By the Numbers

Volume 9, Number 4 The Newsletter of the SABR Statistical Analysis Committee November, 1999

Welcome

Phil Birnbaum, Editor

Well, it's been a year now since BTN came back, and, while five issues is hardly enough for a full-blown retrospective, maybe I can at least get away with a brief synopsis of what we've accomplished in the past 12 months.

A quick count shows that we've published 34 articles (not including monthly comments from Neal and from me). We've had a few reviews, a few criticisms, and a whole bunch of original research studies, all of which, I think, are of quite good quality.

It has been said that the public perception of Sabermetrics is of a bunch of nerds poring over numbers with slide rule in hand, and this perception rankles many of our members. I think it's therefore significant that among the studies we've published are a few that answer questions that typical baseball fans might ponder at the watercooler or at the ballpark.

In this issue

Take, for instance, the issue of whether catchers learn to handle pitchers better as they gain experience. This is a question that has always intrigued me, and I've had discussions with non-statistically-oriented fans who also find the question interesting. Well, last issue, Tom Hanrahan published an excellent study that put forth convincing evidence to provide an answer to that question.

It's also often mentioned that faster runners benefit their teams by causing more opposition errors, a benefit that doesn't appear in their official stats. Is this true? Well, a year ago, Dan Levitt amassed a full two year's worth of play-by-play data to show that it is, indeed, true, and to what extent.

And, in this BTN, a study by John Jarvis answers another oft-discussed question: how much was Mark McGwire pitched around to in his record-breaking year? Is his achievement made even more noteworthy by the fact that he didn't see as many good pitches to hit? Mr. Jarvis marshalls the evidence to reach a conclusion that any casual fan can understand.

I don't mean to say that studies like these are better or more important than studies that are less casual-fan oriented. Indeed, I could point out several articles that we've published that, although they require a bit of sabermetric background to appreciate, have taught me more about the game than these other studies. Rob Wood's article on what drives MVP voting, for instance, is less likely to appeal to the casual fan, because of its mathematical content and because the answer is difficult to simplify to a sentence or two. But I nonetheless like Rob's article very much, as it answers a question I've always wondered about in a way that's probably as mathematically simple as it can get.

And so I certainly don't mean to say that some types of study are more important or more welcome in BTN than others. The only criteria for inclusion in this newsletter are these: does the study teach us something new about baseball, and do the conclusions follow logically from the evidence? If all we ever published were studies aimed at only the casual fan, the field of Sabermetrics would never move forward.

But from the limited standpoint of gaining more acceptance from the public – both within SABR and beyond – for what we do, this kind of study is extremely valuable, and I would encourage the authors of these excellent studies to consider reworking them for SABR's *Baseball Research Journal*, where they would find a wider audience and enlighten another category of researcher and fan.

You can e-mail me at birnbaum@sympatico.ca. Or, you can write me at #608-18 Deerfield Dr., Nepean, Ontario, Canada, K2G 4L1. ♦

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
John Matthew	jmatthew@totalsports.net	Apostrophes
Jim Box	im.box@duke.edu	Statistics
Rob Fabrizzio	rfabrizzio@bigfoot.com	Statistics
Duke Rankin	RankinD@montevallo.edu	Statistics
Keith Karcher	kckarcher@compuserve.com	
Tom Hanrahan	HanrahanTJ@navair.navy.mil	Statistics
Steve Wang	Steve.C.Wang@williams.edu	Statistics
Larry Grasso	l.grasso@juno.com	Statistics
Keith Carlson	kcarlson@stlnet.com	Economics/Econometrics/Statistics
John Stryker	johns@mcfeely.interaccess.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 (www.microsoft.com).

To get on the electronic subscription list, send me (Phil Birnbaum) an e-mail at phil_birnbaum@iname.com. (That's an underscore _ between Phil and Birnbaum.) 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. The electronic copy is sent out two business days after the hard copy, to help ensure everyone receives it at about the same time.

E-Mailing BTN

I have been told that my iname.com and philbirnbaum.com e-mails often don't work. Those go through internet forwarding services before they reach me, and it seems these services are unreliable.

My "real" e-mail address – for now – is birnbaum@sympatico.ca. If that ever fails to work – who knows, I may change service providers again someday – try either phil_birnbaum@iname.com, or BTN@philbirnbaum.com.

Sorry for any inconvenience.

Bill James Worth Reading on Managers

John Matthew IV

Back in the 1980s, every spring I would call my local bookstore asking if they knew when the new Bill James book was being released. It was a major event I looked forward to. I would buy the book as soon as it was available and read it cover to cover immediately. Well, Bill James stopped writing the Baseball Abstracts in 1988. He wrote a book in a different style from 1990-1995 and by then the annual ritual had lost its excitement.

In 1997, Bill wrote a guide to baseball managers and it took two years for me to buy it and read it. I should have read it earlier. While not destined to be a must read like his Abstracts were, it is certainly worth reading.

Everyone has opinions about managers. We could all manage our team better than the guy whose job it is to do so, but how much do we really know about managers beyond a very vague, “he is good” and “he is bad”? James tries, and succeeds, in getting the reader to focus on what makes a manager distinct.

The Bill James Guide to Baseball Managers from 1870 to Today

By Bill James

Scribner, 352 pages, $30
ISBN 0-684-80698-3

He does this in a straightforward way that makes for great but simple reading. For every decade from the 1870s to the 1990s, a “Snapshot” is presented. Each snapshot briefly lists the most successful managers, the most controversial mangers, plus others of note. Then there is a description of what the typical manager was like, the percentage of playing managers, player rebellions, managerial stunts, and evolutions in strategy and in the role of the manager. This paints the background for each decade.

He then analyzes a few managers “in a box” as he calls it. For each of these, he answers a long series of questions. They range from the basics -- year of birth, years managed, and record as a manager -- to detailed questions on what he brought to a ball club, how he used his personnel, what was his game managing and use of strategies, and how he handled his pitching staff. This is a much better way of describing a manager than I have seen before. It also allows you to easily compare managers across eras.

Additionally, throughout the book there are sidebars and short essays to fill things out. This is where Bill James’ excellent writing style shines through. A typical sidebar would be the brief biography of Bob Allen, who managed the 1900 Cincinnati Reds. Another would be the brief review of Paul Richards’ year as White Sox manager in 1976. The essays range from “The Marshalltown Enfant Terrible” on Cap Anson to “The Darrtown Farmer” on Walter Alston to “Ranking Managers.”

The last essay is the one that would get the most attention. Depending on how things are measured, either John McGraw or Joe McCarthy comes out on top as the greatest manager ever. James does discuss how difficult this is to measure. The goal of every baseball manager is to win. Does that make Connie Mack, who lost more games than any other manager, the worst all-time? No, he was one of the best -- but how can you prove that?

One of the most interesting essays is on manager statistics. If you look at *Total Baseball VI*, while there are many stats quoted for hitters and pitchers, managers only get nine per year: Team/League, Games, Wins, Losses, Percentage, Standing, Manager/Year, Expected Wins, and Actual Wins Minus Expected Wins. In a great essay called “The Manager’s Record”, James comes up with sixteen numbers that are much more informative. One such number is “LUp,” the number of different batting lineups used in a season. Some managers use the same lineup every day and some change it regularly. (Johnny Oates used only different 73 lineups for 163 games in 1996 while Bob Boone used a whopping 152 different lineups in 161 games.) How about “H&R” or “hit and run attempts”? Again using 1996, Bob Boone had 172 “H&Rs” and Art Howe had 61. Or what about “DS” (defensive substitutions)? Joe Torre led the majors in 1996 with 55 and Cito Gaston was last with only 11. These numbers tell you much more about a manager than his traditional record would.

A couple of typos were missed in the editing stage. For example, on page 293 there is this: “The baseball leadoff hitter in baseball history is Rickey Henderson.” But these small errors do not detract from this fine book.

What I found surprising was that for someone known for at least inventing the word “sabermetrics” if not the science, Bill referred only to the regular stats when he was describing a player. The man who taught me that batting average was not the best way to rate a hitter would nonetheless use that number when describing him. There is no mention of runs created or any other statistic that Bill James is famous for.

He did come up with Expected Wins. How many wins should a team be expected to win? The formula for the expected winning percentage is half of last year's percentage, plus an eighth of the two previous years' plus a quarter of .500. Below is a table for this year. Interesting that the manager who exceeded his expectations the second most in the AL was fired!

	Projected		Actual		
	W	L	W	L	Difference
Anaheim	82	80	70	92	(12)
Baltimore	83	79	78	84	(5)
Boston	87	75	94	68	7
Chicago	81	81	75	86	(6)
Cleveland	88	74	97	65	9
Detroit	69	93	69	92	(0)
Kansas City	74	88	64	97	(10)
Minnesota	74	89	63	97	(11)
New York	101	61	98	64	(3)
Oakland	75	86	87	75	12
Seattle	80	81	79	83	(1)
Texas	85	77	95	67	10
Toronto	83	79	84	78	1
Tampa			69	93	
Arizona			100	62	
Atlanta	98	64	103	59	5
Chicago	83	79	67	95	(16)
Cincinnati	78	84	96	67	18
Colorado	80	83	72	90	(8)
Florida	69	93	64	98	(5)
Houston	92	70	97	65	5
Los Angeles	84	78	77	85	(7)
Milwaukee	77	85	74	87	(3)
Montreal	74	89	68	94	(6)
New York	84	78	97	66	13
Philadelphia	75	87	77	85	2
Pittsburgh	74	88	78	83	4
San Diego	90	72	74	88	(16)
San Francisco	85	78	86	76	2
St. Louis	82	80	75	86	(7)

In summary, this is a book focusing on a part of the game that is often overlooked. We frequently say, "Joe Torre is a great manager," but, if asked to explain why, we say, "Well, his team wins the World Series." However, we never stop to think why his teams win. This book makes you ask the questions why. It is also a good read. Bill James is, most importantly, a good writer. If he wrote about stamps instead of baseball, I probably would be a philatelist instead of a sabermatrician.

John Matthew IV, 167 Church St. #400, Toronto, Ontario, Canada, M5B 1Y4, john.matthew@home.com. ♦

POP Analysis of Career Walks

Mike Sluss

How best can we evaluate who had the most outstanding career walk numbers? Ranking by raw totals shortchanges those who had shorter careers, and ranking by walk rate shortchanges those who walked not as often but in more seasons. Here, the author suggests ranking by POP – a statistic that scores the probability that an average player would match outstanding season totals.

How best can we evaluate who had the most outstanding careers for getting walks?

Should "most outstanding" be defined as the most walks? If so, Babe Ruth (2056 walks) beats Ted Williams (2019 walks), even though Williams has the better rate of walks.

Should "most outstanding" be defined as the highest rate of walks? If so, Ted Williams (rate of .208) beats Babe Ruth (rate of .197), even though Ruth has more walks.

Should the evaluation of career rates of walks require a minimum number of career walks or plate appearances (PA) to qualify for consideration? If so (e.g. a player needs at least 5000 PA), then Rod Carew (rate of .099 for 10333 PA) ranks above Ferris Fain (rate of .187 for 4834 PA), even though Fain had almost as many walks (904) as did Carew (1018).

Should the evaluation of career walk performances ignore relativity (the differences in leagues' average likelihood of walks from one year to another)? If so, then Mark McGwire (rate of .170 for 6813 walks) beats Topsy Hartzel (rate of .147 for 5685 walks). even though McGwire has played in leagues with an overall average rate of walks 34% higher than did Hartzel.

I am proposing that the Probability of Performance (POP) analysis best evaluates bases on balls performances. The POP analysis defines the "most outstanding" performance as being the least probable (above average) performance.

BB POP determines the probability of an average league player obtaining at least that number of walks (and at least that rate of walks) in a specific number of plate appearances. BB POP uses 3 factors to measure the excellence of a player's walking performance: (1) his total of walks, (2) his plate appearances, and (3) the league average rate of walks. A batter's BB POP is higher if his number of walks is higher, if his rate of walks is higher, or if the league average rate of walks is lower.

BB POP is derived from the binomial probability function, where

W = batter's walks for a season,

A = batter's plate appearances for a season
= at bats + walks,

L = league average rate of walks
= (league walks - W) / (league at bats + league walks - W - A),

P = the probability that a batter would walk at least W times in A plate appearances (for a walk rate of at least A / W) if his assumed walk rate were L, the average of the other players in the league)

$$P = \sum_{w=W}^{A} \frac{w!}{(A-w)!} L^{w} (1-L)^{(A-w)}$$

Because probabilities are so small that average players will achieve outstanding seasons, BB POP is the negative logarithm of the actual probability P:

BB POP = - log10P

It should be noted that the calculation of POP excludes the individual batter's data from the league average. Also, POP does not require arbitrary criteria for consideration (e.g. a minimum 5000 PA). A player's cumulative career POP is the sum of his seasons' POPs, representing the probability of an average player achieving each year's performance. A player's career POP as currently defined is clearly set in historical perspective and cannot decrease. There is no reason to arbitrarily consider only the first 8000 at bats or plate appearances in order to exclude the frequent end-of-career decreases in rate of productivity (although this seems not as common for walks). Poor performances in later years will not decrease a career POP, but neither will they increase it. POP is a positive measure of achievement.

Career Best Batters' POP for Bases on Balls

Rank	Name	BB POP	BB AV	BB	PA	Yrs
1	BABE RUTH	313.77	.197	2056	10455	22
2	TED WILLIAMS	251.63	.208	2019	9725	19
3	JOE MORGAN	195.52	.167	1865	11142	22
4	Rickey Henderson #	186.48	.166	1890	11363	20
5	MEL OTT	185.60	.153	1708	11164	22
6	MICKEY MANTLE	183.85	.176	1734	9836	18
7	Max Bishop	172.80	.204	1153	5647	12
8	Barry Bonds #	156.82	.170	1357	7978	13
9	Roy Thomas	152.07	.164	1042	6338	13
10	Eddie Yost	148.13	.180	1614	8960	18
11	H. KILLEBREW	140.76	.161	1559	9706	22
12	Jack Clark	135.64	.156	1262	8109	18
13	JOHN MCGRAW	131.67	.176	836	4760	16
14	Gene Tenace	130.30	.183	984	5374	15
15	BILLY HAMILTON	130.02	.159	1187	7456	14
16	Darrell Evans	124.03	.152	1605	10578	21
17	Jimmy Wynn	123.35	.155	1224	7877	15
18	MIKE SCHMIDT	120.73	.153	1507	9859	18
19	Eddie Stanky	119.93	.188	996	5297	11
20	Topsy Hartsel	119.40	.147	837	5685	14
21	WILLIE MCCOVEY	118.89	.141	1345	9542	22
22	LOU GEHRIG	116.90	.159	1508	9509	17
23	EDDIE MATHEWS	114.81	.145	1444	9981	17
24	Mark McGwire #	110.21	.170	1052	6183	13
25	C. YASTRZEMSKI	108.63	.133	1845	13833	23
26	Frank Thomas #	103.81	.183	989	5395	9
27	Yank Robinson	97.73	.162	664	4092	10
28	Ken Singleton	97.01	.149	1263	8452	15
29	EDDIE COLLINS	96.78	.131	1503	11454	25
30	Dolf Camilli	94.91	.150	947	6300	12

# = active player   CAPS = hall of famer

Mike Sluss, 2847 Pioneer Drive, Green Bay, WI, 54313-5857, akili2000@aol.com. ♦

Reliability of Statistics

Willie Runquist

A player's statistics can vary from his innate ability if he benefits or is hurt by the breaks of the game. How can we determine, then, if a player's performance is significantly better than another's, or if the difference is due to chance? Here, the author explains how the statistical concepts of Standard Error and Reliability can help answer this question.

Cal Ripken is badly fooled on a pitch but the ball drops behind the second baseman for a base hit, but Omar Vizquel ropes a line drive directly at the shortstop which ends up as a double play. Kenny Lofton is called safe on a close play at first even though the throw was there first, while Edgar Martinez beats the throw but is called out. There are good bounces and bad bounces, and participants and fans alike philosophically accept the breaks and hope that in the long run they will even out, the better team will win, and the brighter star will shine.

The statistics of baseball, since they record the outcome of every plate appearance, are equally contaminated. Cal Ripken had 115 singles in 1996 and Omar Vizquel had 111, but how much of that difference was simply a matter of the breaks? In one sense, it is a moot point. After all, as far as the outcome was concerned, Ripken did indeed have four more singles than Vizquel regardless of how those singles were attained, and as a description of the result of those plate appearances, the numbers represent the facts. But when we attempt to evaluate a player's performance the issue is joined. Every statistic we compute therefore contains two parts: the "true" value of the statistic and chance "error".

What do we mean by *true value*? The true value is that value which would be obtained if the player had an infinite number of at bats. It will include any external conditions that systematically differ from player to player as well as the player's ability, and these factors will not average out no matter how often the player comes to the plate. Chance effects (error), however, are random, and while they influence a particular time at bat, will balance out between players over a large number of plate appearances. It does not matter what the source of these effects is, it is only necessary that they potentially even out. The two components are simply defined in terms of the way they affect the player's performance. In this essay we will deal with two related problems: (1) evaluating the contribution of random error to the value of the statistic for a single player, and (2) evaluating the ability of the statistic to indicate true differences between players.

The Standard Error

The standard error (SE) is the most common statistic for assessing the accuracy of an average. Most baseball statistics are averages over plate appearances, and, for any average, an estimate of the SE may be computed. Confidence intervals derived from the SE will then provide an estimate of the accuracy of that statistic.[1] Since the true value of the statistic will lie within one SE of the obtained value about 68% of the time and within two SEs about 95% of the time, a 20-point standard error for a player's batting average means that the player's average may be thought of as "accurate" within 20 points about two-thirds of the time, and within 40 points about 95% of the time.

Table 1: Standard Errors for Batting

	BA	OBA	SA	IP	OPS
AL	.023	.023	.046	.030	.059
NL	.024	.024	.046	.030	.061

	1B	2B	3B	HR	BB
AL	.020	.011	.004	.009	.013
NL	.021	.012	.004	.009	.015

Standard errors for individual players may vary considerably from player to player, but if the number of at-bats or plate appearances does not vary too much from player to player, it is possible to obtain a pooled SE for that group of players that provides a general estimate of the accuracy of the statistic as a whole.[2]

[1] Standard errors were computed by first computing the variance of the results of each plate appearance or at bat. The estimated standard error of the mean of those measures is then equal to SQRT(var /AB or Var/PA). For proportions, SE is usually computed as SQRT((p*(1-p))/PA . This formula is algbraically equivalent to the longer one above when the values are all 1 and 0. To estimate an SE it is necessary that the statistic be additive over PA or AB, ie., be the average of a value that exists for each PA. Statistics such as the popular runs created do not admit to an E because the sum of the individual values for each PA will not sum to the RC for the total.

[2] Pooling SE is not simply taking their average. In a pooled SE players contribute to the final value in proportion to their plate appearances or at bats.

Table 1 presents the pooled SE for a number of different averages for players from the 1996 season. The results are for 187 American League and 190 National League players that had more than 125 at bats. The statistics were batting average, on base average, slugging average, isolated power, and OPS (on-base plus slugging).[3] The SE for batting runs (linear weights) was also computed (but there are some special problems with this measure and it will be considered separately below). SE was also computed for the separate events in the batting line: singles, doubles, triples, homeruns, walks, and stolen bases, each averaged over either AB or PA. Plate appearances were counted as at bats plus walks so that averages may not appear exactly as in the official statistics.

Table 2 presents the SE for pitcher's counterparts of the five averages. Separate tabulations were made for starting and relief pitchers. A pitcher was considered a starter if he started more than ten games without making half as many relief appearances as starts. Relief pitchers were those who made at least 25 relief appearances. (The American League supplied 81 starters and 71 relievers while the National League supplied 77 starters and 81 relievers.) In addition, standard errors were computed for four pitching averages based on batter's faced (Outs, Bases Given Up = TB +SB+WP, SO, BB) and ERA.[4]

Table 2 – Standard Errors for Pitching

Starters	BA	OBA	SA	IP	OPS
AL	.018	.018	.036	.024	.048
NL	.021	.012	.034	.022	.045
	OUTS	BGU	SO	BB	ERA
AL	.018	.034	.014	.011	.50
NL	.017.	032	.014	.010	.50

Relievers	BA	OBA	SA	IP	OPS
AL	.029	.029	.055	.036	.070
NL	.028	.028	.051	.033	.067
	OUTS	BGU	SO	BB	ERA
AL	.028	.038	.024	.019	.75
NL	.027	.048	.023	.017	.66

In structure, batting runs is just like slugging average, except that each quantity is weighted by its theoretical ability to produce runs rather than bases gained. Batting runs, however, usually includes stolen bases and caught stealing. Without play by play data, it is not known on which plate appearance a stolen base attempt occurs. Therefore, the SE for batting runs was estimated in two ways: one version in which steals were omitted, and one in which some simple assumptions were made about the distribution of attempted steals among different batting outcomes.

For the simple version, the SE was an identical .022 for both leagues for batting runs per plate appearance. With stolen bases included, the standard errors were both .019. Batting runs, however, is normally expressed as a total for the season rather than an average. Translated to full season totals based on 625 plate appearances, batting runs had standard errors of 9.38 for the American League and 7.42 for the National. With stolen bases, the values were 8.02 and 7.42. For individual players, the seasonal value is obtained my multiplying the players SE his plate appearances.

The SEs for the individual events may also be expressed as season totals. Based on 550 AB or 625 PA, these values were nine singles, five doubles, two triples, four homeruns, four walks, and four stolen bases.

The SE also provides an exact picture of the relationship between PA (or AB) and accuracy. SE for a typical BA is .018 for 400 AB, .022 for 200 AB, and .045 for 100 AB. The range and distribution of AB will therefore have an effect on the pooled SE.

However, empirically, the pooled values are remarkably stable for different groups of players. Standard errors for the 50 batters having the most plate appearances in each of five different seasons (1996, 1993, 1990, 1980 and 1970) never differed by more than .001. A definitive statement about a particular player, however, requires a reference to his own SE. For example, Jeff Bagwell had an slugging average of .507 with SE = .044, while Ron Gant slugged .504 with an SE of .054.

Reliability

[3] Estimating SE for OPS is not straightforward because the two components are based on different sets of plate appearances. The problem may be sidestepped by averaging total bases and times reached base to plate appearances then multiplying the mean and variance by a constant that corrects for the different denominators.

[4] The use of outs as a denominator in pitching statistics such as ERA (and in some offensive statistics such as total average) does not lend itself to direct computation of SE because an ERA is not the mean of the values for each out. The variance of ERA used to get SE for each pitcher was obtained by first computing the variance for ER over batters faced (ER /BF) then multiplying this variance by the total BF / O for that pitcher. There is also some inaccuracy introduced in all of the pitching statistics because 3 x IP was used as the value for outs. A measure such as AB- H would probably be more accurate since some of the "innings" included outs made on base including those originally put on base by other pitchers.

While the SE provides a convenient and useful measure of the accuracy of a given statistic, it is most meaningfully interpreted in relation to the differences between players on the statistic in question. These differences are indicated by the variance among the players in a particular group (Var P), while the SE is based on the pooled variance among the values for each plate appearance for a player (Var E).

Reliability (R) may be simply defined as

```
R= 1 - VarE/VarP
```

In other words, R is the proportion of the variance (differences) among players which is not due to random error. If there were no random error, R would be 1.00 and all of the differences between players would represent true differences. Conversely, a value of zero would mean that all of differences between players were random. R is not a general property of a given statistic, however, but depends upon the particular group of players being considered. It is basically an index of how successful a statistic is in reflecting true differences between the players in that group on that statistic. Tables 3 and 4 present the R values for the measures and players from the earlier table.

Table 3 – Reliability of Batting

	BA	OBA	SA	IP	OPS	BR	BR+SB
AL	.52	.73	.73	.82	.68	.73	.81
NL	.37	.68	.70	.73	.64	.68	.77

	1B	2B	3B	HR	BB	SB
AL	.64	.41	.39	.84	.85	.92
NL	.58	.38	.46	.80	.87	.67

The most notable feature of the batting statistics is the low reliability of batting average, especially in the National League where almost 2/3 of the variance between players may be considered random. Indeed, the lack of validity usually attributed to BA may result in part from its inability to differentiate among players. Note that the SE for on-base average is the same as that for BA, but OBA has greater reliability because the differences among players is larger.

Slugging average suffers because it includes batting average as a component. When that component is removed, (giving isolated power, or IP), reliability improves considerably. Nevertheless, the large differences in slugging average override the larger SE for this measure. Batting runs has about the same reliability as slugging average.

R is also affected by the number of plate appearances. The batting-average R for 50 players that had fewer than 270 plate appearances was .33, while for isolated power R was still only .63. However, over a career, even batting average becomes more reliable. A sample of 94 players from the 1970s with over 5000 career plate appearances produced R^{1}s of .91 for batting average, .95 for on-base average and .98 for isolated power.

Table 4 – Reliability of Pitching

Starters	BA	OBA	SA	IP	OPS
AL	.34	.54	.48	.46	.47
NL	.52	.65	.51	.47	.55

	OUTS	BGU	SO	BB	ERA
AL	.53	.50	.85	.71	.66
NL	.63	.52	.88	.81	.71

Relievers	BA	OBA	SA	IP	OPS
AL	.42	.36	.40	.43	.37
NL	.31	.41	.26	.21	.25

	OUTS	BGU	SO	BB	ERA
AL	.48	.33	.82	.50	.60
NL	.29	.27	.75	.78	.52

The fact is that with the exception of a few very good players, most major league players do not differ from one another on batting average to any great extent. A difference in batting average of 20 points is a difference of only 10 hits in 500 at bats, easily within the range of a few good or bad breaks. Regardless of their other talents, players who cannot maintain a minimal average do not play much or soon disappear. A sample of 100 high school players would probably yield higher reliability for even batting average.

For the individual categories, homers, steals, and walks are the most reliable while doubles and triples fare the worst. These results are important for understanding how reliability varies for different combined statistics. The reliability of a statistic is a function of the reliability of the various components and the degree to which component contributes to the whole. In general, adding a highly reliable component will improve reliability and even more so if the new component is heavily weighted. On-base average is more reliable than batting average because of the inclusion of walks, while the reliability of any total offense measure is largely determined by the weight in given to home runs. This fact also explains the counterintuitive finding that including stolen bases results in a smaller standard error and higher reliability.

However, it is not all that simple. The variance between players is greatly enhanced by a few players who outstrip the pack by a large margin. This is particularly true of stolen bases, where 14 players account for 50% of the steals, but is also true in the case of home runs. Most players do not differ by much, but a few do so by a large amount and their speed and/or power is not only a legendary but a reliable commodity. The low reliability of doubles, however, is somewhat surprising. Most players do not differ to any great extent in the frequency of two-baggers and even those who excel in this department do not do so by a large amount. Perhaps, the notion of "gap power" is overdrawn, particularly if players are so identified by an unreliable measure such as doubles. Triples are simply too infrequent to produce much variation.

There is no statistical reason why one could not adjust averages in order to control for identifiable sources of variance such as seasonal differences or park effects. In so doing, however, the result is likely to make players even more homogeneous, since these factors contribute to the variance between players. Taking only road games eliminates some of the park bias. Reliability computed only from road games may be less than that when home games are included, not just because of fewer plate appearances but because a significant source of variance between players has been removed. .

Correcting a player¹s statistics with some multiplicative factor such as a park effect will also affect reliability in that the multiplicative operation will change the player¹s variance between his plate appearances by the square of the constant.

With the exception of strikeouts and walks, the reliability of most pitching measures is questionable, thus underscoring the prevalent opinion among many baseball experts that pitching numbers are pretty much unpredictable. Most pitchers do not even differ sufficiently in their susceptibility to extra base hits to overcome the random error in these measures. Batting average may be somewhat more reliable than that for batters, however, simply because of more batters faced than plate appearances. Most of the reliability figures for relief pitchers are lower than those for starters because of the huge difference in batters faced. The oft cited "runners stranded" measure is so unstable as to be practically worthless. In the face of these difficulties, some scouts seem to have turned to strikeout/walk ratio as a prognostic measure for pitchers. This would seem to be based more on the fact that the K/BB ratio is highly reliable, rather than on its inherent validity.

Final Comment

One should not abandon a useful statistic even if it cannot make fine discriminations between players. On the other hand, there are practical consequences of low reliability. One cannot have very much confidence in any statistical difference between two players that is less than the square root of the sum of the squares of the two players' standard errors. For example, batting averages must differ by at least 28 points. Differences in power measures must be even larger.

Perhaps of more importance, at least to general managers, the relative amount of random error in a statistic severely limits the accuracy of any prediction from one season to the next or from one set of games to another, even if nothing changes from year to year. On average, a player¹s statistics differ by about one standard error from one season to the next and correlations between full season statistics mirror the reliability of the statistic. Taking some of the idiosyncrasies of individual players into account, and adjusting for statistical regression may improve predictions, but the contribution of random error to all known measures is large enough to mitigate any claims for real accuracy in such predictions. The solution lies not in devising more and more statistics from the traditional categories, but in improving the data base from which the statistics arise.

Willie Runquist, Box 289, Union Bay, BC, Canada, V0R 3B0, willipeg@island.net. ♦

BBs, IBBs, HPs and Pitching Around in 1998

John F. Jarvis

The traditional statistics tell us how many times a player was walked intentionally, and how many times he was walked unintentionally. But some of those "unintentional" walks are "pitching around" walks – plate appearances in which the batter was not deliberately walked, but neither was given anything decent to hit. Here, the author studies whether an estimate of the number of "semi-intentional" walks can be determined.

In addition to setting the Major League single season record with his 70 home runs in 1998, Mark McGwire set a National league record with 162 bases on balls (BB). This tied him with Ted Williams, who did it twice (1947, 1949), for second on the ML single season list behind Babe Ruth's 170 BB in the 1923 season. In each of these four seasons, the leader in BBs was also the season Home Run (HR) leader.

To increase chances for outcomes favorable to the defense, pitchers sometimes will purposely throw four balls to a hitter, intentionally walking him. Twenty-eight of McGwire's BBs were of the intentional variety (IBB). In this category he was one behind the ML leader, Barry Bonds.

Beside the IBB, there is the much-commented-on practice of "pitching around" a strong hitter. That is, the pitcher doesn't officially give an IBB, but doesn't give the hitter any legitimately good pitches either. There is no serious attempt to get the hitter out and a BB usually occurs. How many of McGwire's BBs are "pitching around" BBs?

Being hit by a pitch (HP) can be considered to be a very intentional base on balls. McGwire was hit a rather unremarkable six times during the 1998 season. Jason Kendall, the season leader in the HP category, was hit 31 times.

McGwire's 70 home runs (HR) came in just 509 at bats (AB), a rate of one HR every 7.3 AB. Obviously, McGwire's 162 BBs cost him a lot of ABs. While unintentional BBs are an inevitable part of the game, the intentional variety purposely prevents hitting. A measure of McGwire's lost opportunities can be estimated by adding the fraction of his 134 unintentional BBs that represent pitching around to his 28 official IBBs. Of course, this requires identifying "pitching around" BBs. These, of course, are not one of the categories coded in the play-by-play files.

According to TV play-by-play announcers, an IBB is usually indicated any time there is a runner on second and none on first. Actually, managers don't seem to do this as often as the game announcers suggest they should, and in real life the IBB tactic is employed in a somewhat more subtle manner than this. IBBs given outside of well defined tactical situations could be called "gift" BBs. Another question that arises is this: How many of McGwire's official IBBs came in situations where tactics wouldn't suggest one?

If managers use the IBB in a consistent way and if there are not an excessive number of "pitching around" BBs, various statistical pattern recognition methods can be used to create a classifier that labels each situation as matching an IBB tactical situation or not. The actual event is known and is not subject to being changed. Only the situation in which it occurs, the context for the event, is being classified. The particular statistical tool that seems most appropriate for this task is the neural net ("Neural Networks", Laurene Fausett, Prentice Hall, 1994). Creating a neural net requires determining the elements of the event context that contain information that helps in deciding when an IBB should be given. A list of BB and IBB events and the context in which they occur becomes the input to the neural net training procedure. The neural net training procedure adjusts its internal numerical coefficients to minimize the mismatches between its output and the actual events recorded in their contexts.

Using the Retrosheet and Total Sports play-by-play game accounts for the 1980 to 1998 seasons, I tabulated all BB (237062), IBB (24076) and HP (17720) events and their contexts. For each of these events, the following information at the time of the BB, IBB or HP -- the "event context" -- was recorded: inning, outs in the inning, runners on base, relative score, player receiving the pass and player following the pass. (Relative score is the difference between the offensive team and defensive team runs at the time of the event.) Player information was used to access a full season hitting measure, typically the slugging average.

Neural Net training sets were formed by taking all recorded IBBs and a similar number of BBs chosen at random from all the available events. HP events are not used in the training process. Once the neural net was trained, it was (and is) possible to use it to classify the context in which a particular event occurred, not only from the training set but any other events where the same context information has been determined.[1]

This neural net requires nine times the computation of evaluating a linear regression equation for the same input parameters in addition to the nine activation function evaluations. It contains 81 coefficients which, unlike a linear regression formula, do not have an interpretation in terms of the input parameters.

When presented with a particular context, the trained neural net generates a number between +1 -- most IBB like -- and -1 -- most BB like. The decision point in declaring a particular context BB or IBB is set midway between the average of the neural net output values for the training set IBB event contexts and the average of the BB event neural net outputs. Repeating the training using other randomly chosen groups of BBs produces essentially the same results. Splitting the data by league or decade (80s and 90s) also produces equivalent results.

The relative importance of each parameter in the event context for determining if an IBB should be offered can be assessed by training eight additional neural nets each with one of the eight event context parameters not used. The items are ranked in importance by how much the accuracy of the neural net is degraded by its absence. Using this procedure the ranking of the data items in the event context from most to least important is: a runner on second, a runner on third, a runner on first, the relative score, the inning, the number of outs in the inning, the slugging average of the batter following the event and last is the slugging average of the batter receiving the IBB. This ordering is largely confirmed by correlation coefficients calculated from the same data (Table 1). The relative unimportance of hitting prowess in this is somewhat surprising. Using season averages rather than characterizing hot streaks may partially explain the low importance assigned to hitting. Still, there's some vindication in this for my single-minded announcer.

Table 1– Linear Regression Coefficients and Correlation Results

parameter	weight	r2
constant	-1.0453	
runner on first	-0.4305	0.107
runner on second	0.7073	0.393
runner on third	0.5110	0.179
inning	0.0614	0.097
outs	0.1201	0.094
relative score	0.0371	0.088
gets BB	0.8105	0.013
after BB	-0.7835	0.022

(standard deviation = 0.495)

Linear regression can be used in a way equivalent to the neural net to classify BB/IBB situations. A linear regression done on the same data used in the neural network training yields the results given in Table 1. Also included in Table 1 are the correlation coefficients between the input parameters and the BB/IBB values. In the regression bases have the value 1 if a runner is present, 0 otherwise. Innings are in the range 1 - 10, with all extra innings given the value 10. Outs have the values 0, 1 and 2. The remaining three parameters are defined the same way as for the neural net.

Table 2 compares the classification results of the neural net and linear regression on all the 1980-1998 data. The neural net reduces misclassifications by 38% compared to the linear regression based classifier.

Table 2 – Classification of all 1980-1998 BBs, IBBs and HPs by a Neural Net (NN) and by Linear Regression (LR)

	BB as			IBB as			BB+IBB			HP as		
	BB	IBB	frac	IBB	BB	frac	correct	incor	frac	BB	IBB	frac
NN	214827	22235	0.906	23211	865	0.964	238038	23100	0.912	16061	1659	0.906
LR	200787	36275	0.847	22919	1157	0.952	223706	37432	0.857	14983	2737	0.846

Players in Table 3 were selected because they were season leaders in BB, IBB or HP or HRs and are ordered by total BBs. McGwire, of course, was the HR and BB season leader. Jason Kendall was the 1998 leader in HPs with 31. Andres Galarraga was tied for second in HP. Sammy Sosa was second in season home runs. Barry Bonds, Rickey Henderson and Frank Thomas were second, third and fourth in BBs. Bonds was the ML leader in IBBs with McGwire second in this category. Ken Griffey, Jr. was third in home runs. The last line presents the

[1] For those technically inclined, the neural net used is a standard back propagation net with one layer of hidden units. Eight input units are used (corresponding to each parameter in the recorded context where each base was given a separate input unit) and eight hidden units are used. A hidden unit is a weighted sum over the input unit values plus a constant. The result of this sum is passed through an activation function, f(x), to generate the hidden unit output. The activation function in this case is sigmoidal, having asymptotic values of +1 and -1 and a slope of 1.0 at x=0. The output unit is a sum of a constant term and the weighted outputs of the hidden units. Subjecting the output unit sum to the same activation function completes the neural net calculation.

totals for the eight players. For the entire 1998 season, the neural net classifier correctly labels 90% of BBs and 93% of IBBs. About 88% of HP events occur in BB situations, not greatly different than the fraction of BBs classified as BBs by the neural net.

The practice of pitching around a batter can be identified with BB situations classified as IBB. Comparing the totals (line "players") for the selected players to the season totals suggests that this occurred about the same with the selected players as in the season as a whole. While McGwire's fraction of BBs classified as coming in IBB situations by the neural net is slightly less than the overall season average, he didn't receive a disproportionate number of "pitching around" BBs. Rickey Henderson, who didn't get any free passes either, had a slightly higher fraction of BBs that came in IBB situations. Opposing teams will pitch carefully to Henderson but they really don't want him on base.

IBBs classified as coming in BB situations are in the category of gift IBBs. Excepting Bonds and McGwire, the selected batters as a group receive these passes at about the same rate as the league as a whole. Bonds, who received a very rare IBB with the bases loaded, and especially McGwire, are clearly in a different category, receiving many more of these gift IBBs than the season average.

Table 3 – Neural Net Classifications for the 1998 Season and Selected Players

	BB as			IBB as			BB+IBB			HP as		
	BB	IBB	frac	IBB	BB	frac	correct	incor	frac	BB	IBB	frac
all 1998	13861	1509	0.902	996	70	0.934	14857	1579	0.904	1399	187	0.882
Kendall	44	4	0.917	3	0	1.000	47	4	0.922	29	2	0.935
Galarraga	48	4	0.923	11	0	1.000	59	4	0.937	24	1	0.960
Sosa	55	4	0.932	13	1	0.929	68	5	0.932	1	0	1.000
Griffey	59	6	0.908	9	2	0.818	68	8	0.895	7	0	1.000
Thomas	100	8	0.926	2	0	1.000	102	8	0.927	5	1	0.833
Henderson	102	16	0.864	0	0		102	16	0.864	5	0	1.000
Bonds	96	5	0.950	23	6	0.793	119	11	0.915	7	1	0.875
McGwire	117	17	0.873	19	9	0.679	136	26	0.840	5	1	0.833
Players	621	64	0.907	80	18	0.816	701	82	0.895	83	6	0.933

The HPs received by the selected hitters classify as BBs at a slightly higher, but not (statistically) significantly different, rate than the season average. The league leaders in BBs, IBBs and home runs are not among the season leaders in HPs. There is no evidence from this small selection of players that HPs were used instead of IBBs or that they were specifically directed at the home run leaders.

The excellent agreement achieved by the neural net in classifying BB and IBB situations confirms that the IBB is only given in well defined tactical situations. The low incidence of "pitching around" BBs and "gift" IBBs suggests that McGwire's high BB total is more a function of his discrimination at the plate than opposing managerial intent.

Tabulating McGwire's total BB by quarters of the season (as equally as 162 can be divided by 4) yields the following: 48, 34, 50, 30 number of BBs in the 4 quarters. There is no suggestion in this that he was subject to "special" treatment during the latter, the most publicized, part of the home run record chase.

McGwire did not receive a disproportionate number of "pitching around" BBs, estimated as 17 by the neural net. His 28 IBBs -- one less than the season high -- includes 9 that came in situations where the IBB is not normally given. His HP total is comparatively low. He appears to have been given the respect all power hitters command but there is little or no evidence that opponents specifically tried to hinder him during the HR chase.

Combining McGwire's IBBs and "pitching around" BBs, those in contexts labeled IBB by the neural net, yields 45 events. If he had been allowed to hit, 45*509/(509+162-45) = 36 of these events could have been expected to result in ABs. At the rate he hit HRs during 1998 this could have resulted in 4 or perhaps 5 additional HRs.

John F. Jarvis, Department of Mathematical Sciences, University of South Carolina – Aiken, 471 University Pkwy., Aiken, S.C. 29801, jfj@pacer1.usca.sc.edu. ♦

Submissions

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.)

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.

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. Since we've hardly published for the last couple of years, even reviews of older books – say, 1997 or later – would be welcome. The only restriction, please: the book should have, or claim to have, some Sabermetric content.

See Clifford Blau's review in the last issue, or John Matthew in this issue, for the kind of thing we're looking for.

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.