By the Numbers

Volume 12, Number 2 The Newsletter of the SABR Statistical Analysis Committee May, 2002

Review

Intriguing Analysis in 2002 "Baseball Prospectus"

Clifford Blau

The 2002 Baseball Prospectus, in its eighth annual edition, this year includes many worthwhile articles, the author says in this review.

The 2002 Baseball Prospectus, the eighth annual edition, is the first one I have read. I was intrigued by several essays that should be of interest to members of the Statistical Analysis committee.

There are several introductory articles by Clay Davenport that explain the statistical methods used in the book to evaluate players. Of particular interest is his piece on fielding statistics. He starts with the basic concept of Range Factor, that fielders should be evaluated based on the number of plays they make. Then he makes a number of adjustments to remove various outside influences. This makes the resulting figures more reflective of the player's individual accomplishments. Along the way, Mr. Davenport presents lists of the top ten fielders ever at each position (except pitcher). My main complaint about his methodology is that when a statistic is strongly affected by park factors, such as third baseman putouts, he drops it rather than make the appropriate adjustment. Surely sufficient data exist to calculate park factors for fielding. He also introduces a new rating for pitchers ("Stuff") which is useful for predicting young pitchers' futures.

After the main part of the book, which evaluates the recent seasons of over 1,600 players and predicts their 2002 performance, come articles on various topics. A couple of noteworthy ones are by Keith Woolner and Michael Wolverton. Mr. Woolner's piece does an excellent job of explaining the concept of replacement level. He makes clear why average players have value, and he uses historical data to determine the offensive replacement level for each position, and designated hitters. It is usually about 80% of the average rate of production, although a seemingly anomalous finding is that it is only 75% for first basemen. Why a position that has a larger pool of potential replacements should have a lower replacement level isn't clear. Perhaps the fact that almost anyone can play the position leads managers to make suboptimal decisions (they have too many options).

Mr. Wolverton's essay deals with the problem of evaluating players using only a few seasons rather than his whole career. He revisits the issue that Bill James dealt with in *The Politics of Glory* in comparing Don Drysdale and Milt Pappas. Using a methodology that looks at the chance that any given player-season would help a random team win a pennant, he finds that high peak value players are slightly more valuable than low peak value players with similar career totals. One small problem with his calculations is that he uses a single figure for standard deviations throughout the 20th century, even though there was a sharp increase in competitiveness in the first few decades. A surprising finding: even using a measure of the value of a player's peak years, Sandy Koufax ranks as only the 42th best pitcher overall.

There are several other interesting articles, too. Everyone interested in the statistical analysis should consider purchasing this book.

Clifford Blau, 16 Lake St. #5D, White Plains, NY, 10603, brak@erols.com ♦

In this issue

Get Your Own Copy

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

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

To join, or for more information, send an e-mail (preferably with your snail mail address for our records) to Neal Traven, at beisbol@alumni.pitt.edu. Or write to him at 4317 Dayton Ave. N. #201, Seattle, WA, 98103-7154.

E-mail Changes

If you normally receive "By the Numbers" by e-mail, but you found this issue in your physical mailbox instead, it's probably because your e-mail address changed. If you'd like to switch back to an e-mail BTN, please drop me (Phil) a line with your new e-mail address, and I'll switch you back to the electronic version. I'm at birnbaum@sympatico.ca.

Academic Research: Pitching Arm Fatigue

Charlie Pavitt

The author describes two academic studies on how fatigue affects pitching arms, one study of little-leaguers, another of pros.

This is one in a series of occasional reviews of sabermetric articles published in academic journals. It is part of a project of mine to collect and catalog sabermetric research, and I would appreciate learning of and receiving copies of any studies of which I am unaware. Please visit the Statistical Baseball Research Bibliography at *www.udel.edu/johnc/faculty/pavitt.html*, use it for your research, and let me know what I'm missing.

Stephen Lyman, Glenn Fleisig, John Waterbor, Ellen Funkhouser, LeaVonne Pulley, James Andrews, David Osinski, and Jeffrey Roseman, Longitudinal Study of Elbow and Shoulder Pain in Youth Baseball Pitchers, Medicine and Science in Sports and Exercise, Volume 33 Number 11, November 2001, pages 1803-1810

Stephen Lyman is a SABR member and co-founder of the Washington, D.C. area sabermetric discussion group that I mentioned in this column a couple of issues back. The sixth of the listed authors is the well-known fixer of injured pitching arms. After its publication, Reuters picked up on the study, leading to the publication of summaries in some newspapers around the country. No surprise; parents and coaches of Little League pitchers ought to take the findings reported here very seriously.

Stephen and his associates used as participants pitchers in four leagues, two for 9- and 10-year-olds and two for 11- and 12-year-olds, from the Birmingham, Alabama metropolitan area during 1997 and 1998. The bulk of the data came from telephone interviews with the pitchers after each of their appearances and from pitch counts provided by their coaches. Forty seven percent of the pitchers reported pain; 32% in the shoulder and 25.5% in the elbow (presumingly, these number sum to more than 47% due to pitchers reporting pain in both areas). In most cases, the pain was mild, but a few of the youths required medical treatment.

Various factors were found to be associated with elbow and shoulder pain. Older and larger pitchers were more susceptible to elbow pain, probably due to the strain of increased weight on their still-underdeveloped bone structure. Reports of elbow pain also increased with number of pitches thrown in a game, particularly as pitch counts exceeded 75. The relationship between elbow pain and pitches over a season was curvilinear; highest for over 600 pitches but lowest between 300 and 600 pitches. Reported arm fatigue and involvement in baseball outside of the league were also associated with elbow pain. Turning to the shoulder, number of pitches in a game and reported arm fatigue were again directly related with pain, but in this case cumulative pitches over the season was inversely related with pain, perhaps due to increased musculature. These data have clear implications regarding limitations that leagues ought to consider placing on the youths under their jurisdiction.

Tricia A. Murray, Timothy D. Cook, Sherry L. Werner, Theodore F. Schlegel, and Richard J. Hawkins, The Effects of Extended Play on Professional Baseball Pitchers, American Journal of Sports Medicine, Volume 29 Number 2, 2001, pages 137-142

A nice companion piece for Lyman et al., sharing a concern with the effect of relatively long pitching appearances on the arm. During the 1998 and 1999 Cactus League, these researchers videotaped pitchers during their windup and release, and compared the motions of seven pitchers between their first and their last (either fifth or sixth) inning of work. Of twelve physical parameters measured, six differed across the innings, all indicating a decrease in physical strain across time, and accompanied by a five-mile-per-hour decrease in fastball velocity. The author(s) are unclear whether these changes are a direct result of fatigue or a protective mechanism to minimize the risk of injury over the course of their appearance.

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

Informal Peer Review

The following committee members have volunteered to be contacted by other members for informal peer review of articles.

Please contact any of our volunteers on an as-needed basis - that is, if you want someone to look over your manuscript in advance, these people are willing. Of course, I'll be doing a bit of that too, but, as much as I'd like to, I don't have time to contact every contributor with detailed comments on their work. (I will get back to you on more serious issues, like if I don't understand part of your method or results.)

If you'd like to be added to the list, send your name, e-mail address, and areas of expertise (don't worry if you don't have any - I certainly don't), and you'll see your name in print next issue.

Expertise in "Statistics" below means "real" statistics, as opposed to baseball statistics - confidence intervals, testing, sampling, and so on.

Member	E-mail	Expertise
Jim Box	im.box@duke.edu	Statistics
Keith Carlson	kcarlson2@mindspring.com	General
Rob Fabrizzio	rfabrizzio@bigfoot.com	Statistics
Larry Grasso	l.grasso@juno.com	Statistics
Tom Hanrahan	HanrahanTJ@navair.navy.mil	Statistics
Keith Karcher	kckarcher@compuserve.com	General
Chris Leach	chrisleach@yahoo.com	General
John Matthew IV	john.matthew@rogers.com	Apostrophes
Duke Rankin	RankinD@montevallo.edu	Statistics
John Stryker	johns@mcfeely.interaccess.com	General
Dick Unruh	runruhjr@dtgnet.com	Proofreading
Steve Wang	scwang@fas.harvard.edu	Statistics

Convention News

The 32nd Annual SABR Convention, SABR32, will take place Wednesday, June 26 through Sunday, June 30, in Boston.

The Statstical Analysis Committee will meet Sunday, June 30, at 10:30am.

Neal Traven reports that the meeting will include a special presentation from Dick Cramer, who will be coming to his first SABR meeting in years. Mr. Cramer's paper deals with the Voros McCracken "balls in play" analysis of pitching, and then relates it to the defensive abilities of the team behind that pitcher.

Don't miss it!

Strength of Opposition a Starting Pitcher Faces: Part II – Historical Data

Rob Wood

In part I of this study last issue, the author provided an alternative statistic to estimate a pitcher's intrinsic winning percentage even in the face of a very small number of starts. Here, in part II, he uses his statistic to estimate the quality of teams faced by a selection of all-time great pitchers.

In part I of this article last issue, I introduced a new statistic to measure the strength of opposition a starting pitcher faced in a season. The statistic reflects the strength of the opposition teams as well as the strength of the opposition starting pitchers.

By using Bayesian inference on pitchers' and teams' won-loss records, I showed that the expected winning percentage of the opposition team (with win pct Q) when the opposition starter has won-loss record W-L is [W+(Q*(T-1))]/(W+L+T-1).[1]

The only remaining variable in the formula is T, a reflection of the weight placed on the team's win pct compared to the individual starter's win pct. I have found that a value for T of 15 is reasonable. Roughly speaking, this implies when a pitcher reaches 15 decisions, our best estimate of the "true" win pct of his team when he starts is the average of his own win pct and the win pct of his team. As the number of his decisions increases, the more faith we put in his own win pct; and the farther below 15, the more faith we put in the team win pct.

This formula, then, gives the strength of the opposition that a starting pitcher faces in each of his games. Taking the average over all of his games gives a good measure of the strength of opposition a starting pitcher faces in a season or in a career.

In this article I want to report the strength of opposition faced by some of the all-time great pitchers. It will turn out that some pitchers faced tougher than expected opposition whereas others faced easier than expected opposition during their careers.

Table 1: Career Averages

	Expected Opp Team WPct	Actual Opp Team WPct	Derived Opp Strength (Bayes)
Whitey Ford	486	493	502
Carl Hubbell	492	506	507
Lefty Grove	488	497	499
Bob Feller	491	497	500
Bob Gibson	496	495	501
Tom Seaver	500	498	505
Pete Alexander*	497	501	501
Roger Clemens**	498	498	502
Walter Johnson*	495	499	499
Warren Spahn	494	495	496
Don Drysdale	494	497	496
Sandy Koufax	492	490	492
Greg Maddux**	496	495	495
Juan Marichal	494	496	491
Randy Johnson**	499	493	488
Average	494	497	498

* Only includes 1920+ seasons ** Through the 2001 season

All-Time Greats

By using the game-logs recently posted on the Retrosheet website, I was able to compile the required information for several all-time great pitchers.[2] I will post the results in a series of tables.[3]

[1] Throughout I use the end-of-season records of teams and pitchers. Also, I do not subtract out the game in question from the opposition starter's won-loss record.

[2] I wish to thank Dave Smith, Tom Ruane, and all the Retrosheet volunteers for making this information available to researchers. At present the game logs do not include the starting pitchers prior to 1920. So I do not have the pre-1920 data for Walter Johnson and Pete Alexander, or any data for Cy Young, Kid Nichols or Christy Mathewson.

[3] I am happy to share my raw data with anyone who is interested.

Table 1 reports the strength of opposition each all-time great pitcher faced throughout his career.[4] The first column presents the expected opposition team win pct averaged over each start in the pitcher's career, taking into account the win pct of the pitcher's own team. Of course, the league as a whole must play 500 ball each season, so a pitcher on a 600 team faces opposition likely to be less than 500.[5] We will use this first column as the baseline to compare to later on.

The second column is the opposition team's actual win pct averaged over each start in the pitcher's career. The third column is the derived Bayes measure of strength of the opposition averaged over each start in the pitcher's career, taking into account both the strength of the opposition team as well as the strength of the opposition starting pitcher. This comes from the "Bayes" formula given in the introduction above. We will see below why the pitchers are sorted in table 1 as they are.

Table 2 presents the key strength of opposition data. The columns of table 2 are ratios of the career averages reported in table 1. The first column of table 2 presents the ratio of the average of the opposition teams' actual win pcts (second column of table 1) to the average of the opposition teams' expected win pcts (first column of table 1). This reflects the strength of the opposition teams faced by the pitcher over his career. A number greater than 1000 indicates that he faced tougher-than-expected opposition teams, and a number less than 1000 indicates easier-than-expected.[6]

Table 2: Career Ratios

	Ratio of Actual Opp Team WPct to Expected Opp Team WPct (reflects strength of opp team)	Ratio of Bayes to Actual Opp Team WPct (reflects strength of opp starters)	Ratio of Bayes to Expected Opp Team WPct (reflects both opp team and opp starters)
Whitey Ford	1014	1018	1033
Carl Hubbell	1029	1003	1031
Lefty Grove	1018	1004	1022
Bob Feller	1013	1005	1018
Bob Gibson	998	1012	1010
Tom Seaver	997	1013	1010
Pete Alexander*	1008	1001	1009
Roger Clemens**	1000	1007	1008
Walter Johnson*	1007	999	1007
Warren Spahn	1001	1003	1005
Don Drysdale	1005	999	1004
Sandy Koufax	995	1005	1000
Greg Maddux**	999	1000	999
Juan Marichal	1003	991	994
Randy Johnson**	988	989	977
Average	1005	1003	1008

The second column of table 2 presents the ratio of the average Bayes measure of the strength of the opposition (third column of table 1) to the average of the opposition teams' actual win pcts (second column of table 1). This reflects the strength of the opposition starters faced by the pitcher over his career, relative to the strength of the opposition teams. A number greater than 1000 indicates that he faced tougher-than-expected opposition starters, and a number less than 1000 indicates easier-than-expected.

The third column of table 3 is the key, presenting the ratio of the average Bayes measure of the strength of the opposition (third column of table 1) to the average of the opposition teams' expected win pcts (first column of table 1). This reflects both the strength of the opposition teams faced by the pitcher over his career as well as the strength of the opposition starters he faced. A number greater than 1000 indicates that he faced tougher-than-expected opposition, and a number less than 1000 indicates easier-than-expected. By construction, the third column is the product of the first two columns.

[4] Juan Marichal and Don Drysdale, not really all-time greats, are included in the study since I was also interested in them.

[5] To derive this expected win pct, I used a simple formula that ignores unbalanced schedules.

[6] For simplicity the decimal points are omitted, so that 1.014 is written as 1014.

You can see that among these all-time greats, Whitey Ford faced the toughest opposition. His opposition teams had a win pct 1.4% higher than expected and his opposition starters had a win pct 1.8% higher than expected (relative to the opposition teams). The combination implies that he faced opposition that was 3.3% tougher than expected during his career. In the next section, we will discuss the implications of facing tough opposition.

Table 3: Miscellany

	Toughest Opposition ever Faced (Bayes)	Easiest Opposition ever Faced (Bayes)	Most Faced in Career
Whitey Ford	766 (Don Mossi 6-1, 1954 Indians)	208 (Don Larsen 3-21, 1954 Orioles)	Billy Pierce (15)
Carl Hubbell	759 (Dizzy Dean 30-7, 1934 Cards)	174 (Ben Cantwell 4-25, 1935 Braves)	Larry French, Lon Warneke (12)
Lefty Grove	752 (Johnny Allen 15-1, 1937 Indians)	213 (Gordon Rhodes 1-8, 1932 Red Sox)	Ted Lyons (15)
Bob Feller	746 (Whitey Ford 9-1, 1950 Yanks)	226 (Lou Knerr 3-16, 1946 Athletics)	Hal Newhouser (16)
Bob Gibson	763 (Sandy Koufax 25-5, 1963 Dodgers)	230 (Roger Craig 5-22, 1963 Mets)	Tom Seaver (11)
Tom Seaver	726 (John Candelaria 20-5, 1977 Pirates)	224 (Tommie Sisk 2-13, 1969 Padres)	Steve Carlton (18)
Pete Alexander*	720 (Dolf Luque 27-8, 1923 Reds)	218 (Les Sweetland 3-15, 1928 Cards)	Burleigh Grimes, Eppa Rixey (13)
Roger Clemens**	764 (Ramiro Mendoza 10-2, 1998 Yanks)	254 (Scott Aldred 0-4, 1996 Tigers)	Jimmy Key (10)
Walter Johnson*	719 (Carl Mays 27-9, 1921 Yanks)	192 (Roy Moore 1-13, 1920 Athletics)	Slim Harriss (12)
Warren Spahn	786 (Preacher Roe 22-3, 1951 Dodgers)	182 (Ron Kline 0-7, 1952 Pirates)	Bob Friend (21)
Don Drysdale	735 (Juan Marichal 25-6, 1966 Giants)	191 (Craig Anderson 3-17, 1962 Mets)	Juan Marichal (16)
Sandy Koufax	749 (Bob Purkey 23-5, 1962 Reds)	167 (Bob Miller 1-12, 1962 Mets)	Don Cardwell (10)
Greg Maddux**	753 (Randy Johnson 10-1, 1998 Astros)	243 (Kyle Abbott 1-14, 1992 Phillies)	Andy Benes, Doug Drabek (10)
Juan Marichal	763 (Sandy Koufax 25-5, 1963 Dodgers)	167 (Bob Miller 1-12, 1962 Mets)	Don Drysdale (16)
Randy Johnson**	733 (Jimmy Key 17-4, 1994 Yanks)	238 (Jim Abbott 2-18, 1996 Angels)	Kevin Brown (7)
Average	752	205	13.5

As you probably know, Casey Stengel frequently saved Ford for the Yankees' toughest foes. Indeed, Ford's strength of opposition under Stengel is even tougher than his entire career figures shown in the table. The relevant ratios are 1023, 1022, and 1046, implying that Ford faced opposition that was 4.6% tougher than expected, when Stengel was his manager.[7] R.J. Lesch is doing further research into how Stengel used Ford and his other starting pitchers.

Carl Hubbell is another pitcher who faced tougher-than-expected opposition. In King Carl's case, it was almost entirely due to which teams he faced, not which starters he faced on those teams. Of Hubbell's 433 career starts, 75 were versus the St. Louis Cardinals, generally the Giants' toughest foe, whereas only 49 were versus the lowly Boston Braves. All things considered, Hubbell faced opposition that was 3.1% tougher-than-expected.

Lefty Grove and Bob Feller are the next two all-time greats on the list. They faced opposition that was 2.2% and 1.8% tougher-than-expected, respectively.

I was somewhat surprised to see that Warren Spahn checked in as having faced nearly average opposition. As you know, Spahn (a lefty) rarely faced the Dodgers at Ebbets field (a predominant right-handed hitting team in a friendly park). In fact, Spahn started exactly one game against the Dodgers from 1954-1957 (June 5, 1956 in Milwaukee). I anticipated that "ducking" the tough Dodgers would sway Spahn's opposition towards the easy end of the scale. All things taken into account though, Spahn faced about average opposition throughout his career; the tougher opposition he

[7] I conjecture that the pitcher's own manager has more control over what opposition teams he faces, whereas the opposition manager has more control over what opposition starter he faces.

faced in other years compensated for the easier opposition he faced in 1954-1957.

Among the 15 pitchers for whom I have performed these analyses, only Randy Johnson faced significantly easier-than-expected opposition over his career. The Big Unit has benefited from both easier-than-expected opposition teams (1.2% easier) as well as easier-than-expected opposition starters on those teams (1.1% easier). The combination implies that Johnson faced opposition 2.3% easier-than-expected thus far in his career.

The second factor may be a reflection of the opposition manager "sacrificing" one of his lesser starters facing Johnson, games in which his team is apt to lose in any event. Although I find this hypothesis attractive, if true I wonder why it does not seem to have manifest in the case of the other all-time greats. One possibility is that today's fairly loose five-man rotations make it easier for opposing managers to swap their starters to get matchups they prefer (sacrificing) whereas yesteryear's four-man rotations were too rigid for this.

Table 3 presents some miscellaneous information I compiled. The first two columns report the toughest and easiest opposition the pitcher ever faced in his career, based upon my new Bayesian formula. The toughest of them all was the opposition that Warren Spahn faced on May 12, 1951; he was facing Preacher Roe of the Brooklyn Dodgers. Roe wound up going 22-3 that year for the 97-60 Dodgers (96-58 in the regular season). The easiest opposition that any of these pitchers ever faced is deemed to have been the woeful 1-12 Robert L. Miller, the right-hander of the two Bob Millers on the 1962 New York Mets. It turns out that Sandy Koufax and Juan Marichal both faced Miller that year, so he shows up on both of their lists. The table also reports the starting pitcher(s) that each pitcher faced the most times during his career.

Table 4: Seasonal standard deviations (Averages)

	Expected Opp Team WPct	Actual Opp Team WPct	Derived Opp Strength (Bayes)
Whitey Ford	8	16	19
Carl Hubbell	7	20	20
Lefty Grove	10	19	24
Bob Feller	8	17	12
Bob Gibson	5	12	18
Tom Seaver	6	9	16
Pete Alexander*	10	15	14
Roger Clemens**	4	12	17
Walter Johnson*	10	10	7
Warren Spahn	8	17	24
Don Drysdale	7	12	16
Sandy Koufax	6	9	16
Greg Maddux**	5	9	13
Juan Marichal	4	10	17
Randy Johnson**	4	6	12
Average	7	13	16

Table 4 reports the standard deviation of each pitcher's seasonal measures of strength of opposition, for which the career averages were reported in table 1.[8] I report the seasonal standard deviations for a couple of reasons. First, it can help calibrate readers to better interpret these new stats. Second, sometimes career averages do not tell the whole story.

Whitey Ford's entry shows that his overall average strength of opposition varied by about 19 points each season (19 points compared to a career average of 502). In fact, Ford's seasonal opposition strengths ranged from 478-542, so you can see that there is a fair amount of variability in this stat from year to year. The other pitchers exhibit roughly the same amount of seasonal variability.

Table 5 reports the standard deviation of each pitcher's seasonal measures of strength of opposition, for which the career averages were reported in table 2. Whitey Ford's entry shows that his overall ratio of the strength of opposition varied by about 43 points each season (43 points compared to a career average of 1033). In fact, Ford's seasonal opposition strength ratios ranged from 981 (in 1964) to 1140 (in 1954), so you can see that there is a fair amount of variability in this stat from year to year. The other pitchers exhibit a little less than Ford's seasonal variability – possibly due to how Stengel utilized Ford.

[8] These standard deviations are taken over only the full seasons for each pitcher.

Impact of Opposition

Now that we have a measure of the strength of the opposition each of these all-time greats faced in their careers, what do we do with it? First let's see if there is evidence that a pitcher's stats are affected by my new strength of opposition stats. Of course, we would expect that if the pitcher faced easier opposition, his winning pct will be higher and his ERA will be lower, and vice versa if he faced tough opposition.

Such is indeed the case. Using each pitcher's seasonal data, I find that the correlation across all pitchers between their ERA+ and the Bayes strength of opposition is -0.24.[9] The tougher the opposition (the higher is Bayes), the lower the pitcher's ERA+. Similarly, the correlation across all pitchers between their seasonal win pct and the Bayes strength of opposition is -0.32; the tougher the opposition, the lower the pitcher's win pct.

Standard sabermetric formulas consider a pitcher's seasonal performances as a whole (e.g., his seasonal ERA relative to league average). No account is typically made of the strength of the opposition that the pitcher faced during the season.

Table 5: Seasonal standard deviations (Ratios)

	Ratio of Actual Opp Team WPct to Expected Opp Team WPct (reflects strength of opp team)	Ratio of Bayes to Actual Opp Team WPct (reflects strength of opp starters)	Ratio of Bayes to Expected Opp Team WPct (reflects both opp team and opp starters)
Whitey Ford	37	18	43
Carl Hubbell	34	19	34
Lefty Grove	27	21	39
Bob Feller	32	26	30
Bob Gibson	25	29	38
Tom Seaver	15	25	30
Pete Alexander*	23	7	19
Roger Clemens**	22	20	33
Walter Johnson*	16	19	25
Warren Spahn	26	26	39
Don Drysdale	15	27	31
Sandy Koufax	10	15	23
Greg Maddux**	14	24	23
Juan Marichal	22	22	36
Randy Johnson**	12	20	20
Average	22	21	31

Suppose I told you that Whitey Ford went 19-6 with a league-leading 2.47 ERA in 1956. You would say that he had a great season. Now suppose I told you that he faced very tough opposition that season (7.6% tougher-than-expected). Wouldn't you elevate your view of Ford's season, at least a little bit? In this section I hope to derive some rules-of-thumb for the utilization of these strength of opposition figures.

In order to try to quantify the relationship, I performed a couple of regressions. The first regression simply regresses the pitchers' seasonal ERA+ on his seasonal Bayes strength of opposition. I find that the coefficient on Bayes is -0.45.[10] This implies that a 22 point increase in the strength of opposition, say from 500 to 522, would lower the pitcher's ERA+ by 10 points, say from 120 to 110. This is a rather large impact, more than I would have expected.[11]

The next regression I performed was on a pitcher's seasonal win pct. I regressed this on each pitcher's own team's win pct, the pitcher's seasonal ERA+, and the Bayes strength of opposition he faced during the season. The coefficients are 0.69 on the team win pct, 1.90 on the pitcher's ERA+, and -0.52 on the Bayes measure.[12] This implies that a 19 point increase in the strength of opposition, say from 500 to 519, would lower the pitcher's win pct by 10 points, say from 600 to 590. This is smaller than I would have expected.[13] Clearly, more research has to be done in this area, and I welcome all comments or ideas.

[9] ERA+ is Total Baseball's measure of the pitcher's ERA relative to his league, adjusted for his home park. A low ERA corresponds to a high ERA+.

[10] Standard error of the coefficient is 0.13.

[11] The Pythagorean relationship between runs scored and win pcts yields an estimate that a pitcher's ERA would be affected by roughly half of the impact on his win pct, since only half the change in win pct is likely due to the opposition team's offense (the other half due to its pitching/defense). For example, if a pitcher faced 10% tougher than expected opposition (in terms of win pcts), then his ERA would be expected to have been 5% higher than otherwise.

[12] The standard errors of the coefficients are 0.10, 0.18, and 0.35, respectively.

[13] Using the "log-5" method of estimating the win pct when two teams face off, we would expect the coefficient to be around -0.90 for this set of pitchers (and around -1.00 for all pitchers as a group).

Concluding Remarks

In Part I of this study, I introduced a new stat that measures the strength of opposition a starting pitcher faces over the course of a season or a career. In Part II, I have reported the stat for several of the greatest pitchers of all-time. We have seen that over the course of most pitchers' careers, the strength of opposition "evens out" so that standard sabermetric methods are likely to give accurate measures of the pitcher's true value to his team.

We have seen that there were a few pitchers, however, whose strength of opposition was significantly different than expected over the course of their careers. Most notably, Whitey Ford, Carl Hubbell, Lefty Grove, and Bob Feller faced tougher-than-expected opposition. These pitchers deserve even more credit than the standard sabermetric methods indicate. On the other hand, only Randy Johnson appears to have faced significantly easier-than-expected opposition over the course of his career to date.

I have demonstrated that my measure of strength of opposition does correlate with a pitcher's ERA and win pct, though I am not yet satisfied that we have a good handle on the exact relationship. More work is needed in this area.

Rob Wood, 2101 California St. #224, Mountain View, CA, 94040-1686, robert_wood@standardandpoors.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 (but no death threats, please) are all welcome.

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

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

I usually edit for spelling and grammar. (But if you want to make my life a bit easier: please, use two spaces after the period in a sentence. Everything else is pretty easy to fix.)

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

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

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

Send submissions to:
Phil Birnbaum
18 Deerfield Dr. #608, Nepean, Ontario, Canada, K2G 4L1
birnbaum@sympatico.ca

Week-to-Week Consistency in Individual Offensive Performance

Charlie Pavitt

Sportswriters and broadcasters will often speak of a hitter having a "hot bat," meaning that he's on a hot streak and more likely, therefore, to continue performing well. Other sportscasters will suggest that a player on a cold streak is "due," being more likely to perform well to make up for a previous slump. Do either of these effects exist? Here, the author sifts through many years of play-by-play data to see if there is any evidence for either effect.

How consistent is offensive performance? Before we can try to answer this question, we must refine it. We need to indicate whether we are interested in consistency across seasons, across months or weeks, or day by day. Across seasons, we know that offense over a major league career can be approximated by a curve with its greatest height for most players between the ages of 25 and 29, and with a standard deviation of approximately 25 points in batting average for individual seasons. Turning to consistency across months, Jim Albert found no evidence for significant differences in performance comparisons between the first and second halves of the season as reported in STATS publications (see Albert and Jay Bennett's book *Curve Ball*, Chapter 4, and Albert's article in the *Journal of the American Statistical Association*, 1994, volume 89, pp. 1066-1074).

As for more short-term consistency, we can phrase the question as follows: are batting streaks and slumps real phenomena or random fluctuations? In his 1986 *Baseball Abstract* (pages 230-231), Bill James reported on a small 1985 study of Astros hitters by Steven Copley, who compared performance the game after a "good game" (2 for 6 or better) with performance after a "bad game" (an oh-fer), and found a slight improvement after good games (.280 versus .268). James claimed that evidence here for any correlation across games is "very questionable." James was right; I performed a chi-square test on that data and found it to equal .37, which translates to over eighty percent odds that random data would produce a similar result. In the 1987 *Elias Baseball Analyst* (pp. 97-99), Siwoff and the Hirdt brothers showed several indicators supportive of the random fluctuation hypothesis; that players riding a five-game hot streak do not perform better in the next game than players after a five-game cold streak, that players who are streaky one season are just as likely as not to be steady the next. In the 1989 Elias Analyst (page 164), they noted a slight tendency (translating to .015 in batting average improvement) for players to get a hit following an at bat producing a hit than following an at bat producing an out. They did not present data allowing a test for statistical significance. This tendency could easily be due to consistencies across at bats in pitching quality (i.e., two consecutive at bats against the same poor pitching), ballpark, or some analogous situational factor independent of a batter's skill.

The same question led to a debate in the pages of the Journal of the American Statistical Association (1993, volume 88, pages 1175-1196). S. Christian Albright did a number of analyses along the lines of, but more sophisticated than that reported by the Elias group. The odds of getting on-base after a successful at bat versus an unsuccessful at bat varied among the players in his data set closely to what would be expected from random fluctuation, and streakiness among players was uncorrelated across seasons. Respondent Jim Albert, using the same data, also found variation among players to approximate random fluctuation, although he correctly noted that these fluctuations could still signal a real effect. Respondents Hal S. Stern and Carl N. Morris criticized Albright's procedures for absence of statistical power (in other words, they are not precise enough to find any real effects existing in the data) and for bias toward stability. They concluded that they remained convinced that streaks and slumps are real, but present no evidence other than their experience playing and watching sports. More recently, Albert has described methods for detecting streakiness and consistency in two outlets; Chapter 5 in Albert and Bennett's *Curve Ball*, and an article by Albert and Patricia Williamson (*The American Statistician*, 2001, volume 55, pp. 41-50; many thanks to Chris Kenaszchuk for alerting me to this article). To the best of my knowledge, Albert has not yet used these methods to perform a study of performance comparable to Albright's.

To conclude; Albert could well be correct that what appears to be random fluctuation may be a real phenomenon, but the existing evidence is overwhelmingly in support of the claim that streaks and slumps in offensive performance are random fluctuations. Given this evidence, why continue studying the issue? To be honest, I began collecting data for this study in 1991, well before the Albright study, and if I had known then what I think I know now, I probably would not have begun it. But I remain convinced that there may be more to learn about the issue. In short, I now have 11 years (1991 to 2001) of week-by-week performance relevant to batting and slugging averages, and the following is a report of some preliminary analyses of that data with the hopes of addressing the question of offensive consistency across weeks.

First, let me describe the data set. It consists of players who had at least 10 Sunday-through-Saturday weeks in which they registered 10 or more at bats (not plate appearances), along with their hits and total bases for that week, over at least four seasons, which I gathered from the

Sporting News and USA Today (except for a missing four-week stretch for which Dave Smith supplied me data; a SABR Salute to him). Although I was not always consistent in doing so, I tended to cut out stretches in which players did not play those weeks approximately consecutively; so, for example, if Joe Schmo got 10 or more ABs in the first two weeks and then (perhaps due to injury) not again until weeks 8 through 26, I tended to cut out the first two and only include the rest. As an exception to the 10-at-bat rule, I also included a few platooning catchers (Kurt Manwaring is an example that pops to mind) who consistently got 7 to 13 or so ABs per week over a space of several seasons.

The goal of the data analyses was to examine week-to-week fluctuation in batting average and slugging average as measures of offensive performance. The ideal way to perform this analysis is to look at fluctuations across the entire career of a player, as this provides the biggest sample size and thus the greatest chance of finding patterns in these fluctuations. One problem that immediately suggested itself was the predictable change in performance during a career. If, as is normal, a player performs best during the "middle years," and one uses average weekly performance for the entire career as the criteria for judging fluctuation, then the first and last years will be consistently below average and the middle years consistently above. This would lead to artifactual findings of patterns that are irrelevant to the question at hand. Another problem is changes caused by differences in playing conditions due to changing teams and new ballparks. After discussing this issue with the attendees at one of the meetings of the Washington, D. C. area "Risks and Rewards" sabermetrics discussion group, I decided that I would be better off examining consistency at both the annual and career levels. The former would escape this problem but yield small sample sizes; the latter would of course have this problem but would include larger sample sizes. In this preliminary report, I will only describe findings at the annual level.

I also decided to conduct the analysis two ways; through Wald-Wolfowitz runs tests and regression-based time series analysis. A "run" in a data set is a stretch of one or more consecutive data points all of which are either above or below the average of that data set. In a runs test (described in W. J. Conover's Practical Nonparametric Statistics, pages 349-356 and the various editions of William Hays's Statistics for the Social Sciences; in the 2nd edition it can be found on pages 775-777), the number of "runs" of scores both above and below a criterion, usually the median, is transformed into a z-score. A statistically significant positive z-score, or more runs than chance would predict (for example, a pattern of coin flips such as HTHTHTHTHT, which contains 10 runs) would indicate a circumstance in which good and bad weeks alternated non-randomly, reminiscent of the stereotypical sportscaster belief that after a bad week a player is "due" for a good one. A statistically significant negative z-score, or fewer runs than by chance (a pattern of coin flips such as HHHHHTTTTT, with only two runs) would indicate a pattern of long streaks and slumps. The regression-based time series (see the book on this subject by Charles W. Ostrom) begins with the computation of the regression equation, which defines a line running through the "center" of the data that represents expected performance across weeks. This analysis in and of itself allows a secondary study of whether performance increased or decreased across weeks, and I did an additional examination of this issue. Of primary interest, however, is a statistic called the Durbin-Watson test, which examines consistency of performance above and below the regression line across weeks. The Durbin-Watson test results in an index that ranges from 0 to 4. A large index indicates more runs than chance would allow, and a small index indicates fewer runs than by chance. Unfortunately, at least at the time that Ostrom's book was published (1978), the exact points indicating statistical significance was uncertain, with both conservative and liberal critical values. In this essay, I report indices beyond the conservative levels as significant and indices beyond the liberal levels as questionable.

Table 1 – Runs Tests

	Batting Average		Slugging Percentage	
Significance level	.05	.10	.05	.10
Number of z's	549	549	549	549
Significant z's by chance	27	55	27	55
Significant z's in data	17	33	14	26
Significant z's indicating inconsistency	8	15	6	15
Significant z's indicating consistency	9	18	8	11

The time series analysis is the "better" procedure of the two, in the sense that it is more sensitive to existing patterns, but technically it should be limited to data with no missing values. However, there are missing weeks in this analysis, such that some statisticians would disapprove of my using it. The runs test does not have this limitation. Another difference between the two is that, by using the median as the basis for comparison, uses of the runs test implicitly presume that average performance remains at that level across a season; in contrast, by using the regression line as the basis for comparison, the Durbin-Watson test does not. This can lead to differing conclusions based on each method. For an extreme example, if Phil Thrill has a terrible first week, slightly better second week, even better third week, and continues to perform a little better every successive week for the rest of the reason, then the runs test will note the improvement as unusual consistency (the first half of the season below median, the second half above), whereas the time series analysis will find a significant positive regression line but neither consistency nor inconsistency in the Durbin-Watson test.

The data for the present study includes 93 players whose careers, to the best of my knowledge, have ended (please contact me if you are curious about who was included). Further, the table I had for examining the significance of the Durbin-Watson index (from Jan Kmenta's Elements of Econometrics, page 625) included critical values for sample sizes of 15 through 100. As a consequence, I limited annual analyses to seasons with 15 weeks for which I had data. Across the 93 players, I ended up with a sample size of 549 seasons. (Incidentally, one reason I do not look at data across seasons in the present analysis is that I do not know what the critical values are for careers lasting over 100 weeks; if anyone reading this is knows how to find out, please let me know.) The results were as follows:

Table 2 – Data for Durbin-Watson Indices

	Batting Average significant	Batting average significant & questionable	Slugging Average significant	SLG significant & questionable
Total number	549	549	549	549
# unusually large	25	72	22	65
# unusually small	10	30	10	37

The runs tests showed absolutely no systematic tendencies for either consistency or inconsistency. For both batting average and slugging average, there were fewer statistically significant z's than would be expected by chance, and no observable tendency for statistically significant z's to indicate either unusual consistency or unusual inconsistency (see Table 1 for relevant data). Further, the mean z for BA was -.006 and the median z was 0, implying a symmetrical distribution and thus no overall tendency for consistency or inconsistency across the 549 seasons. Similarly, the distribution of z's for SA was symmetrical, with a mean z of -.002 and a median z of 0. The Durbin-Watson test leads to a slightly different conclusion, with more unusually large indices than unusually small indices, showing some evidence that players are more inconsistent than chance would allow (see Table 2 for relevant data). Further, for batting average, the mean across all indices was 2.10 and the median was 2.08. Although 2.10 might not appear much different than 2, the difference is clearly statistically significant (one-sample t equals 5.03, which is significant at the .001 level). In the case of slugging average, the mean and median were both 2.09, with the mean again significantly different from 2 (one-sample t equals 4.26, which is significant at the .001 level).

Table 3 – Data for Overall Performance Change

	Batting Average		Slugging Average	
Significance Level	.05	.10	.05	.10
Number of equations	549	549	549	549
Number significant by chance	27	55	27	55
Number significant in data	24	50	25	55
Number significant indicating decline	16	30	13	35
Number significant indicating improvement	8	20	12	20

The tendency for inconsistency across weeks could possibly be due to the alternation between home stands and road trips. If a player's home park was either particularly conducive or particularly detrimental to hitting, then there would of course be predictable BA and SA discrepancies between home and away games. If home stands and road trips lasted approximately one week, then these could translate into alternate good and bad weeks in offensive performance. Although most home stands and road trips are not one week in length, enough may approximate this time period to produce the cross-week inconsistency in this data. I believe this explanation to be at least feasible, and look forward to any alternative suggestions from interested readers.

Finally, there were no significant tendencies for overall improvement or decline in performance across the weeks. Having said this, however, there was a tendency for those equations that were significant to be negative, indicating declining performance across the year (see Table 3 for relevant data). This could possibly be the type of case that Albert has warned us about, in which data that look to be random actually are not. One possible explanation, which I am fairly confident is correct, is that borderline starting players who start slowly will often end up on the bench and thus not have the opportunity to improve on their performance over the year, whereas borderline starting players who start out quickly will often maintain their position even as their performance deteriorates over the year. As a result, there is more opportunity for significant decreases in offensive performance than there is for significant increases.

These findings provide at least some warrant to search for specific players who appear to have been either particularly consistent or inconsistent across weeks. There did seem to be a few, although not many. John Jaha's Durbin-Watsons were large, at least in the questionable range for batting average every year between 1993 and 1996 and for slugging average between 1994 and 1996, indicating unusual inconsistency within seasons. It was more difficult to find a player clearly exhibiting consistency; perhaps not surprisingly, Tony Gwynn was one of the closest over his last qualifying five seasons (1995 through 1999). For batting average, his Durbin-Watson's during that stretch were never larger than 1.77, and 3 of the 5 reached the questionable range; for slugging average, all but one were 1.74 or less and 2 reached the questionable range. Although there is very little warrant in looking any further at overall performance across the weeks, there were a couple of players who seemed to be consistently inconsistent. Lance Johnson's slugging average increased at a rate significant at .10 during 3 of 8 seasons for which I had data (1991, 1995, and 1996), which would be roughly equivalent to an average increase of something in the vicinity of 10 SA points per week from beginning to end during those seasons. Joe Carter's batting average decreased at a similar rate for 3 of 8 seasons (1991, 1996, and 1997), which is somewhere around an average decrease of 5 points per week.

In conclusion, although there is some evidence in this data that individual offensive performance across a season may not be totally random, that evidence is far from conclusive. In the future, I plan to repeat these analyses with a larger sample size (including players still active). Further, if I find out how to evaluate Durbin-Watsons for sample sizes over 100, I would like to look at data across seasons, the problems in doing so notwithstanding, to see if the conclusions from this analysis still hold.

The author would like to thank our editor, Phil Birnbaum, for suggestions than notably improved this paper. Charlie Pavitt, 812 Carter Road, Rockville, MD, 20852, chazzq@udel.edu ♦

Receive BTN by E-mail

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

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

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