By the Numbers

Volume 18, Number 4 The Newsletter of the SABR Statistical Analysis Committee November, 2008

Review

Review: Baseball Prospectus 2009

Koy Kosek

The author reviews the 2009 edition of the sabermetrically-oriented annual, Baseball Prospectus.

If you're like me, you hustle to the mailbox every evening in mid-February to see if your copy of this year's Baseball Prospectus has arrived yet and, on that lucky day when it finally does, you tear open the box and turn straight to the Milwaukee Brewers section. You are willing to overlook things like a mathematical table delineating of why it looks like your favorite team's talented quintet of star hitters has peaked to get to a heart-swelling summary of the team's greatest season in a quarter-century, and are even left somewhat in awe on occasion with lines like "Deadline deals end 26 years of wandering in the wilderness, as Dale Sveum plays Joshua to Ned Yost's Moses." Indeed.

You, of course, have only a slim probability of the Brewers being the first team you look at, but the phenomenon is nonetheless the same for statheads and, increasingly, for regular baseball fans across the nation. Baseball Prospectus has cut its way through the insanity of pre-sabermetrics preseason publications to emerge as a reliable, thorough set of commentaries on every team in major league baseball with some special nuggets interspersed throughout the book and at the end.

In this issue

It is those special nuggets that are the source of this review. They are generally meant to add insight into the game of baseball, independent of any specific team. This is the fourth year in a row that I have read the Baseball Prospectus preseason tome, and I read their website regularly, so this review ought to be taken as being from someone who has some familiarity with the publication. This is a rundown of what's inside this year's edition:

The foreword is kicked off by Keith Olbermann. For all of the whining that his inclusion in the publication inspired from some people initially, the write-up was really quite conventional. It's more about Keith Olbermann than the book (we learn, for instance, that Olbermann was a considered a math prodigy as a child, up until he had to do something other than add and subtract), but it is still an interesting, two-page, this-is-what-baseball-means-to-me-type essay that pays homage to Baseball Prospectus in general. It wasn't inspiring, but it does give the reader a bit of a sense of anticipation that Spring Training is about to begin, which probably makes it worth reading.

Similar to previous years, 589 pages are devoted to chapters of about 20 pages each for the 30 individual major league baseball teams. The back portion of the publication contains Kevin Goldstein's top 100 prospects, followed by Clay Davenport's park adjustment ratings for every minor and major league baseball stadium that he could get his hands on, and finally with leaderboards for last season's individual statistical standouts ushering us into the rather lengthy acknowledgements section. Mixed in with these lists in the back this year are four articles, which I will now briefly review one by one.

First up among the stand-alone articles is one authored by David Laurila called "Discovering America, South to North" which details the differences between Latin America and the United States. It focuses primarily on the adjustment that Latino players have to go through when coming to play in the United States. It is insightful, at least compared to what you'd read in a typical, 500-word internet post, in large part because it contains

extensive quotes from people who have been in the trenches, including Melvin Mora (Orioles third-baseman), Neil Huntington (Pirates General Manager), and Ed Romero, Jr. (Red Sox Latin Program Coordinator), and several others. Some of the story touches on cultural adjustments as well, including the prevalence of banned substances in Latin America – something which seems almost prescient given the revelations that have come out about Alex Rodriguez since the book went to press.

Following that article is "Something Old and Something New: Stadium Update" by Neil deMause, which details the stadium situation in major league baseball. The article is framed nicely by recaps of past and present stadium situations – remember in 1989, when the SkyDome was new and trendy? – and the meat of the article has some disturbing details of specific stadium deals, including those involving the Yankees, Mets, Marlins, Athletics, Rays, and Twins stadium deals. The article is interesting, but it does not go into enormous depth on any of the situations, choosing instead to go just deep enough to show the reader that portions of these deals are unsettling without naming many names or getting bogged down in the gritty details of any of them. It reads more like a well-researched general outline for an aspiring investigative journalist to pick a target out of and build a story around than it does like actual muckraking.

Next on the docket is Baseball Prospectus' ancient ancestor, Gary Huckabay, making a comeback with an intriguing essay called "An Unfair and Uneven Look at MLB's Marketing." The article essentially focuses on changes in American media in general as its primary topic for the first two-thirds or so of the essay. It gradually works major league baseball into the mix, and by the time it's over, major league baseball is being given rather high praise for how it has positioned itself over the last 25 years from a marketing standpoint. The article is complex and, while Mr. Huckabay references his experiences more than hard numbers, it is a thought-provoking piece.

The fourth and final stand-alone article on baseball analysis is sandwiched between tables listing the Park Factors and the PECOTA Leaderboards. Upon first flipping through the book, this seemed like an odd place to put the final article, but upon examining its content and author, it had to be thus: for the article, dubbed "The Times, They Are A-Changin'," is by Clay Davenport and is about – surprise! – the new WARP (Wins Above Replacement Player) equation for valuing players in 2009, which factors in defensive ratings and park adjustments. The math in this gets pretty textbookish, but it is a noble attempt to improve the way WARP is calculated, and it does make sense. To make a long story short, the type of batted ball and park-specific factors are now factored into defensive ratings, Clay Davenport is (with a healthy amount of grumbling) changing his definition of defensive replacement level to something that conforms to the present norms of the sabermetric community, and the equations for a player's WARP now factor in his defense based on objective data in a pretty credible way. Slowly, the puzzle of defensive value gets filled in.

What stands out about the articles in the back of the book this year is that only one of is based on divining truths about the actual game of baseball from mathematics and logic. The rest are well-written, but they are no different in form than what could be found in a feature article in Sports Illustrated or The Sporting News. As such, statheads may be disappointed with this year's stand-alone articles. There is no new ground broken here in a mathematical sense, except in Davenport's article, and even that is as much on the order of refinement of his previous system as it is about radical innovation, albeit high-quality, well-reasoned refinement.

On the other hand, what Baseball Prospectus (arguably) has that Sports Illustrated, The Sporting News, and their ilk (arguably) do not have is independence from the direct influence of major league baseball. As the more obvious elements of quantifiable baseball analysis become better understood by the public at large, it is worth a moment's reflection to appreciate the commitment to independence and truth that has gotten us this far. Thank you, Baseball Prospectus, for continuing to bring that truth to baseball fans.

Koy Kosek, koykosek@wrestlingadvantage.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
Shelly Appleton	slappleton@sbcglobal.net	Statistics
Ben Baumer	bbaumer@nymets.com	Statistics
Chris Beauchamp	cbeaucha@asinc.ca	Statistics
Jim Box	jim.box@duke.edu	Statistics
Keith Carlson	kcsqrd@charter.net	General
Matt Deutschman	mdeutschman@gmail.com	Proofreading
Dan Evans	devans@seattlemariners.com	General
Rob Fabrizzio	rfabrizzio@bigfoot.com	Statistics
Larry Grasso	l.grasso@juno.com	Statistics
Tom Hanrahan	Han60Man@aol.com	Statistics
John Heer	jheer@walterhav.com	Proofreading
Dan Heisman	danheisman@comcast.net	General
Christopher Jehn	cjehn@cox.net	Statistics
Bill Johnson	firebee02@hotmail.com	Statistics
Mark E. Johnson	maejohns@yahoo.com	General
David Kaplan	dkaplan@education.wisc.edu	Statistics (regression)
Keith Karcher	karcherk@earthlink.net	Statistics
Chris Leach	chrisleach@yahoo.com	General
Chris Long	clong@padres.com	Statistics
John Matthew IV	john.matthew@rogers.com	Apostrophes
Nicholas Miceli	nsmiceli@yahoo.com	Statistics
Barry Nelson	uawmftdwmf@aol.com	Proofreading
John Stryker	john.stryker@gmail.com	General
Tom Thress	TomThress@aol.com	Statistics (regression)
Joel Tscherne	Joel@tscherne.org	General
Dick Unruh	runruhjr@iw.net	Proofreading
Steve Wang	scwang@fas.harvard.edu	Statistics

How Leadoff Hitters are Sabermetrically Overrated

Tom Hanrahan

Do run estimators accurately evaluate all hitters, regardless of where they hit in the batting order? Here, the author argues that because they hit with so few men on base, leadoff hitters are systematically overrated by the traditional sabermetric offensive statistics.

Run Estimators. Those tools we use to turn individual batting events into the unit that is most useful for determining how baseball teams win games: runs. Linear Weights, Base Runs, Runs Created (LW, BsR, RC) – they assign values to singles, walks, home runs, outs, etc., either individually or in combination, to estimate how many runs a team will score. All of the above tools (as well as others not mentioned) have been shown to be accurate for what they are used for, on a team level. Much effort has been made to come up with better run estimators; tweak a coefficient here, adjust something there, looking for a small edge in accuracy. The sabermetric community has gotten to the point where we mostly agree that not much more will be gained by new formulae; we've reached a point of diminishing returns, and we can assess the particular strengths and weaknesses of each estimator pretty well.

It is assumed by almost all of us that because LW, BsR, and RC are accurate for estimating team run totals, that they also are good measures of individual players' contributions in terms of runs. I agree with the above assumption, with one big caveat: Leadoff hitters are overrated, by about 10%, by all of the above run estimators. Provocative statement? Yes, it is. The rest of the article explains and defends this assertion.

Run estimators measure quality and quantity; as they should. An OPS of .800 will always generate more runs than an OPS of .700, given the same amount of playing time. And given two hitters with great stats, the one who comes to the plate more often has created more runs for his team.

While MLB managers are responsible for determining batting order, it's obvious that most choose their batters based on fairly standard characteristics; high on-base guys at the top, power in the middle, with the poorer hitters lower in the order. So, while the 6th place hitter will have an advantage of maybe 30 plate appearances (PA) per season over the 8th place hitter, it is reasonable to give credit for this to the 6th place hitter; he was judged to be good enough to hit there, rather than down at the bottom.

The best hitters typically hit in spots 1 through 4, or sometimes 5th. The leadoff batter will, in the same number of games, come to the plate about 7% to 8% more frequently than the cleanup hitter[1]. So, if the batters had similar results per PA, the leadoff hitter will by any run estimator have 8% more runs to his credit than the typical hitter. This would be fine, if these extra PA actually added 8% more runs to his team. But they don't.

This really is not very hard to calculate; and in fact, it's easy to visualize. Everyone knows that the reason good hitters with power bat 3rd and 4th is because they come to the plate more frequently with runners on base; and so, their hits are more important (and their outs more costly!). The leadoff batter, on the other hand, comes up with fewer runners on than every other lineup spot -- primarily because he leads off the game with no one on, but also because the 7th thru 9th batters who precede him typically do not reach base as often.[1]

In BTN's Feb 2001 issue, I showed that using data from the 1999 season provided by Tom Ruane, the typical leadoff batter came to the plate with an average of .50 runners on.

Full data for each batting order spot is in Table 1 on the next page. There is a fairly smooth increase and decline from spots 2 to 9, but the paucity of runners on for leadoff men sticks out like a sore thumb.

That article also showed that the relative importance of each plate appearances varied almost perfectly linearly with the number of runners on, plus the batter: hits and outs are 4 times more crucial to run scoring when the bases are loaded (batter plus 3) than when they are empty (batter only).

[1] If a team averages 39.5 men batting per game, that means the cleanup hitter will come to the plate an average of 4.5 times, as the 5th time through the order stops at his spot. Each hitter preceding him will come up 1/9th of one time more often per batting spot. The leadoff man will average 4.5+3/9ths or 4.83 times up per game, which is 7.4% more than the cleanup hitter's average of 4.5.

Putting these two together and it is seen that the cleanup man typically bats for 1.74 men per PA, while the leadoff hitter bats for 1.50; which is 14% less leverage! The first batter may get more PA per season, but do those extra PA really generate more runs for the team?

Of course, the above analysis could miss some key points. Maybe the leadoff batter's initial trip in the first inning is more crucial, as there are always no outs. Maybe because he bats in front of other good hitters, his times reached base are more important. In order to assess these and other possible errors in the above analysis, I ran some lineup simulations to see how much difference there is in improving the leadoff spot in the lineup, versus other batting order positions.

I used the free (albeit somewhat old) program Star v1.2, a lineup simulator, to generate oodles of games with different lineups. It may not be the most complex product on the market, but it seems to do a reasonable job. And since I was not testing nuances like the baserunning, bunting, or the relative value of walks versus doubles, I think the results are very solid.

Table 1 – Runners on base per PA by lineup spot

Lineup spot	1	2	3	4	5	6	7	8	9
Avg runners on base	.50	.62	.67	.74	.72	.70	.69	.68	.69

Step 1: I created a typical lineup, and ran one million games with it. Results, in average per 162 games:

Position	AVG	OBP	SLG	AB	H	2B	3B	HR	BB	R	RBI
leadoff	.283	.375	.392	649	184	35	4	9	94	104	54
#2 man	.283	.373	.435	632	179	34	4	18	91	102	74
3rd guy	.276	.366	.472	615	170	33	4	27	89	98	101
cleanup	.274	.365	.513	600	165	32	4	35	87	98	118
5 hole	.263	.330	.451	613	161	31	4	25	61	82	97
sixth	.257	.318	.414	604	155	31	4	19	53	72	80
seventh	.253	.315	.387	588	149	30	3	14	52	65	68
eighth	.241	.298	.356	577	139	29	3	10	45	60	60
last	.208	.260	.298	565	117	28	3	5	38	56	48
TOTALS	.261	.336	.414	5443	1419	283	33	163	610	737	699

Runs per season: 737.1

Step 2: I made the leadoff batter much better. Per 550 PA, I gave him 20 more walks, 20 more singles, 20 more home runs, and 60 fewer outs. And ran another million games.

Position	AVG	OBP	SLG	AB	H	2B	3B	HR	BB	R	RBI
leadoff	.381	.481	.625	631	240	35	4	37	122	149	98
#2 man	.282	.372	.435	642	181	34	4	19	93	104	80
3rd guy	.276	.365	.472	624	172	33	4	27	90	100	109
cleanup	.275	.365	.514	609	167	33	4	35	88	100	124
5 hole	.263	.330	.451	622	164	32	4	26	62	84	101
sixth	.257	.318	.414	613	157	31	4	19	54	73	81
seventh	.253	.315	.386	597	151	30	4	14	53	69	69
eighth	.241	.297	.355	587	141	30	3	10	46	67	60
last	.208	.260	.299	575	120	29	3	6	38	66	49
TOTALS	.272	.348	.441	5500	1495	288	33	193	646	811	771

Runs per season: 811.0

So, replacing the typical leadoff batter with an awesome one added 73.9 runs per season to the team's totals. You can see the totals for the other batters remained basically the same, except they each got to bat more often, since the #1 man in the lineup was on base much more often.

Step 3: I did the same for other batting spots; the #3, #4, #6, and #9 hitters, always increasing their rate stats by the same as in the above example. I realize it is unrealistic to consider a MLB manager hitting a man with an OPS of .876 in the 9th spot all year, but this does show how the lineup would work if such a plan were implemented.

Team totals for each of these lineups, as compared to the ones already given above:

Lineup	AVG	OBP	SLG	AB	H	2B	3B	HR	BB	R	RBI
Original	.261	.336	.414	5443	1419	283	33	163	610	737.1	699
Great #1	.272	.348	.441	5500	1495	288	33	193	646	811.0	771
Great #3	.271	.348	.439	5507	1492	288	33	191	644	814.9	775
Great #4	.271	.347	.439	5502	1490	288	33	190	643	813.9	774
Great #6	.270	.347	.438	5498	1487	288	33	189	642	806.0	766
Great #9	.270	.346	.436	5495	1482	287	33	187	639	800.5	761

Observations:

1. Team totals get slightly worse as you go down the rows; because the earlier in the lineup you put the best hitter, the more he gets up.
2. No surprise that putting your best hitter in the 9th spot generates fewer runs.
3. Putting a great hitter in spots 1, 3, and 4 generates about the same results; so doesn't this make us conclude that, like studies before have shown, that lineup construction is no big deal? Answer: Well, yes, but the point to this article is *not* whether lineup construction is important.

Here is the full result of the great #4 hitter in the lineup:

Position	AVG	OBP	SLG	AB	H	2B	3B	HR	BB	R	RBI
leadoff	.282	.374	.391	659	186	35	4	9	95	112	55
#2 man	.283	.373	.435	641	181	34	4	19	92	113	75
3rd guy	.276	.366	.472	624	172	33	4	27	90	112	102
cleanup	.372	.470	.751	584	217	33	4	60	113	137	169
5 hole	.264	.330	.452	622	164	32	4	26	62	83	105
sixth	.257	.318	.413	613	157	31	4	19	54	73	86
seventh	.253	.315	.386	597	151	30	4	14	53	66	72
eighth	.240	.297	.355	587	141	30	3	10	45	61	61
last	.208	.260	.298	575	120	29	3	6	39	58	49
TOTALS	.271	.347	.439	5502	1490	288	33	190	643	814	774

If I were to assign values to each individual lineup spot by any of the commonly used run estimators, it would show that the difference between the great leadoff and the typical one is *larger* than the differences between the great and typical #3, #4, #6, and #9 hitters. Why? Because the leadoff batter had more PA. In the 'great leadoff hitter' results, we see he averaged 753 PA per season. In the recent table above, the great #4 hitter averaged 697 PA. It is absolutely true that the great leadoff hitter came to the plate more often. It is just as true that, *per plate appearance,* the leadoff hitter's performance was worth less. Much less.

Table 2, on the next page, takes the team run net change in scoring results gained by making each lineup spot "great", and combines them with how often each batter came up, to assign a team runs per PA value to each of the great hitters.

Improving the middle-of-the-lineup hitters (3 and 4) had a slightly larger overall impact on how many runs the team scored; even though they came up to the plate less often. The cleanup hitter's impact on the lineup was 12 percent higher (= 12 percent more "leverage") per PA (.110/.098) than that of the leadoff hitter. While this number only came from a simulation using certain parameters, it is line with the theoretical results (14% more leverage) I had determined previously using simply the average number of base runners for each spot.

Practically, what does this mean?

1. When evaluating hitters' contributions to their teams, the hitter in the leadoff spot is worth about 10% less per PA than most other lineup spots. Thus, their value is 10% lower than the results you get from most every Run Estimator. There are some other smaller corrections that maybe could be made (#2 spot a little lower, cleanup batter a little higher), but those are small by comparison.

2 Rating great players: Let's take the Hall of Fame cases of Tim Raines and Andre Dawson. Contemporaries, teammates, similar career length, both fine peaks and slow declines. They even played virtually the same position and both played it fairly well. Really, these should

be two of the easiest players to compare. Both were on the Hall ballot in 2009. Raines was considered Hall-worthy by 23% of the BBWAA, while Dawson got 67% of their vote.

The next two tables show their offensive career totals; Table 3 shows the traditional ones, and Table 4 shows their careers as seen by some of the more popular batting metrics. For Table 4, I dropped off a few of the years at the tails of each man's career that had very little value.

Baseball-reference.com's Batting Runs is a measure above average. It does not take into account stolen bases, which should give Raines another 120 run advantage or so. Win Shares (WS) is Bill James' creation, and I multiplied them by 3.33 here to turn them into Runs, the same unit as the other metrics. WS measures value above a low replacement level. Baseball Prospectus gives Batting Runs above average (BRAA) and above Replacement (BRAR).

Table 2 – Value in Runs Per PA by Lineup Spot

Lineup Spot Improved	Extra Team Runs Scored	PA for lineup spot	Extra Runs per PA
1	73.9	753	.098
3	77.8	714	.109
4	76.8	697	.110
6	68.9	666	.103
9	63.4	612	.104

All of these metrics show Raines was clearly the better hitter over his career. But Raines batted in the leadoff spot for about 63% of his career (other time was spent as a #2 and #3 hitter, with a few appearances elsewhere), while Dawson was almost always a middle-of-the-order man. Raines' extra PA gained by batting leadoff led to additional runs, as measured by each of these metrics, but his true value above average I believe ought to be downgraded by about 7% (since he spent almost 2/3rds of his career batting first). I still personally feel he has a stronger Hall of Fame case than the Hawk, but I also think it is closer than some of my statistically-oriented brethren see it; and maybe this is what many voters implicitly sense?

Table 3 – Tim Raines vs. Andre Dawson – Traditional statistics career comparison

Player	BB	AB	AVG	HR	SB	R	RBI
Raines	1330	8872	.294	170	808	1571	980
Dawson	589	9927	.279	438	314	1373	1591

Similarly, if one were to try to rank the newest Hall of Famer, Rickey Henderson, against many of the other all-time great corner outfielders, in many peoples' assessment he comes in below the obvious superstars Williams/Bonds/Musial/Ruth/Aaron, but possibly equal to or better than such luminaries as Frank Robinson and Mel Ott. In a recent (January 12) article on Baseballprospectus.com, Will Carroll and Nate Silver opined that "Rickey Henderson is one of the 20 greatest players in baseball history". Because Rickey spent his whole career as a leadoff hitter, I think the current set of runs estimators (likely what Mr. Carroll and Mr. Silver used to formulate their opinion) overstate his value, and I believe this pushes him below Mr. Robinson and Mr. Ott in terms of wins contributed to his baseball teams.

Table 4 – Tim Raines vs. Andre Dawson – Run Estimator offense comparison

	Batting Runs, bb-ref.com	Win Shares * 3.33 (offense only)	BRAA/BRAR, Baseball Prospectus
Raines, 1981-98	337.3	1105	605 / 891
Dawson, 1977-92	241.8	931	318 / 651

In summary, all Run Estimators overstate leadoff batters contributions to team runs scored. By about 10%. That ain't small potatoes when figuring out what makes teams win.

Tom Hanrahan, Han60Man@aol.com ♦

Submissions

Phil Birnbaum, Editor

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

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

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

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

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

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

Get Your Own Copy

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

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

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

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

Great Streaks

Jim Albert

An article in this year's Baseball Research Journal *argued that hitting streaks are achieved more frequently than if there were no "hot hand" effect. Here, the author acknowledges that finding, but argues that the effect is so small that it can be ignored for practical purposes. In addition, he uses the original researcher's technique to identify the most seemingly-unlikely (although not necessarily longest) streaks of the past several baseball seasons.*

1. Introduction

Recently I wrote an article for *Quantitative Analysis of Sports* where I looked careful at the streaky patterns of hitters during the 2005 season. After I wrote this, I vowed never again to write again about streakiness. But after reading the recent article by Trent McCotter in the *Baseball Research Journal*, Volume 37 (2008), I had to break my vow. McCotter's interesting look at streaky hitting and the statements made in the article deserve some explanation and comments. Also, he describes an attractive method for assessing streakiness and it is straightforward to apply his statistical approach to identify extreme hitting streaks in recent seasons. Using this methodology, I find some "great streaks" during the 2004 through 2008 baseball seasons.

2. Comments on BRJ article by McCotter

In the BRJ article, McCotter wishes to construct a test of the common hypothesis that the individual batting outcomes of a particular player during a season represent independent, identically distributed trials. (We'll call this the "IID assumption" or the "IID model".) Essentially this hypothesis says that the batting outcomes are similar to flips of a coin where the chance of a hit on a single at-bat is equal to the batter's "true" batting average.

To test this hypothesis, the author looks at the pattern of game hitting streaks of all players for the seasons 1957 through 2006. Suppose we collect the game-to-game hitting records of Mickey Mantle during the 1961 season. We record all his hitting streaks of Mantle for this season – maybe he started with a hitting streak of one game, a second hitting streak of three games, a third hitting streak of four games, and so on. If the IID assumption is true, then the pattern of hitting streaks for Mantle is simply a byproduct of chance variation. If we randomly rearranged his game-to-game hitting statistics, then that wouldn't change the pattern of streaks. The general question is whether Mickey Mantle's observed pattern of hitting streaks (and the streak patterns for other players) is consistent with the random patterns from a model with the IID assumption.

The author performs a computationally-intensive simulation of the patterns of hitting streaks for all players and seasons for the years 1957-2006. For each player's game log for a season (for all 50 seasons), he randomly arranges the batting lines. Then he computes the lengths of all hitting streaks for all players for all seasons. Then he repeats this simulation process for a total of 10,000 iterations. When he is done, one has an empirical distribution of the lengths of hitting streaks under the IID assumption, and one can see if the actual lengths of streaks are consistent with this distribution.

The conclusions of the paper can be summarized by two tables that show that the actual number of long hitting streaks (of length 5 and greater) are consistently larger than the mean number of long streaks predicted from the IID model. Moreover, the differences are highly statistically significant. The author concludes by saying:

> "This study seems to provide some strong evidence that players' games are not independent identically distributed trials, as statisticians have assumed all these years, and it may even provide evidence that things like hot hands are a part of baseball streaks. … From the overwhelming evidence of the permutations, it appears that, when the same math formulas used for coin tosses are used for hitting streaks, the probabilities they yield are incorrect." Much of the article is devoted to a discussion of this conclusion, giving some possible explanations for the presence of long streaks.

Generally I'm fine with the statistical methodology used in the paper. As I'll illustrate later, the permutation test procedure is an attractive method for testing the IID assumption and the results described in the article are interesting. But I am concerned about some of the author's statements and conclusions about this analysis.

First, the author seems to make the implicit assumption that all statisticians believe in the IID assumption. The IID assumption is an example of a statistical model that we may use to fit baseball data. Any model we apply is actually wrong – that is, the real process behaves in a much more sophisticated manner than the model suggests. For example, take the standard IID assumption that individual at-bats are coin-tossing outcomes with a constant probability of hitting success p. Do I believe this is true? Of course not. I believe that the hitting talent of a player goes through many changes during the season and it depends on many other variables such as the quality of the pitcher, the game situation, whether the game is at home, etc. So if the IID model is wrong, why do we use it? Well, the IID model has been shown to be useful in understanding the variation of baseball data. One thing that I have found remarkable in my baseball research is that good simple models (like the IID model) are really helpful in predicting future baseball outcomes.

Second, the author gives the impression that this statistical analysis gives evidence for the hot hand effect in baseball. Suppose you reject the IID assumption – what does this mean? It could mean that there is a dependence structure in the batting sequence. That is, one's performance in one at-bat is helpful in predicting the performance in successive at-bats. But there is a second possible explanation. Maybe the outcomes are independent, but the chance of getting a hit changes across the season. Either explanation, a dependence pattern or a change in hitting probability, would explain the presence of long streaks. Also these two characteristics are confounded and it is difficult statistically to isolate their effects. So it is wrong to say that long streaks imply a dependence pattern in the hitting sequences. People love to believe in the hot hand and I'm concerned that this paper adds fuel to their hot-hand belief.

Last, what is the practical significance of the results? To find these streaky effects, the author had to consider all hitting sequences in 50 seasons of baseball data. This is a ton of data – the author was likely considering streaks present in over 30,000 player seasons! But we live in the context of a single season and these results really don't say that the IID assumption is inappropriate for understanding the lengths of hitting streaks for a single season. I suspect that in the context of a single season, these streaky effects are relatively small and can safely be ignored. The author is concerned about the difficulty of devising a more accurate modeling method since he has shown that the IID assumption is incorrect. But that's okay, since statisticians don't need "exact" models. If the streak effect is "real" but small in size, then I'll continue to use the IID model since I believe it is an attractive approximate model that works.

3. Is there evidence of long streaks for the last five seasons?

After reading this paper, it seemed natural to explore the presence of long streaks in the context of a single season. McCotter demonstrated that there was a streakiness effect, but didn't measure the size of this effect. If the streakiness effect was substantial, then I would think it should manifest itself in a single season.

So I replicated the author's analysis for each of the five recent baseball seasons from 2004 to 2008. I'll carefully outline what I did for the 2004 season which may help explain the author's method in the BRJ article.

1. Using play-by-play files from Retrosheet (www.retrosheet.org), I collected the game-to-game hitting data (number of hits and number of at-bats) for all 959 players who had at least one official at-bat in the 2004 season.

2. For each player's game log, I collected the lengths of all hitting streaks. For example, for the 2004 John McDonald, I record if he got a hit (Y) or not (N) for each of the 40 games he had an official at-bat in the season.

```
Game    0000000001111111111222222222233333333334
Number  1234567890123456789012345678901234567890
Hit?    NNNNYNNNYNNNNNYNYYYNNNNNNNNYYYYYYYYNNNNNNN
Streak      1   1     1 12        1234567
```

I collect the hitting streak lengths 1, 1, 1, 2, 7. Likewise, I collect the streak lengths for all other 958 players.

3. Next, I wish to simulate batting logs for all players under the assumption that the game order in each player's batting log is recorded is not important. For each player's batting log, I randomly permute the Y's and N's. For the simulated batting log, I again collect all of the streak lengths for all players. When I am done, I collect the number of streaks of length 1, the number of streaks of length 2, and so on.

4. I repeat this simulation method in part 3 one thousand times, obtaining 1000 sets of streak lengths.

5. Last, I compare the distribution of simulated streak lengths with the actual streak lengths observed in the 2004 season. A sample of results is displayed in the following table. Suppose we are interested in the number of "long" streaks that are five or longer. Under the "Actual" column, we see that we observed 1707 streaks of length 5 or higher in the 2004 season. In the simulation, the mean number of

streaks that were 5 or higher was 1690 and the standard deviation of the number of 5+ streaks was 24.2 – these numbers are placed in the "Mean" and "Stand Dev" columns. We notice that we observed more streaks than one would expect under the IID model. Is this significant? To answer this question, we compute the p-value which is the probability that the simulated number of 5+ streaks is at least as large as the observed number of 5+ streaks. If the p-value is small (say, under 0.05), then we reject the IID model. Here we compute that the p-value is 0.25 – the conclusion is that we have insufficient evidence to say that the data rejects the IID hypothesis. (By the way, the BRJ article didn't contain p-values and I think the inclusion of those numbers would help the exposition.)

The above procedure was repeated for each of the five seasons and the results are displayed in the following five tables. In each table, we look at the number of streaks of length 5 or more, the number of streaks of length 10 or more, the number of length 15 or more, and the number of length 20 or more. The p-values indicate the consistency of the strength lengths with the IID model – small p-values indicate that the observed strength lengths are longer than one would expect under the IID model.

2004 Season

Streak Length	Actual	Mean	Stand Dev	P-value
5 or more	1707	1690	24.2	0.25
10 or more	235	227.4	12	0.28
15 or more	35	35.8	5.6	0.58
20 or more	7	6.2	2.4	0.42

2005 Season

Streak Length	Actual	Mean	Stand Dev	P-value
5 or more	1707	1665.3	23.7	0.04
10 or more	228	214.5	11.6	0.14
15 or more	29	32.7	5.2	0.79
20 or more	9	5.8	2.3	0.13

2006 Season

Streak Length	Actual	Mean	Stand Dev	P-value
5 or more	1729	1689.8	24.4	0.07
10 or more	231	227.6	11.9	0.40
15 or more	34	35.6	5.4	0.65
20 or more	9	6.2	2.3	0.16

2007 Season

Streak Length	Actual	Mean	Stand Dev	P-value
5 or more	1712	1691.6	23.9	0.20
10 or more	238	226.3	11.8	0.16
15 or more	46	35.4	5.5	0.04
20 or more	11	6.2	2.4	0.04

2008 Season

Streak Length	Actual	Mean	Stand Dev	P-value
5 or more	1663	1688	24	0.87
10 or more	236	227.1	12	0.24
15 or more	38	35.5	5.4	0.35
20 or more	4	6.2	2.3	0.87

What do we learn from this analysis? The p-values for the 2004 and 2008 seasons are large, indicating that for these seasons the streaks were consistent with the IID model. In contrast, the 2006 and 2007 p-values are small, suggesting that the streakiness is significant for these seasons, and the 2005 p-values are less conclusive. From this brief analysis, the IID model appears useful in explaining the variation in strength lengths for some seasons. The size of the streakiness effect is small enough that it is not detectable statistically for particular seasons. McCotter did find significant streakiness in his study of 50 seasons of data, but the practical significance of his result is questionable by this analysis.

4. Using a permutation test to identify great streaks

In baseball, we simply define a long streak by the consecutive number of official games in which a player gets at least one base hit. The website http://www.baseball-reference.com/bullpen/Longest_Hitting_Streaks lists all of the hitting streaks in baseball history of length 30 or greater. In my QAS article, I explain that the length of a hitting strength is confounded with two variables. Better hitting players are more likely to have long streaks since they are more likely to get a hit in a game. Also, regular players who play all the games in a season are more likely to have long streaks than utility players who have fewer opportunities to hit. It is desirable to get a measure of streakiness that is not related to hitting success or number of games played.

The permutation test described in the BRJ article provides a simple method of assessing the size of a particular hitting streak that adjusts for player ability and number of games played. We illustrate the calculation using John McDonald's data for the 2004 season.

We show again his game data. We see that his longest hitting streak was 7 games. Was this a noteworthy streak? (On the surface, you probably would say no, since 7 doesn't sound very large.)

```
Game    0000000001111111111222222222233333333334
Number  1234567890123456789012345678901234567890
Hit?    NNNNYNNNYNNNNNYNYYNNNNNNNNYYYYYYYNNNNNNN
Longest hitting streak = 7 games
```

As in the previous analysis, we simulate hitting sequences assuming the IID model. For each of the ten lines below, we randomly arrange the sequence of 12 Y's (games with a hit) and 28 N's (games with no hit), and then compute the length of the longest hitting sequence in each of the random permutations.

Simulation Number	Sequence	Length of longest hitting streak
1	NNNNNNNYNNNYNNYYNNNNYNNNNYYYNYNNYNNNNNYY	3
2	NYNYNNYNNNYNNNYNNNNYNYNNYYYNYNNNNNNNNNNY	3
3	NNNNYYNNNNNNNNNNYYNYNNYNYYNNYNNNNNNYYNYNN	2
4	NYYNNNYNNNNYNNNNYYNYNNNNNNNYNYNNNNYYNNNY	2
5	NNNNYNNNNYYNNYYYNNNNNNNNNNNNYYNYYNYYNNNNN	3
6	NNNNNNYYNNYNYNNNNNYNYNYNNNNYNNNYYNNNNYYN	2
7	NYYYNNYNNNYNNNYNNYNYNNNNYNNNYNNNYYNNNNNN	3
8	NNNYYYNYNYNNNYYNNYNYNNNNNNNYNYNNNNNNYNNNN	3
9	NNYYYYNNYNNNNNYNNNNNYNYNNNNYYNNNNNNNNYYN	4
10	YNNNNYNNYNNNNNNNYNYYNNNNNNNYNNNNNYYYNYYN	3

If we repeat this exercise for 10,000 simulations, we obtain the empirical distribution of the longest hitting streak for McDonald if the game results were truly a random sequence. To see if McDonald's streak of 7 is extreme, we compute the p-value, the probability that the longest streak length in the random sequence is 7 or higher. We see from the output below that the p-value is 0.0017, a pretty small number. We conclude that McDonald's hitting streak of 7 is pretty impressive since the chance of getting a streak this large by chance is so small.

```
Streak Length     1     2     3     4    5    6    7    8    9
-------------------------------------------------------------
        Count    78  4344  4010  1197  291   63   15    1    1

p-value = P(Random Streak Length >= Observed Length)
        = P(Random Streak Length >= 7) = (15+1+1)/10000 = 0.0017

-log10(p-value) = 2.77
```

I used this procedure to assess the greatness of the longest streak of hits for every player in the five seasons 2005 through 2008. To pick out a relatively small number of streaks, I arbitrarily decide that a streak is "great" if the p-value is smaller than 0.0032 (that is, if -log10(p-value) exceeds 2.5).

The following table displays 13 great streaks in this five season period that satisfy this criterion. There are some obvious great streaks listed such as Jimmy Rollins' streak of 36 games in 2005, Chase Utley's streak of 35 in 2006, and Willy Taveras' streak of 30 in 2006. But there are several surprising names on this list including John McDonald, Mike Napoli and So Taguchi. But remember that this streaky measure

automatically adjusts for the hitting ability and number of games of the player. This measure essentially lists the most surprising hitting sequences as identified by the permutation test.

5. Closing comments

Have we learned anything new about streakiness in baseball? McCotter proposes an interesting method of detecting streakiness using a large dataset (50 years of baseball) and he did show that ``true'' streakiness existed. But I believe his conclusions are similar in spirit to the conclusions in my JQAS paper. We see much streaky behavior in baseball data, but most of the variability in this behavior of it can be explained using simple probability models such as the IID model here. Although simple models explain most of the behavior, I concluded in my article that some players exhibited more streakiness than the models would predict. Moreover, it seems hard to find statistical evidence for players who are consistently streakiness across seasons.

Table 1 – Hitting streaks in the seasons 2004-2008 where -log10(p-values)>2.5

Season	Player	-log10(p-value)	Length of streak
2004	Robb Quinlin	3.1	21
	John McDonald	3.05	7
	Ross Gload	2.72	16
	Carlos Lee	2.72	28
2005	Jimmy Rollins	> 4.0	36
	Maicer Izturis	2.64	13
2006	Chase Utley	4	35
	Willy Taveras	3.4	30
	Chris Gomez	3.1	18
	Manny Ramirez	2.89	27
2007	Javier Valentin	2.82	14
	Mike Napoli	2.66	14
	So Taguchi	2.54	18

One interesting byproduct of this work was the use of the permutation test to identify unusually long hitting streaks. By looking all at players instead of the regulars, one can identify players such as John McDonald, who exhibit strong streaky performances despite hitting for a poor average.

References

Albert, J. (2008), "Streaky Hitting in Baseball", *Journal of Quantitative Analysis of Sports,* http://www.bepress.com/jqas/vol4/iss1/3

McCotter, T. (2008), "Hitting Streaks Don't Obey Your Rules"*, Baseball Research Journal.*

Jim Albert, albertcb1@gmail.com ♦