anchored for players within each league. (The column designatated with "#" shows the number of league hitters with these data)

Transforming one of the MM (League 3) players into AA (League 4) downgrades BB's POP from 2.56 to 2.50. But according to RW-POP, BB's performance seems to be excessively downgraded from 2.56 to 1.95.

League 3

player	#	H	AB	BA	LA	POP	RW-POP
BB	1	155	500	.310	.254	2.56	2.56
MM	99	127	500	.254	.255	0.28	0.28

League 4

player	#	H	AB	BA	LA	POP	RW-POP
AA	1	165	500	.330	.255	4.01	2.72
BB	1	155	500	.310	.255	2.50	1.95
MM	98	127	500	.254	.255	0.26	0.28

League 5

player	#	H	AB	BA	LA	POP	RW-POP
AA	1	165	500	.330	.255	4.01	2.12
BB	1	155	500	.310	.255	2.50	1.32
CC	19	145	500	.290	.255	1.38	0.81
DD	19	135	500	.270	.255	0.62	0.47
MM	22	127	500	.254	.255	0.26	0.29
PP	19	119	500	.238	.255	0.08	0.16
QQ	19	109	500	.218	.256	0.01	0.06

League 6

player	#	H	AB	BA	LA	POP	RW-POP
AA	1	165	500	.330	.258	3.70	2.46
BB	1	155	500	.310	.258	2.27	1.65
CC	5	145	500	.290	.258	1.22	1.06
DD	10	135	500	.270	.259	0.53	0.53
MM	83	127	500	.254	.259	0.21	0.24

League 7

player	#	H	AB	BA	LA	POP	RW-POP
AA	1	165	500	.330	.251	4.33	2.73
BB	2	155	500	.310	.251	2.75	1.97
MM	83	127	500	.254	.252	0.32	0.33
PP	10	119	500	.238	.252	0.11	0.12
QQ	5	109	500	.218	.252	0.01	0.02

Compare the RW-POP results of League 4 with League 5. It is not obvious why the RW-POPs for AA and BB are significantly lower in League 5 compared to League 4.

Now, compare League 6 (skewed positively) and League 7 (skewed negatively). POP clearly reflects, as does RW-POP, this difference in distribution.

In my opinion, the mean (average) is still the most useful one number to describe a population. In addition, changes in the distribution of the population will almost always change the mean. However, as Rob points out, POP uses only the mean (league average) to compare a player to the rest of the league. RW-POP is a good attempt to incorporate additional league data into the comparison. Let's look at some real results.

Reality Check

It will be even more instructive to go from the hypothetical to a real situation: determining the top 10 batting average performances in the 1930 National League. In calculating these rankings, both POP and RW-POP consider both total production (hits) and rate of production (batting average). Chart A ranks batting averages by RW-POP. Chart B ranks batting averages by POP.

Chart A shows that RW-POP lists the third best BA performance in the 1930 NL as George Puccinelli's 9 for 16 (with 3 home runs, 1.188 slugging average, and .563 batting average as a Cardinal outfielder in his first year). Likewise, Ray Blades (.396 in 101 at bats) is ranked seventh best and Showboat Fisher (.374 in 245 at bats) tenth best. Four of the season's top ten batters (Puccinelli, Blades, Fisher, and George Watkins) were outfielders for St. Louis in 1930, but all had fewer hits and at bats than two other Cardinal outfielders, Hafey and Douthit.

A closer look at Chart A (RW-POP) shows that Paul Waner is ranked just barely above Showboat, even though Waner had more than twice as many at bats as Showboat and a batting average only 6 points lower than Showboat.

It appears that RW-POP excessively rewards batting average at the expense of total hits and total at bats.

Perfect but not that Good

POP generally is resistant to ranking too highly performances that consist of good rates but relatively small amounts of success.

For example, consider rankings of the 20 best stolen base seasons (POP Chart C). Both Paul Moliter and Kevin McReynolds had perfect seasons, yet their SB POPs do not place their season performances into the top twenty. Stealing 51 bases in 53 attempts when others had a ,537 rate of success in the 1922 NL, Max Carey is the class of the field with an SB POP of 11.68.

Chart A – Top 10 BA RW-POP: 1930 NL (ranked by RW-POP)

Rank	Player	Hits	ABs	BA	POP	RW-POP
1	Bill Terry	254	633	.401	6.09	1.70
2	Babe Herman	241	614	.393	5.00	1.42
3	Geo. Puccinelli	9	16	.563	1.48	1.37
4	Chuck Klein	250	648	.386	4.54	1.26
5	Lefty O'Doul	202	528	.383	3.56	1.16
6	Fred Lindstrom	231	609	.379	3.72	1.12
7	Ray Blades	40	101	.396	1.35	0.98
8	George Watkins	146	391	.373	2.29	0.97
9	Paul Waner	217	589	.368	2.77	0.93
10	Showboat Fisher	95	254	.374	1.71	0.92

Comparing Apples and Grapefruit: Limitations of Relative Means Comparison

The traditional relativity tool is relative means comparison (RMC). This method was used by Shoebotham[7] when he compared hitters' batting averages from one league to another. RMC uses the ratio of a player's BA divided by his league's BA and compares different players' ratios to determine the best batting averages.

The first drawback of RMC is that it does not inherently incorporate the amount of success (number of hits) into the comparison. As a result, RMC makes no distinction between a hitter who hits .300 in 600 at bats and a hitter who hits .300 in 10 at bats. With this type of comparison, players receive no credit for the amount of their production. In addition, there is usually an arbitrary restriction of which players will be considered for comparison.

Chart B – Top 10 BA POP: 1930 NL (ranked by POP)

Rank	Player	Hits	ABs	BA	POP	RW-POP
1	Bill Terry	254	633	.401	6.09	1.70
2	Babe Herman	241	614	.393	5.00	1.42
3	Chuck Klein	250	648	.386	4.54	1.26
4	Fred Lindstrom	231	609	.379	3.72	1.12
5	Lefty O'Doul	202	528	.383	3.56	1.16
6	Paul Waner	217	589	.368	2.77	0.93
7	Pie Traynor	182	497	.366	2.30	0.89
8	George Watkins	146	391	.373	2.29	0.97
9	Kiki Cuyler	228	642	.355	2.02	0.76
10	Hack Wilson	208	585	.356	1.92	0.75

The second drawback of RMC was alluded to by Bob Costas during the 2000 All Star Game telecast. Costas commented on the difficulty of comparing players' performances from eras with markedly different league average performances--"apples" (e.g. players from the 1920 AL, where the league average HR rate was .00879) and "grapefruit" (e.g. players from the 1998 NL, where the league average HR rate was .0289). On the average, during the 1998 NL season, home runs were hit per at bat at more than triple the rate than during the 1920 AL season.

Using the RMC method, compare Babe Ruth's 1920 HR / league HR ratio (13.4) with Mark McGwire's ratio (4.8). Because of the widely different league HR averages between the two leagues, McGwire would have to have hit 208 home runs in 1998 to have matched Ruth's 1920 performance, by RMC standards. This is not reasonable.

Because of the markedly different home run environments (summarized by the leagues' HR averages), I think that there is no question that Ruth's 1920 HR achievement surpasses McGwire's 1998 HR achievement. Using POP, McGwire would have to have hit 93 home runs in 1998 to have matched Ruth's 1920 performance. POP provides a more reasonable result.

[7] David Shoebotham, "Relative Batting Averages", *Baseball Research Journal*, 1976, pp. 37-42.

RMC seems at first glance to provide reasonable comparisons when apples are compared to apples -- that is, when performances are compared with leagues with similar league averages. Using RMC to compare batting averages from different eras provides superficially reasonable results because the highest league batting average (1894 NL) is only 34% greater than the lowest (1968 AL). However, as the league differences become more pronounced (as with league home run averages), RMC results become more obviously unreasonable, as shown by the Ruth-McGwire HR comparison.

Linear vs. Curved: Carefully Stepping off the Straight and Narrow

As a relativity tool, POP introduces a new dimension (with considerations of both amount and rate, along with league average rate) and so POP results require us to think in curved terms rather than with Shoebothamian linearity.

To show that there is order in the universe described by POP, consider the following graphs.

Chart C – All-Time Best Season SB POP

Rank	Player	Year	Lg	SB POP	SB	CS	Ave	LgAv
1	Max Carey	1922	NL	11.68	51	2	.962	.527
2	Maury Wills	1962	NL	9.40	104	13	.889	.633
3	Vince Coleman	1986	NL	7.20	107	14	.884	.673
4	Willie Wilson	1979	AL	6.84	83	12	.874	.633
5	Willie Wilson	1980	AL	6.82	79	10	.888	.643
6	Max Carey	1923	NL	6.80	51	8	.864	.544
7	Bobby Bonds	1969	NL	6.58	45	4	.918	.587
8	Rickey Henderson	1983	AL	5.84	108	19	.850	.662
9	Rickey Henderson	1988	AL	5.78	93	13	.877	.677
10	Tim Raines	1983	NL	5.56	90	14	.865	.665
11	Lou Brock	1968	NL	5.41	62	12	.838	.589
12	Amos Otis	1970	AL	5.37	33	2	.943	.597
13	Lou Brock	1966	NL	5.27	74	18	.804	.582
14	Ron LeFlore	1979	AL	5.27	78	14	.848	.635
15	Jim Wynn	1965	NL	5.22	43	4	.915	.623
16	Eric Davis	1986	NL	5.20	80	11	.879	.675
17	Amos Otis	1971	AL	5.17	52	8	.867	.601
18	Willie Wilson	1984	AL	5.05	47	5	.904	.632
19	Mitchell Page	1977	AL	4.94	42	5	.894	.604
20	Lou Brock	1969	NL	4.93	63	14	.818	.585
	Paul Molitor	1994	AL	3.28	20	0	1.000	.686
	Kevin McReynolds	1988	NL	3.15	21	0	1.000	.708

Graph A shows the season BA POP (vertical y-axis) for hitting at least .400, as a function of league batting average (horizontal x-axis) and number of at bats (one of 7 curves, with the top curve representing 700 at bats and the bottom curve representing 100 at bats).

For example, Graph A shows that a player hitting .400 in 400 at bats with a .300 league average earns a BA POP of 5. Another player hitting .400 in the same league, but having 600 at bats, earns a BA POP of 7.

For those feeling insecure without linearity, Graph B should provide some solace. With increasing amounts of success, POP quickly approximates a linear function of the amount of success, when personal and league rates are held constant. For example, Graph B shows BA POP as a function of hits, with personal BA held constant at .400 and league BA held constant at .300.[8]

As Rob points out, Carl Yastrzemski's 1968 batting achievement (BA POP 3.36) is 600 times more probable than Bill Terry's 1930 achievement (BA POP 6.09). Mathematically this is so because of the relationship of the players' batting averages with their league averages and because Terry had nearly 100 more at bats. However, we need to be careful not to appeal to our linear intuitions when looking at POP. Although Terry's performance is shown to be significantly better than Yastrzemski's, POP does not allow us to conclude that he was 600 times better.

POP should be used when determining how well (above average) a player has performed. However, POP is not very helpful for assessing below average performances, so a corollary statistic (nPOP) can be used for poor performances.

The probability that a player (with an average league batting average of L) would get no more than H hits in A at bats is P_n, where

[8] Although not depicting a continuous function, Graph B has its plotted points joined by a line to improve visualization; labeled axis coordinates are for scale but may not designate actual data points.

$$P_n = \sum_{h=0}^{H} \frac{A!}{\left(h!(A-h)!\right)} L^h (1-L)^{A-h}$$

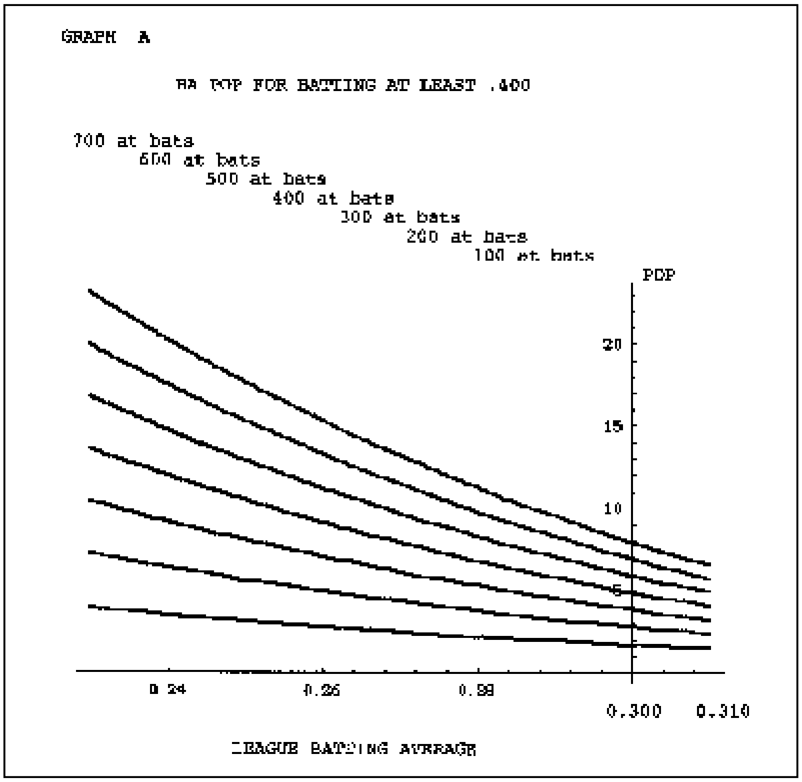
and nPOP = $\log_{10}P_n$

nPOP is a negative number. For example, Ozzie Smith's 1996 home run performance (2 HR in 227 at bats) is better described by nPOP (-1.38) than by POP (0.00). nPOP is a more sensitive measure of below average performance, and POP is a more sensitive measure of above average performance.

The term 'linear' has been used in this paper to denote straight line results and is not referring to linear weights. In fact, POP can be used to supplement linear weight evaluations.

Future POP: Slugging Average POP and Beyond

As Rob supposed, slugging averages can be compared by using POP. The multinomial formula for SA POP is as follows:

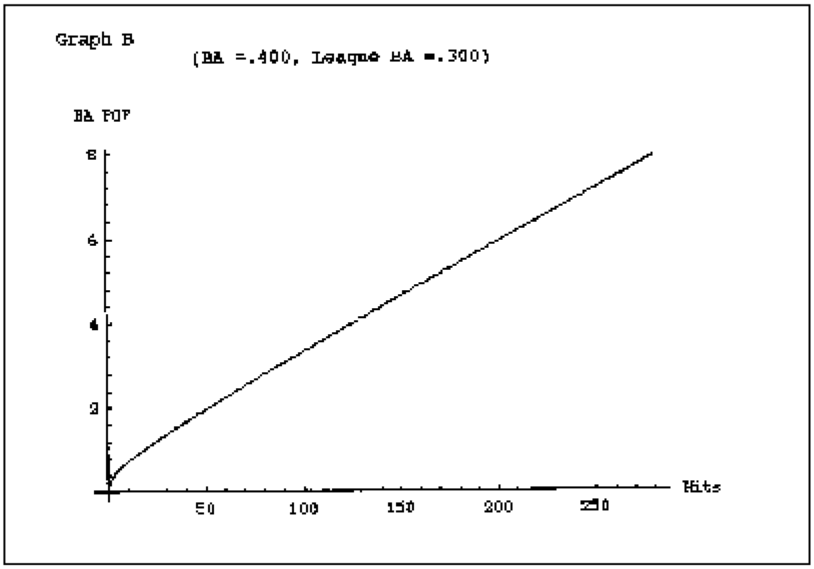


$$P_n = -\log_{10}\left(\sum_{h}\sum_{d}\sum_{t}\sum_{s}\sum_{u} \frac{A!}{h!t!d!s!u!} L_4{}^{h} L_3{}^{t} L_2{}^{d} L_1{}^{s} L_0{}^{u}\right)$$

where A = batter's actual at bats, H = batter's actual home runs, T = batter's actual triples, D = batter's actual doubles, S = batter's actual singles, = batter's actual outs, L_4 = average league rate of home runs, L_3 = average league rate of triples, L_2 = average league rate of doubles, L_1 = average league rate of singles, L_0 = average league rate of outs, and all league averages exclude the specific batter's data, and where u, s, d, t, and h are nonnegative integers that are used in every possible combination meeting the conditions

```
s+d+t+h+u = A and
(1s+2d+3t+4h+0u)/A >=
(1S+2D+3T+4H+0U)/A.
```

As for the POP comparison of Jeff Bagwell's 1994 .750 SA to Babe Ruth's 1920 .847 SA, it will be left as an exercise for the reader.[9]



[9] Sorry, but I always wanted to write that.

If the last equation uses weighting factors other than 0,1,2,3,4; if plate appearances are used instead of at bats; and if walks, HBP, sacrifices, and hit into double plays are added, then a batting linear weights POP can be calculated, using an expanded multinomial formula.

In my opinion, calculation of slugging average POPs or linear weight POPs would be a better use of supercomputers than modeling nuclear explosions or producing sophisticated but still inaccurate meteorological forecasts.

In Conclusion

Rob Wood's RW-POP is a commendable effort to incorporate more information into POP comparisons. His formula seems sound, but RW-POP results do not seem as reasonable as POP.

Compared to traditional relativity tools (relative means comparisons), POP has two advantages:

- First, POP considers both rate of production (batting average) and amount of production (hits), whereas more traditional tools typically consider either rate or amount.

- Second, when comparing performances from widely different eras, more traditional relativity tools produce less reasonable results compared to POP. POP allows us realistically to compare apples to grapefruit.

Mike Sluss is a neurologist in Green Bay, Wisconsin. Mike Sluss, 2847 Pioneer Dr. Green Bay, WI, 54313, mpsluss@aol.com. ♦

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	kcarlson@stlnet.com	Economics/Econometrics/Statistics
Rob Fabrizzio	rfabrizzio@bigfoot.com	Statistics
Larry Grasso	l.grasso@juno.com	Statistics
Tom Hanrahan	HanrahanTJ@navair.navy.mil	Statistics
Keith Karcher	kckarcher@compuserve.com	General
John Matthew	john.matthew@home.com	Apostrophes
Duke Rankin	RankinD@montevallo.edu	Statistics
John Stryker	johns@mcfeely.interaccess.com	General
Steve Wang	Steve.C.Wang@williams.edu	Statistics