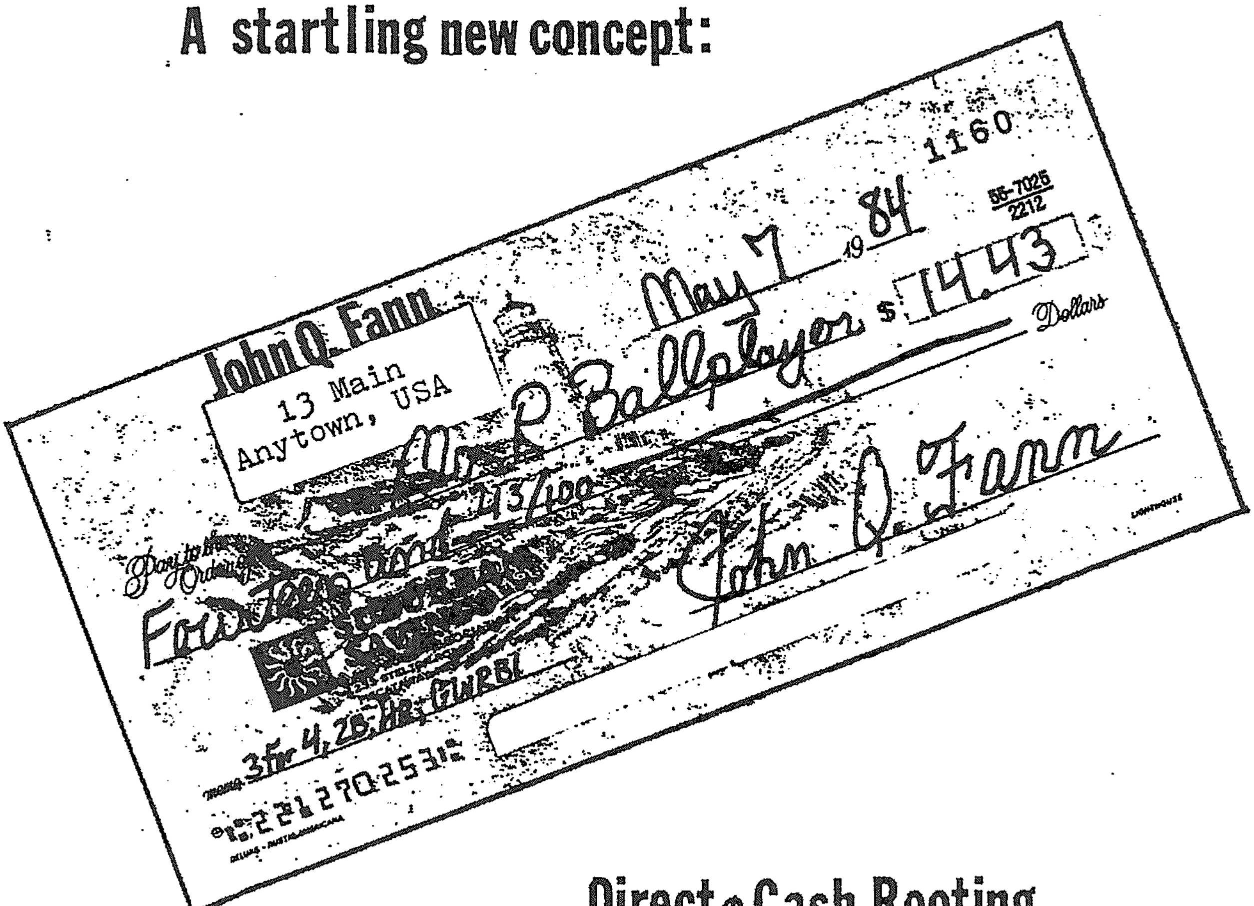
Baseball Analyst

December 1983 | Issue #9

ANNOUNCING!

A startling new concept:



Direct$ Cash Rooting

Editor's Note:

First of all, sorry we're late with this issue, but December is Baseball Abstract month around here, so you can imagine how hectic things have been. But look at it this way: the February issue will be in your mail box in no time. There's lots of good news to report. Our subscribers have doubled in number and the articles are coming in quite rapidly. My thanks to Cliff Kachline at SABR for running the announcement about us in a recent bulletin. It really helped make more people aware of what it is we're doing here. I'd like to return the favor by plugging SABR, but I'm sure you're all members. However, I can remind you to re-up with SABR, as it's that time of year again. Also you could get a friend interested in joining, then you'd have someone to go to the convention with, right?

There is other great news, and that is that Pete Palmer (along with John Thorn) has a book coming out in April. It's called The Hidden Game of Baseball (Doubleday, hardcover), and it should hit the bookstores in time for the start of the '84 season.

Pete is a frequent contributor to the Analyst and has done a lot to advance the study of sabermetrics. I say without reservation that anyone who takes his baseball seriously should absolutely have this book. If ever there were a more appropriate audience for a book, I can't think of one. Hidden Game is right up your alley. It contains a history of sabermetrics as well as a detailed review of baseball history through its rule changes and their effects on the game. Pete studies the 19th Century at length, something few (if any) sabermetricians ever bother to do. There is year by year statistical analysis starting in 1876 and including the season just past. The relative merits of various records are also charted at great length. At last we have a single source for all of this information. I hope you'll add it to your libraries. --Jim Baker

World Headquarters
945 Kentucky Street
Lawrence, KS 66044

$12.00 per annum

All Submissions ready for reproduction, please

baseball ANALYST baseball

December '83

Founder/Publisher: Bill James
Editor: Jim Baker
Business Manager: Susie McCarthy

Issue no.9

IN THIS ISSUE:

COVER STORY: Reader Fann has proposed a new concept in major league team management. Fann hates owners and all they stand for, and would like to see the patrons of the sport take a more direct approach in their involvement with the game. Fann suggests that instead of buying tickets to individual games, loyal rooters merely pay the players on a 'per-performance' basis. The sample payment on the cover is a pretty good one, but then 'Mr. P. Ballplayer' had a pretty good night, and deserved tobe payed off handsomely. Fann sees this as a way to eradicate 'greedy baseball owners' and inspire baseballers to give 100% effort at all times. He sadly reports that his efforts to reach the league presidents and commisioner's office have gone for naught, but he remains persistent and hopeful. We gladly give him this space to air his views.

THE BEST FIELDING SECOND BASEMEN SINCE 1925

Dan Finkle

The Fielding Index (FI) measures skills of fielders using fielding statistics available in the Baseball Guide -- fielding games, putouts, assists, double plays, and fielding percentage. These statistics are modified by the strikeouts made by each fielder's pitching staff and by a measure of the staff's effectiveness. The statistics developed for the individual players are compared to the same statistics for the entire league. As a result the FI is a measure of performance of players against the conditions and norms for all other players of the same league. In this way performances from one year to the next can reliably be compared.

The FI for second basemen is a combination of three factor indices: The Range Index, the Double Play Index, and the Misplay Index. I have calculated the FI for all second base regulars in the major leagues from 1925 to 1982. "Second base regulars" is defined as players who have appeared in at least 100 games at second base during the season and who are not simply late inning replacements. For those players who have been regulars at least five years I have compiled records to determine the best fielding second basemen.

"Best" can be tested in a variety of ways. Let's take a look at some of the possibilities. One possible measure is outstanding performance over a short period of years. Bill James, in developing the concept of the defensive spectrum, has shown that certain positions make larger defensive demands. On his spectrum shortstop is the most demanding defensive position, second base next. Defensive skills, however, tend to disappear more quickly than offensive so an outstanding fielder may have a shorter career at the defensively demanding position. If this argument is valid, a fair measure for best fielding second basemen would be over a short period of years.

I have selected the second base regulars with the highest average FI for four consecutive years. (Not necessarily four consecutive calendar years, but four consecutive years when the player performed as a second base regular.)

BEST FIELDING SECOND BASEMEN FOUR CONSECUTIVE YEARS		
Mazeroski	1963 - 66	1.204
Schoendienst	1951 - 54	1.132
Critz	1930 - 34	1,128
Melillo	1930 - 33	1.110
Frey	1940 - 43	1.104

Bill Mazersoki not only appears as number one, but his average FI is far above any other. He is in a class by himself.

From 1925 to 1982 fifty-one players have been second base regulars for five or more years. The average tenure for these is 7.8 years. (Currently active second basemen are excluded from this calculation to avoid a downward bias.) Another reasonable test for "best" might be: Who stands out during the average period of a second basemen's career?

BEST FIELDING SECOND BASEMEN EIGHT CONSECUTIVE YEARS		
Mazeroski	1960 - 67	1.165
Critz	1926 - 34	1.109
Schoendienst	1946 - 54	1.101
Doerr	1942 - 50	1.088
Herman	1933 - 40	1.080

Again, Bill Mazeroski is not only best but far the best. Hugh Critz and Red Schoendienst have exchanged places but the differences are small. Oscar Melillo has dropped from fourth in the four year list to ninth in the eight year list. Lonnie Frey had only six years as a second base regular.

Some may reason that real excellence is only established when the player demonstrates it over a long period. Certainly long time stardom has proven to be a major factor in the selection of players to the Hall of Fame. In fact, few players spend most of their careers at the positions with high defensive demands. If they also possess strong offensive talents they tend, as Bill James has demcnstrated, to move across the spectrum to less demanding defensive positions. If, therefore, a player spends many years as a second base regular, a prima facie case may be made that his fielding skills are exceptional. In the period 1925 - 1982 only

seven players have performed as many as twelve years as second base regulars. Incidentally, the player with the most years as a second base regular is Joe Morgan at 17 and still counting.

BEST FIELDING SECOND BASEMEN TWELVE CONSECUTIVE YEARS

Mazeroski	1958 - 70	1.139
Doerr	1938 - 50	1.082
Schoendienst	1946 - 58	1.078
Gehringer	1927 - 39	1.058
Herman	1932 - 43	1.044

The question of the greatest fielding second baseman is settled. Bill Mazeroski once again stands out by a large margin. One name from the previous list, Hugh Critz, drops out because he had nine years from 1925 to 1934 as a second base regular. The replacement is Charlie Gehringer. The order once again changes.

In an attempt to find some objective measure of best fielder, I devised an arbitrary formula that gives greatest weight to the eight year average FI, equal weight to the four and twelve year average FI, and proportionate reduction for less than twelve years as a second base regular. Using this device the five best fielding second basemen since 1925 are:

BEST FIELDING SECOND BASEMEN

Mazeroski	1.168
Schoendienst	1.103
Doerr	1.090
Gehringer	1.078
Herman	1.076

Another interesting way to examine the skills of the best fielding second basemen is to look at their factor indices: The Range Index, the Double Play Index, and the Misplay Index. Calculation of the FI requires that each of these factors be assigned a weight in proportion to its perceived importance. The weights cannot be assigned in any completely objective manner, so a look at the factors may give some new insights on the best fielders.

The standouts in fielding range are:

SECOND BASEMEN WITH BEST RANGE INDEX

Four Consecutive Years		Eight Consecutive Years	
Mazeroski	.574	Mazeroski	.553
Trillo	.544	Critz	.526
Critz	.539	Herman	.521
Melillo	.536	Gordon	.520
Herman	.533	Trillo	.520

As expected, Bill Mazeroski once again leads and dominates. One name is surprisingly missing. Red Schoendienst. He is eighth on the four year list, tenth on the eight year list. Two new names appear. Manny Trillo, the first active player to have appeared on any of our lists, makes a strong showing as one of the greatest second basemen in fielding range. Trillo has not made a stronger showing in the FI because his Misplay Index is not outstanding and his Double Play Index is less than average. The other new name is Joe Gordon.

The standouts in the Double Play Index are:

SECOND BASEMEN WITH BEST DOUBLE PLAY INDEX

Four Consecutive Years		Eight consecutive Years	
Mazeroski	.289	Mazeroski	.277
Gordon	.275	Gordon	.256
Herman	.261	Critz	.244
Critz	.256	Herman	.242
Cash	.247	Schoendienst	.238

Mazeroski, of course. Joe Gordon emerges as a standout in turning the double play. Although he had not appeared on any of the lists for best FI, he was number six in the list of best fielding second basemen. Observers who would give greater emphasis to double plays than I have would certainly include Joe Gordon among the five best fielding second basemen. His consistently low fielding averages have kept him from a higher FI. Missing from both the lists of Range Index and Double Play Index are Bobby Doerr and Charlie Gehringer. Their strengths are more in low misplays, as you will see, and in their longevity and consistency.

The standouts in the Misplay Index are:

SECOND BASEMEN WITH BEST MISPLAY INDEX

Four Consecutive Years		Eight Consecutive Years	
Schoendienst	.375	Schoendienst	.359
Robinson	.367	Bishop	.356
Doerr	.363	Doerr	.353
Frey	.358	Gehringer	.349
Bishop	.357	Frisch	.345

The high rating of Red Schoendienst as a fielder leans more heavily on his consistently high fielding average than on the other factors although he did not do poorly on any measure. Two interesting new names appear on this list. Both might have received much higher ratings as all time greats of fielding at second base but for different reasons. Jackie Robinson had only five years as a second base regular. If he had been able to come up to the majors when he was younger; if he had not played under such extrodinary pressure . . .! there are many ifs in his great career. Frankie Frisch began his major league career in 1920. Five years are cut off because this analysis begins only in 1925. His rating very possibly would be among the top five if those years were considered. Another new name is Max Bishop of the great A's teams of the thirties who stands out as one of the most accurate of all fielders.

Bill Mazeroski's name is missing for the first time. No stronger argument could be made for the irrelevancy of the emphasis on fielding average as a primary measure of fielding skill. There is no question but that Bill Mazeroski is the greatest fielding second basemen from 1925 to 1982. Yet the conventional measure would not have even ranked him in the top five.

Bill Mazeroski is eligible for the Hall of Fame. In the 1982 voting he received just 48 votes, seventeenth on the list. Maybe we shouldn't be surprised, because Bill Mazeroski is best remembered, not for his remarkable fielding ability, but for one hit, his winning home run in the last half of the ninth inning of the seventh game of the 1960 World Series.

continued on page 19

THE WORST

by Joe Ferrere

A PREFATORY NOTE SHOULD WARN THE READER THAT THIS ESSAY CONTAINS ONLY TWO DECIMALS! ALL OTHER SIMILARLY APPEARING DOTS ARE, IN FACT, PERIODS.

HOW MANY TIMES HAVE YOU HEARD AN ANNOUNCER REFER TO A BALLPLAYER AS "...ONE OF THE BEST...."? THAT STATEMENT IMPLIES--THOUGH THEY ARE NEVER MENTIONED--THAT PLAYERS DO EXIST AT THE OTHER END OF THE SCALE. I HAVE STUDIED THESE 'LOWER DEPTHS' AND HAVE SELECTED FROM AMONG THEM MY CANDIDATE FOR **WORST** BALLPLAYER. IN SO DOING I INVITE THE READER TO CONSIDER OTHERS WHO MAY EXCEED MY MAN'S NON-ACHIEVEMENTS. MY OWN CREDENTIALS IN THIS PARTICULAR AREA ARE IMPECCABLE: I WAS THE PERENNIAL 10th MAN ON ALL BASEBALL TEAMS I WAS ALLOWED TO JOIN (pre-DH). MY STATUS DID EVENTUALLY CHANGE WITH A LATE CAREER SWITCH TO SOFTBALL WITH ITS SHORT-FIELD POSITION; I WAS THEN DROPPED TO 11.

NOW IT IS DIFFICULT TO BE TRULY BAD IN THE BASEBALL BUSINESS BECAUSE THEY DON'T ALLOW YOU TO STAY AROUND LONG ENOUGH TO DEVELOP THE NUMBERS TO QUALIFY FOR CONSIDERATION. THE CRITERION OF POOR PERFORMANCE OVER AN EXTENDED PERIOD IMMEDIATELY ELIMINATES SUCH AS ED GAEDEL OR THE WHOLE COLLEGE TEAM DRAFTED TO PLAY A MAJOR LEAGUE GAME DUE TO A PLAYER STRIKE. ALSO, A WEALTH OF CANDIDATES PERFORMING DURING THE WAR YEARS OF THE 40'S CAN NOT BE RECOGNIZED BECAUSE THEY PLAYED AGAINST EACH OTHER, THUS STATISTICALLY UPGRADING THEIR RELATIVE LACK OF SKILLS.

IN ADDITION TO DURABILITY, OTHER QUALIFICATIONS INCLUDE WEAK HITTING AND SLOPPY FIELDING. A "GOOD HIT,NO FIELD" LABEL DISQUALIFIES A PLAYER BY DEFINITION. OBVIOUSLY, THESE CRITERIA PRECLUDE PITCHERS FROM SERIOUS, NOT ALL, CONSIDERATION; BUT, ONE WHO PLAYED OTHER POSITIONS, CLINT HARTUNG COMES TO MIND, MAY BE ABLE TO PROVIDE AND ADDED DIMENSION TO INEFFICIENCY.

ENOUGH ALREADY!! MY SELECTION IS JOHN GOCHNAUR, BORN SEPT. 12, 1875. HE PLAYED WITH THE CLEVELAND INDIANS IN 1902-1903, HIS GLORY YEARS.

IN 1902 CLEVELAND FINISHED IN FIFTH PLACE WITH A 69-67 RECORD; GOCHNAUR WAS THE SHORTSTOP FOR 127 OF THOSE GAMES. HE COMPILED A .185 BATTING AVERAGE IN 459 AT-BATS. THE NEXT YEAR CLEVELAND IMPROVED TO 77-63 AND THIRD PLACE. HOWEVER GOCHNAUR, A MODEL OF CONSISTENCY, AGAIN HIT .185 (HIS 16 DOUBLES, 4 TRIPLES, AND NO HOMERS EXACTLY DUPLICATED HIS 1902 EXTRA BASE PRODUCTION). HE PLAYED 134 GAMES AT SHORT AND HIS 438 AT-BATS ATTEST TO HIS FULL-TIME STATUS. THESE NUMBERS CONFIRM AN INHERENT INEPTITUDE RATHER THAN BENCH-SITTING RUSTINESS, ANOTHER MARK IN HIS FAVOR---- HE WAS A NATURAL!

DEFENSIVELY, HIS FIELDING DOUBLY QUALIFIES HIM FOR THE TITLE OF "WORST". IN 1902 HE MADE 48 ERRORS, UNFORTUNATELY NOT TOO BAD IN AN ERA OF POOR QUALITY EQUIPMENT (GLOVES MADE IN THE U.S.A.) AND GRASS. THE NEXT YEAR, THOUGH, HE BOUNCED BACK TO MAKE 98 ERRORS; THIS APPEARS TO BE THE MOST ERRORS EVER MADE BY ONE PLAYER IN A SINGLE SEASON. 1903 WAS GOCHNAUR'S LAST SEASON IN THE MAJORS (HE HAD COME UP WITH BROOKLYN FOR 3 GAMES IN 1901). WITHOUT BENEFIT OF OUR PRESENT DAY PSYCHOLOGICAL INSIGHTS, HE DID NOT REALIZE THAT HE WAS PRESSING AND LETTING HIS FIELDING AFFECT HIS HITTING----OR WAS IT THE OTHER WAY AROUND??

INTERESTINGLY ENOUGH THE CLEVELAND TEAMS IN 1902-193 DID NOT HAVE BAD ECORDS. CLEVELAND HAD THE HIGHEST TEAM BATTING AVERAGE IN EITHER LEAGUE IN 1902; A GOOD PART OF THE REASON BEING THAT THE OTHER HALF OF THEIR DP COMBINATION WAS NAP LAJOIE, HALL-OF-FAMER AND BATTING CHAMP 1901-1904. GOCHNAUR'S REPLACEMENT WAS TERRY TURNER WHO REMAINED WITH CLEVELAND FOR THE NEXT 15 YEARS.

Functions for Predicting Winning Percentage from Runs

Charles Hofacker

Introduction and Review

The most fundamental quantity in sabermetrics is the run, and for any team, the most important distinction pertaining to runs is the critical difference between runs that it scores and runs that it allows. The Pythagorean equation was devised by Bill James to predict winning percentage from these two basic quantities. One important use of this tool is to be able to generate won-loss records for offensive players. The formula is simple:

$$(1) \qquad P(\text{Winning}) = \frac{O^{\alpha}}{O^{\alpha} + D^{\alpha}} .$$

In (1), the symbol O is meant to represent the number of runs a team scores (Offensive runs) and D is intended to stand for the number of runs it allows (Defensive runs). The parameter α is a number that can only be discovered by looking at actual data. Bill James suggests that a value around 2.00 for α leads to good prediction, with 1.83 being slightly better. For the purposes of this paper, I would like to refer to (1) as the Pythagorean equation with the understanding that α is not necessarily fixed at 2.0.

The primary goal of this paper is to offer an explanation as to why the Pythagorean equation works so well. The explanation involves the shape of the frequency distribution for the ability to produce leads by each team. The argument also depends on the relationship between the Pythagorean equation and a technique widely used by researchers in several sciences, called logistic regression. But to do this, we need to take a brief walk down the murky path of mathematical statistics.

Linear Regression

The Pythagorean equation is one of many functions that might possibly be used to do the following task. What we want is to have a function f for which, over the course of a season, P(Winning) = f(O,D). One might have three criteria for picking f. First, f should have the property

Acknowledgements

Paul Hoffman loaned a helping hand computerizing some of the data used in this paper. Also, I would like to thank the Office of Academic Computing at UCLA for providing the computing time necessary to do the analyses reported in this paper.

that it comes close to predicting the correct winning percentage for most teams. Another way to say this is that the empirical fit of the function to the data should be good. Second, the form of f should reflect the processes that we believe enter into winning. A possible third criterion is ease of use, a criterion most relevant if computations are done by hand.

No doubt the first thing that would pop into the mind of most social scientists would be linear regression. Linear regression assumes that if we plotted P(Winning) as a function of O and D, we could fit the resulting scatter of points in space with a plane tilted at an orientation determined by the data. If we believe that O and D are identically important we could plot P(Winning) as a function of the difference between O and D. In this simpler case, we could fit the scatter of points with a line whose slope is β, a parameter to be estimated from the data. The line would cross the P(Winning) axis at .5. Algebraically, we would have:

$$(2) \qquad P(\text{Winning}) = .5 + O\beta - D\beta = .5 + (O - D)\beta.$$

There are (at least) two problems with this formulation. First, the probability of winning must be between 0 and 1. There is no guarantee in (2) that we would always get predictions between 0 and 1. I suspect that such a result would make most sabermetricians and laypersons alike suspicious that the process that gives rise to the data is not being correctly represented. Second, most people familiar with the game would object to (2) on the basis that in actuality you need more of a change in O - D to climb the same number of percentage points once you get over .6 or .7. Likewise, you need to lose more competence to drop from .4 to .3 than you do to drop from .5 to .4. In other words, probability of winning is a nonlinear scale, with very small and very high probabilities stretched relative to probabilities around .5. Any function that predicts winning percentage from the difference between O and D would seem to need to be S-shaped. In the next section I would like to propose a rationale for the feeling that the function should be S-shaped.

Distributions of Leads

It is patently obvious that some teams are better than others offensively. But the number of runs any team scores in a given game depends on a large number of circumstances. Some of these circumstances are known, or are at least knowable, including who is in a slump at the moment; who they are facing on the mound; what stadium they are playing in; and who is hurt. And of course, there is always a random or luck factor representing all the things we do not know. In either case, the number of runs the team scores is a random variable which nevertheless behaves according to a particular distribution having a central tendency depending on the overall offensive ability of the team. A parallel argument can be made for the number of runs that the team gives up.

Figure 1 shows the frequency distribution of runs scored at the end of nine innings per team per game for both leagues in 1983. In this figure and the other figures in this paper, Americal League data are represented by triangles, and National League data by squares. It is safe to say that each team has a distribution shaped something like Figure 1 for the number of runs it scores per nine innings, although for offensively superior teams the average is likely to be to the right of the average for its league taken as a whole. Lets call this distribution the team's offensive distribution. Again, a parallel argument can be made for defensive runs. Each team would have a defensive distribution that would look something like Figure 1.

In the discussion that follows, I would like to ignore what occurs in extra innings. The percentage of games tied after nine innings in the 1983 season was 8.9%. The simple model I am about to introduce could be made to accomodate extra innings, but at the cost of increasing the complexity of the following discussion.

The probability that a given team wins the i-th game of the season might be expressed as

$$(3) \quad P(i) = p(O_i > D_i) = p(O_i - D_i \geq 1)$$

where O_i is a score randomly drawn from the team's offensive distribution and D_i is randomly drawn from its defensive distribution. The difference in the parentheses in the second part of (3) merely says that the probability of winning depends on the probability of having any "positive lead." In Figure 2, I have shown the distribution of leads at the end of nine innings for each game in the 1983 season. The lead is expressed from the point of view of the home team. For example, a lead of -2 gives the probability the home team loses by 2 runs. Any particular game played in 1983 is a sample drawn from this distribution. If the sampled value is at or to the right of the +1 point, the home team wins. If it is at or to the left of the -1 point, it loses. To find the p(Winning), we would cumulate the function pictured in Figure 2 from one to plus infinity, plus a small unspecified amount for the probability the home team wins in extra innings.

As a brief aside, the National League seems to play more close games than the American League. Looking back at Figure 1, low scoring seems to be more prevalent in the NL and this could explain the difference. Another odd feature of Figure 2 is that the frequency of ties after nine innings somehow looks lower than it should. The dip for leads of 0 is possibly evidence of managerial strategy in games tied in late innings. Successful execution of single run strategies such as bunting and stealing might account for the dip.

Each team has a lead distribution that looks more or less like Figure 2, although for teams playing worse than .500 ball, the center of the distribution is to the left of 0, and for better teams, to the right. If we move a distribution slowly from the far left to the far right, more and more games fall to the right of one, leading to higher and higher winning percentages as we move.

As can be seen from Figure 2, the lead distribution is bell-shaped. Most of the mass is piled up in the middle, and there is a tail on each side. It is this shape that leads to an S-shaped curve as we move the distribution past the +1 point. During the time the tails of the bell are moving past the +1 point, P(Winning) is increasing relatively slowly.

Figure 2 bears close resemblance to what statisticians call a normal distribution. Such a result is not surprising as there is a famous statistical theorem that states the difference between any two distributions will tend to resemble a normal distribution.

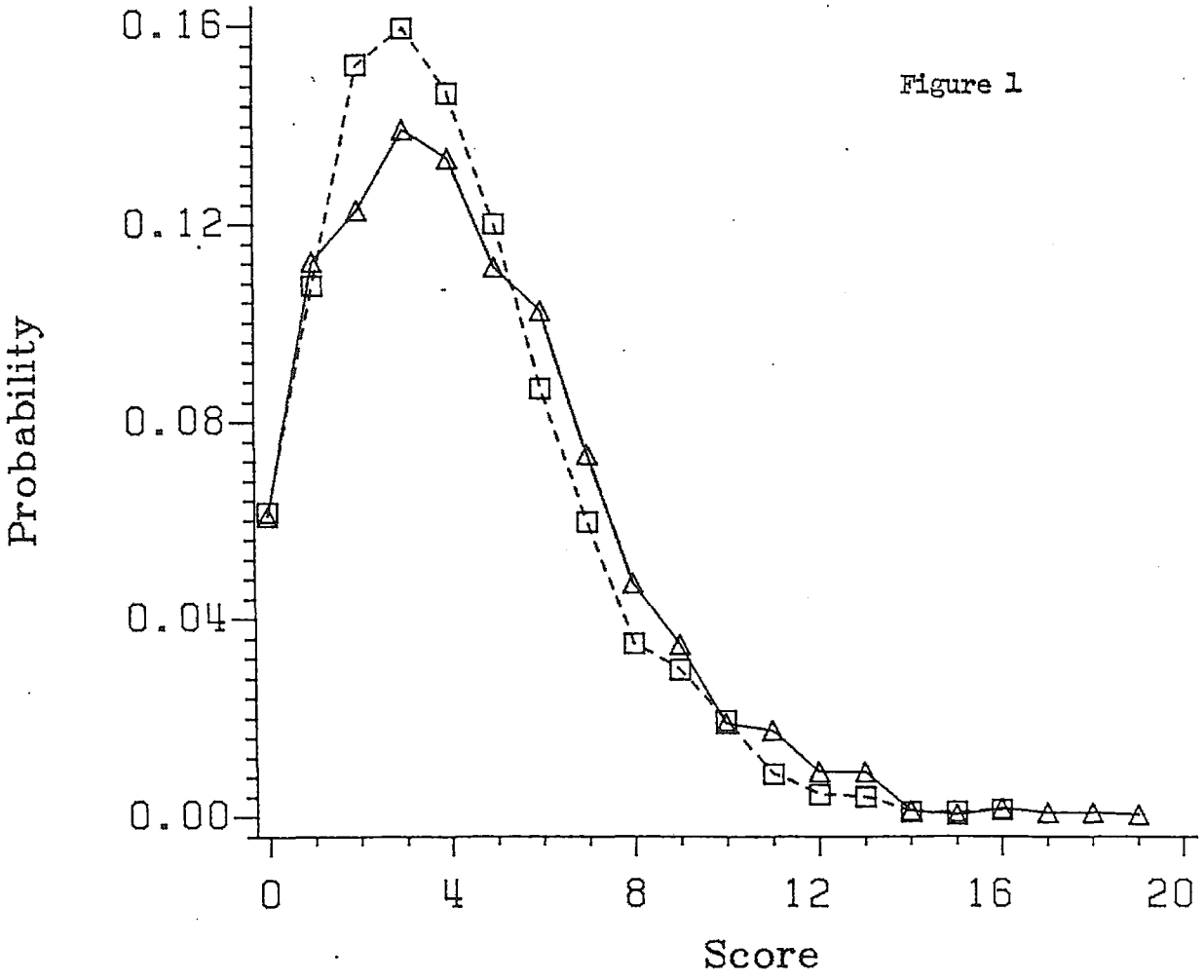


Figure 1

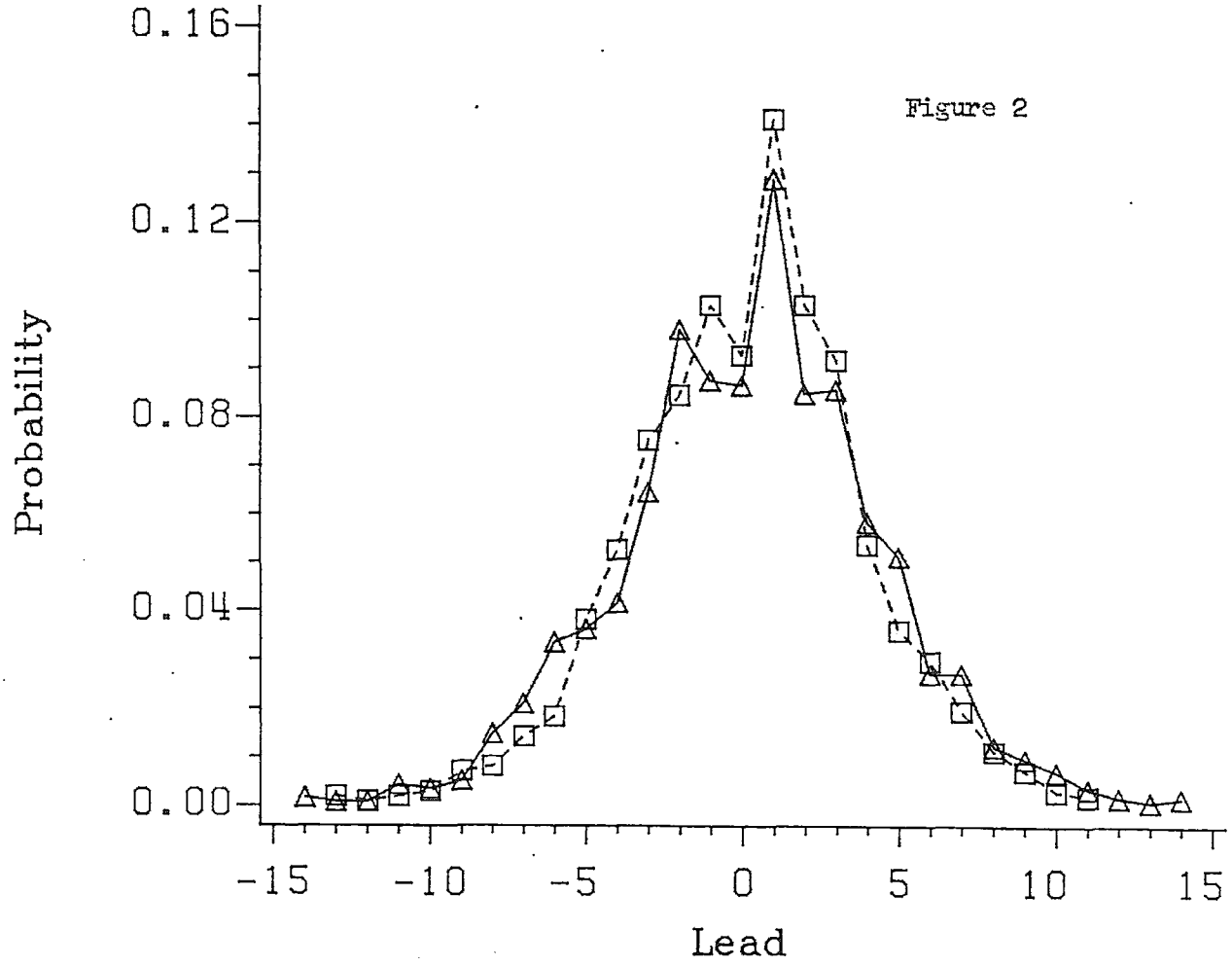


Figure 2

Logistic Regression

There are two natural choices for f once it is agreed that f will look like a cumulative normal distribution. Both are frequently used in biology and economics for modeling binary outcome situations under normality assumptions. One is called probit regression and the other logistic regression. In practice, there is little difference between the two. I will talk about logistic regression primarily due to its similarity to the Pythagorean equation. Here is the formula for logistic regression:

$$(4) \quad P(\text{Winning}) = \frac{\exp(O\gamma - D\gamma)}{1 + \exp(O\gamma - D\gamma)}.$$

The function exp raises the argument in parentheses to the power of e, a constant approximately equal to 2.7. Here, γ is a parameter to be estimated from the data. Equation (4) might make more sense to readers after some algebraic manipulation. First, take logs of both sides:

$$\ln[P(\text{Winning})] = (O\gamma - D\gamma) - \ln[1 + \exp(O\gamma - D\gamma)].$$

Here, ln takes the log of its argument (in brackets) using e as the base for the logarithmn. Next, subtract ln[1 - P(Winning)] from both sides:

$$\ln[P(\text{Winning})] - \ln[1 - P(\text{Winning})] =$$

$$(O\gamma - D\gamma) - \ln[1 + \exp(O\gamma - D\gamma)] -$$
$$\ln(1) + \ln[1 + \exp(O\gamma - D\gamma)],$$

and simplify both sides to yield

$$\ln[P(\text{Winning}) / (1-P(\text{Winning}))] = O\gamma - D\gamma .$$

Note that the log of a ratio is equal to the log of the numerator minus the log of the denominator. Note also that the log of one is zero.

Logistic regression is similar to linear regression except that instead of predicting P(Winning), logistic regression attempts to predict the log of the ratio of winning to losing, called the logit.

Next, lets compare the Pythagorean equation to logistic regression. First, take logs on both sides of (1):

$$\ln[P(\text{Winning})] = \ln(O^{\alpha}) - \ln(O^{\alpha} + D^{\alpha})$$

Next, subtract the log of 1 - P(Winning) from both sides:

$$\ln[P(\text{Winning})] - \ln[1 - P(\text{Winning})] =$$

$$\ln(O^{\alpha}) - \ln(O^{\alpha} + D^{\alpha}) - \ln(D^{\alpha}) + \ln(O^{\alpha} + D^{\alpha}).$$

Now, simplify to

$$\ln[(P(\text{Winning}) / (1-P(\text{Winning}))] = \ln(O^{\alpha}) - \ln(D^{\alpha})$$

$$= \ln(O)\alpha - \ln(D)\alpha .$$

The Pythagorean equation can be seen to be logistic regression after O and D have been log transformed, a very minor change from (4).

The Pythagorean equation predicts the same winning percentage for any two teams with the same ratio of O to D. For example, if the ratio of runs scored to runs given up is either 2 to 1 or 300 to 150, the Pythagorean formula predicts a winning percentage of .8. Such a property is useful for estimating winning percentages for offensive players, when it becomes necessary to estimate using a small number of offensive runs.

The logistic regression model presented in (4), on the other hand, predicts the same winning percentage for any two teams with the same difference between O and D. Such a property makes it less attractive as a tool for offensive players because you have to recalibrate the model every time you wish to predict winning percentage based on a different number of games. Another way to put this is that γ depends on the metric of O and D. For example, we would get a different number for γ in (4) depending on whether we used average number of runs or total runs. The fit of the function to the data would be the same in both cases, however.

Estimation of Models

If one has access to a computer with the appropriate statistical software, it becomes an easy matter to find the best estimates for either α or γ using logistic regression. One reason to use a computer in this fashion might be to test the idea of the symmetry of influence of O and D on winning. Another reason is to graph the best prediction in order to get an intuitive feel for how well the equation fits actual data. Data from McMillan's Baseball Encyclopedia were used to both these ends. Each observation consisted of one team's data for one year. The years included in the study were 1962 - 1975. When total offensive and defensive runs were used as predictors of winning percentage, the best value for γ turned out to be .0027. When separate parameters were estimated for offensive and defensive runs, the two estimates agreed almost perfectly. The best value for α turned out to be 1.74, close to the 1.83 reported in the Abstract. The fit of (4) was very slightly better than for the Pythagorean formula with $\alpha = 1.74$.

Figure 3 shows the best prediction curve for (4), along with a point for each team-year. To get a feel for the prediction of the Pythagorean equation, P(Winning) has been graphed against the ratio of O to D in Figure 4. The function pictured in Figure 4 is not S-shaped only because I used the ratio of O to D along the abscissa, rather than the log of the ratio.

Conclusions in Review

I have offered a possible explanation for the success of the Pythagorean equation in predicting success in baseball. The Pythagorean equation generates an S-shaped prediction curve for winning percentage when the difference between the log of O and the log of D (or ln[O/D]) is used as the horizontal axis. Such a shape is reasonable when the shape of the distribution of team ability to produce leads is considered. In particular, the function relating probability of winning to the difference between offensive and defensive runs is S-shaped because the distribution representing the difference between offensive and defensive runs is bell-shaped. Additionally, the hypothesis of symmetry between offense and defense implicit to the Pythagorean equation is confirmed.

Figure 3

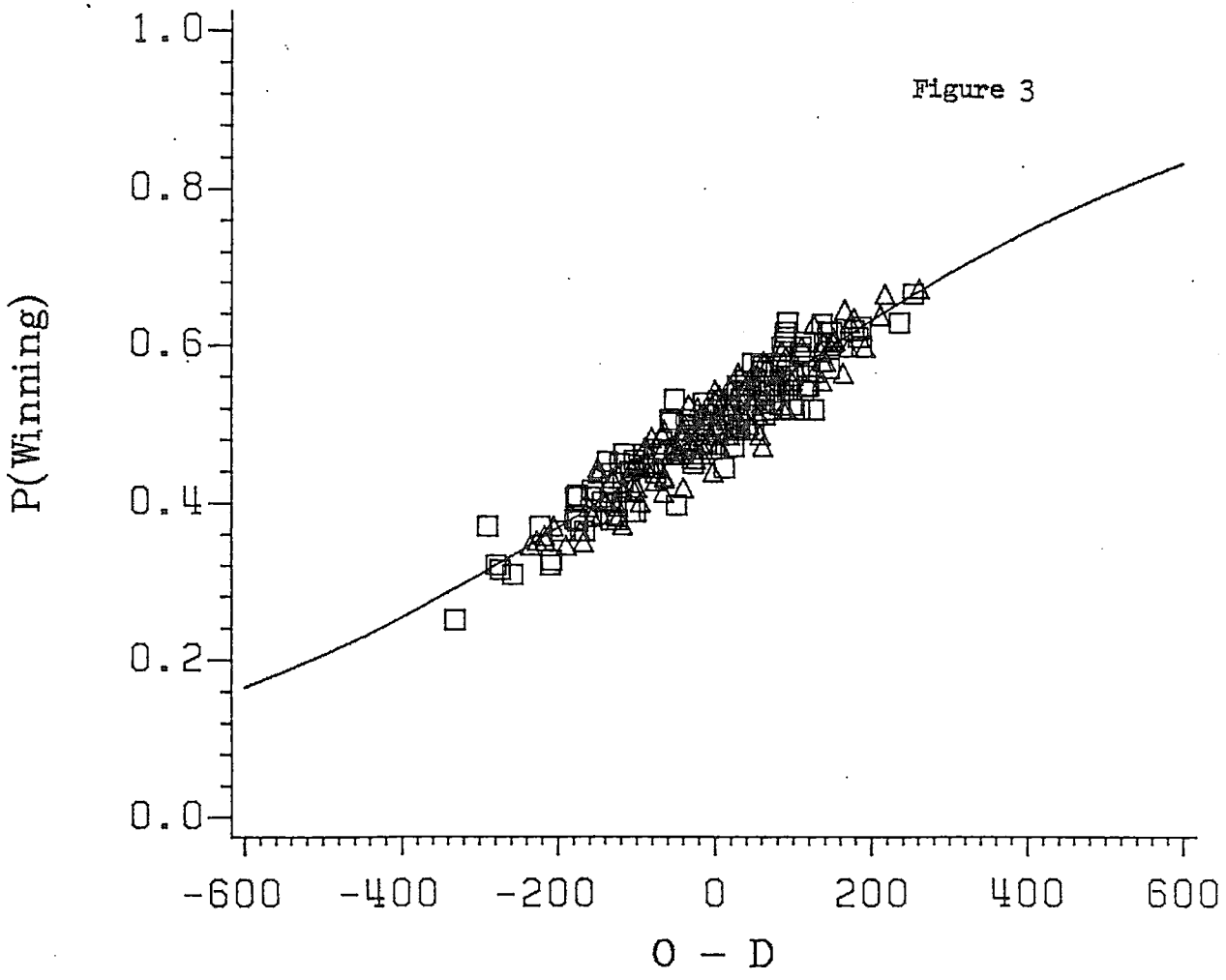
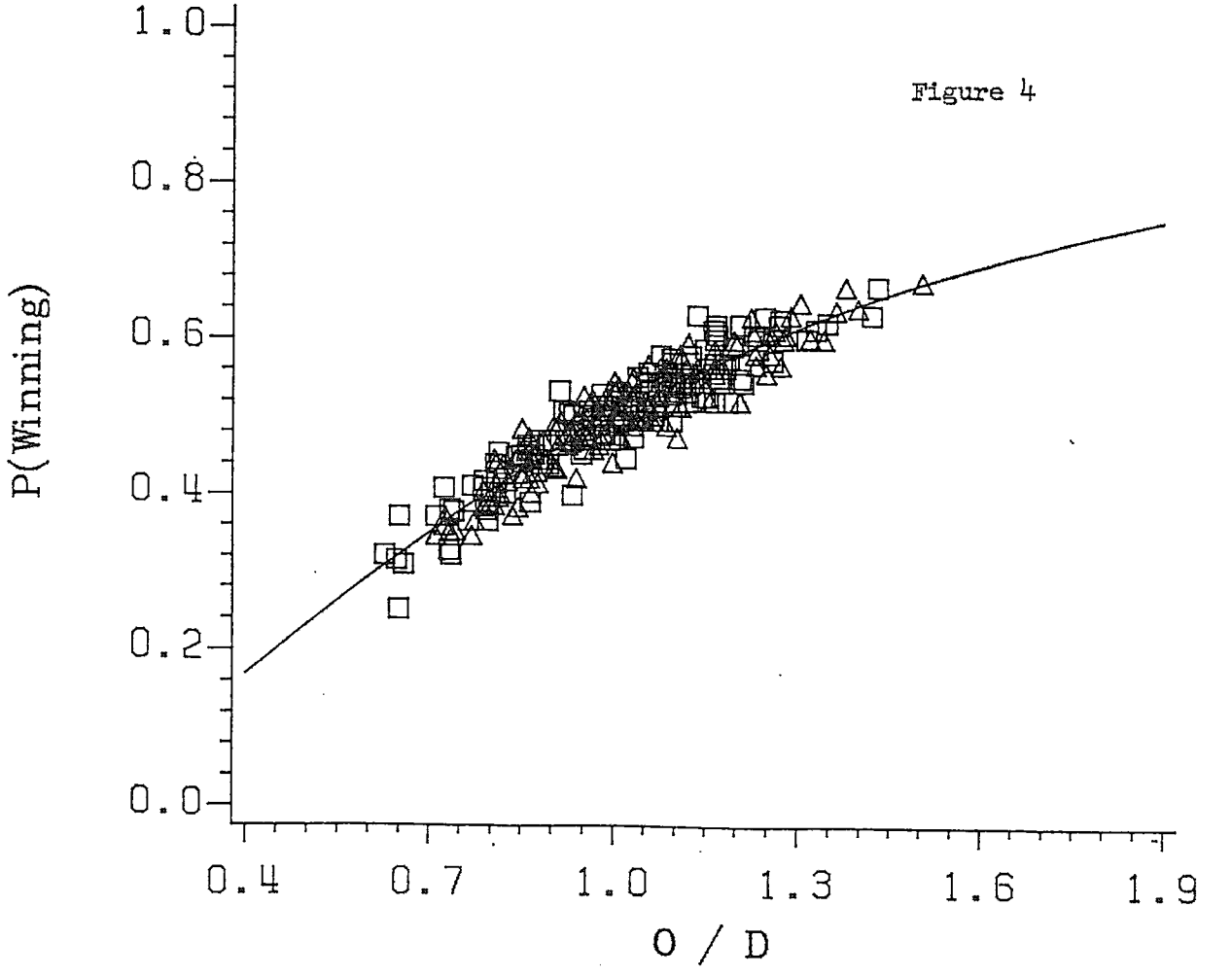


Figure 4



Barry Mednick on

Assists Versus Strikeouts

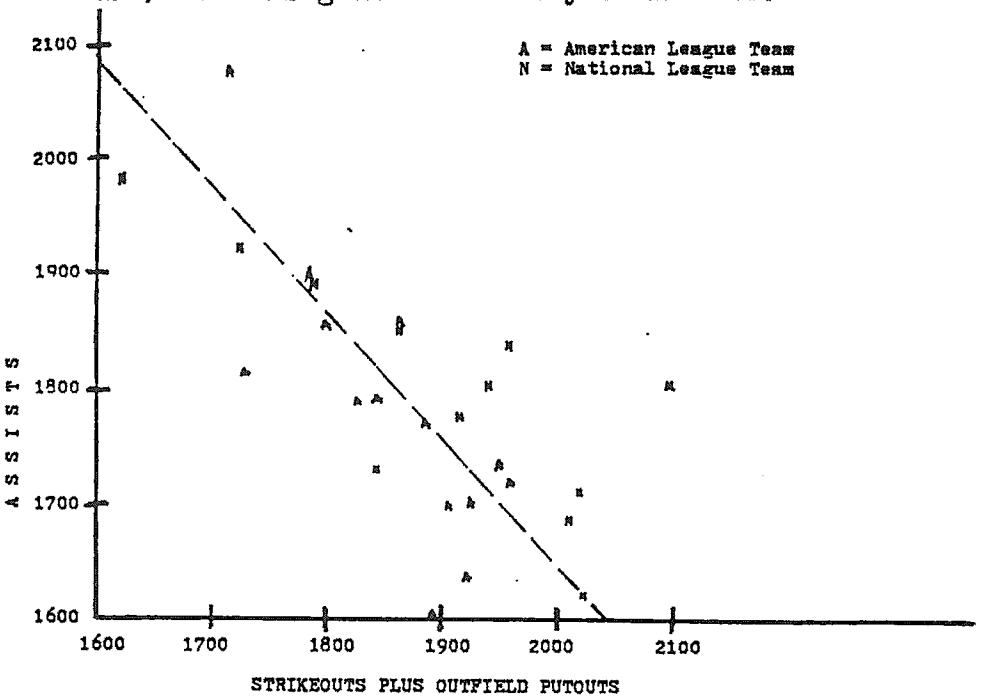
It makes sense to say that every strikeout means one less out by another method. The more strikeouts, the fewer ground outs and flyouts. Conversely, the more ground balls that infielders can turn into outs, the fewer opportunities a pitcher has for strikeouts.

This simple bit of logic has implications for calculation of range factors for fielders. The range factor (defensive plays per game) depends in part on the style of a teams pitching staff. The range factor calculation is based on the assumption that over the course of a full season, the number of balls put into play will be about the same for each position. Although the number of putouts per team is quite similar (from a low of 4255 for the Mariners to the Angels 4422), the number of strikouts covers a wide range from Kansas City's 593 to the Phillies 1092. This results in a disparity in the number of assists for each team. Oakland had only 1602 assists, compared to California's total of 2077.

The graph below plots assists as a function of a team's total of strikeouts plus outfield putouts. The numbers used for outfield putouts are extimates and could be low. The trend is what one would expect. Teams with a high number of strikeouts and flyouts have fewer assists. This implies that infielders on such teams have lower range factors than they would if they played for teams whose pitching staffs tended to produce more ground balls.

Using the laast squares method.and assuming a linear relationship, the American League data show a relationship of one assist per 0.94 strikeouts plus flyouts. The National League value is 1.20. The difference is probably due to the designated hitter, which gives National League pitching staffs their margin of 113 added whiffs per team. The line on the graph is the average of the two leagues.

If there is a bias in range factors, what do we do about it? It is possible to weight ranges based on the quality of the pitching staff. It would be better to calculate an efficiency rating showing the ability of each player to handle balls put into play in his area. This requires statistics on where balls are hit, something not currently available.



AN ANALYSIS OF WIN PERCENTAGE

by Bill Deane

The Sports Encyclopedia Baseball (Neft, Cohen & Deutsch), 1981 edition, contains an interesting study of pitchers' won-lost percentages as compared to those of their teams. As remarked in the edition, "baseball executives have long used (this) comparison...as a measure of the pitcher's performance."

At the time, the all-time leader was Tom Seaver, whose .635 career percentage was 124 points higher than the .511 of his teams (the average of the winning percentages of the teams he pitched for each year weighted by his number of decisions each year). Following Seaver were Grover Alexander (108), Walter Johnson (103), and Lefty Grove (95).

Interesting though it may be, this study has at least three basic flaws:

(1) The pitcher's record should be subtracted from his team's before the team percentages are computed (however, I doubt that this would significantly alter the ranking of pitchers in this category).

(2) The system assumes that everyone else on a given staff is "average", penalizing a pitcher on an outstanding staff, and rewarding one on a weak staff. (Again, I am not convinced of the significance of this inequity; the law of competitive balance should even things out in the long run.)

(3) The system gives equal value to a pitcher who posts a .600 percentage for a .500 team, as one who logs a .700 mark for a .600 team, although the former pitcher has more room for improvement.

Allow me to push aside the first two flaws for now, and concentrate on the third.

In 1972, Steve Carlton had an amazing 27-10, .730 record for the last-place Phillies, who were a woeful 32-87, .269 in games Carlton did not get a decision. Carlton's percentage, therefore, exceeded his team's by a whopping 461 percentage points.

It is hard to imagine any pitcher having a better season than Ron Guidry in 1978, when he went 25-3 with an ERA (1.74) more than two runs better than the league's. Yet, Guidry's .893 win percentage was "only" 337 points above that of the Yankees (75-60, .556 without Guidry's decisions); and even if Guidry had been a perfect 28-0, he would have fallen short of Carlton's 461 point differential.

The key is that Carlton had more room for improvement— 731 possible points (1.000 minus .269)— than Guidry, who had 444 possible points (1.000 minus .556). Taking nothing away from Carlton's incredible '72 performance, he does have an unfair advantage in this example.

This formula— the Pitcher Performance Percentage (PPP)— can help resolve the issue:

$$PPP = \text{Average Pct.} + \left[\left(\frac{\text{Pitcher's Pct.} - \text{Team Pct.}}{\text{Perfect Pct.} - \text{Team Pct.}}\right) \times (\text{Perfect Pct.} - \text{Average Pct.})\right]$$

Since we know that "average pct." is always equal to .500, and "perfect percentage" is always equal to 1.000, we can simplify the formula as follows:

$$PPP = .500 + \left[\frac{\text{Pitcher's Pct.} - \text{Team Pct.}}{2(1.000 - \text{Team Pct.})}\right]$$

With this formula, the effective win percentage, or PPP, for the two pitchers would be as follows:

$$PPP\ (\text{Carlton}) = .500 + \frac{.730 - .269}{2(1.000 - .269)} = .500 + \frac{.461}{2(.731)} = .500 + .315 = \underline{.815}$$

$$PPP\ (\text{Guidry}) = .500 + \frac{.893 - .556}{2(1.000 - .556)} = .500 + \frac{.337}{2(.444)} = .500 + .380 = \underline{.880}$$

What this presumes to tell us is that, if Carlton in '72 and Guidry in '78 had pitched for .500 teams, their win percentages would have been .815 and .880, respectively. Absolute fact? Probably not. Considerable possibility? I hope so!

Now, adapting this formula to career records, we can compare pitchers on an even basis. Of the 64 pitchers who have won 200 or more games since 1900, here are the top 15 finishers in PPP:

PITCHER	PCT.	TEAM	PPP
Lefty Grove	.682	.584	.618
Pete Alexander	.642	.534	.616
Tom Seaver*	.616	.505	.612
Whitey Ford	.690	.604	.609
Walter Johnson	.599	.496	.602
Christy Mathewson	.665	.587	.594
Cy Young**	.600	.513	.589
Juan Marichal	.631	.553	.587
Carl Hubbell	.622	.559	.571
Steve Carlton*	.600	.534	.571
Bob Feller	.621	.562	.567
Ted Lyons	.531	.459	.567
Jim Palmer*	.643	.591	.564
Bob Gibson	.591	.532	.563
Mordecai Brown	.648	.598	.562

* Active; stats through 1983.
** Does not include pre-1900 stats.
Statistics from The Baseball Encyclopedia (MacMillan)

THE BEST FIELDING SECOND BASEMEN....continued from page 8

And maybe we should retitle the place as the Hitting and Pitching Hall of Fame. Certainly any offensively skilled player who so dominated during a long career, who stood out in the statistics, would long ago have been in the Hall of Fame. Until more recognition is given to fielding as one of the critical and essential skills of the game, not just by managers who know and live by that concept, but also by fans and pundits, then great fielders, among them the peerless fielding second baseman Bill Mazeroski, will not receive their due.

A NALYST Back Issues...

Due to popular demand, the archives here at the Analyst are being thrown open so that we may bring new subscribers a chance to get their hands on back issues. All will appear just as they were originally issued except that they will be marked "reprint" on the front cover.

ANALYST INDEX

Issue #1, June 1982: "Ballpark Effects on the Production of Infield Errors and Dps"...Paul Schwarzenbart; "Distribution of Runs Scored"...Dallas Adams; "Analysis of Nolan Ryan's Fifth No-Hitter"...Tom Jones; "Wins and Losses For All Players"...Mark D. Pankin; "Home Runs--A Matter of Attitude" ...Robert H. Kingsley

Issue #2, August 1982: "Some Patterns of Age and Skill" and "The Effects of Overwork on Rookie Pitchers"....Dallas Adams; "Ballpark Effects on Fielding Performance: Further Evidence"...Craig Wright; "Run Production by Batting Order Position" and "Clutch Hitting" by Dick O'Brien; "In Search of the 'True' Slugging Percentage"...Jim Morrow

Issue #3, October 1982: "More on the 'True' Slugging Percentage"...Jim Reuter; "Batting Average Comparisons"...Ward Larkin; "Effects of Overwork on Rookie Pitchers, Part II"...Dick O'Brien; "Player Development Study" ...Craig Wright

Issue #4, December 1982: "Measuring Relief Performance"...John Billheimer; "Some Additional Aspects of the Distribution of Runs Scored"...Dallas Adams; "A New Look at 'Hard Luck' Pitchers"...Mark Lazarus; "Thoughts on Isolated Power"...Jim Reuter

Issue #5, February 1983: "Effect of Batting and Pitching Changes on Team Won-Lost Records"...Dick O'Brien; "Home Park Factors"...Jim Reuter; "Balls and Strikes"...Pete Palmer; "Some Additional Aspects of the Distribution of Runs Scored, Part II"...Dallas Adams

Issue #6, April 1983: "Ballpark Effects on Fielding Statistics, Part II" ...Paul Schwarzenbart; "Quality versus Quantity in Comparison of Career Stats"...Dan Heisman; "Team Won/Lost Percentage as a Function of Runs and Opponent's Runs"...Dallas Adams; "Adjusting Home Park Factors".... Pete Palmer; "Evaluating Pitchers Performances"....(The) Cutbert Magnolia

Issue #7, August 1983: "Run Production by Batting Position, an Update" ...Dick O'Brien; "The Probability of Hitting .400"...Dallas Adams; A Trend Analysis of Batting Averages"...Gary T. Brown; "Assigning Relative Values to Relief Wins, Losses and Saves"...John Schwartz; "Distribution of Runs" ...Pete Palmer

Issue #8, October 1983: "Announcing Project Scoresheet"...Bill James; "Scoring Sequences"...Barry Mednick; "On Handedness and Pitcher's Fielding" ...Warren Johnson; "Pitcher's Range Factors"...Clem Comly; "Power Hitter's Strikeout/Home Run Ratios"...Dick O'Brien; "Foul Balls Effect on Batting Average"...David Aceto; "The Left-Handed Hitter's Advantage"...John Scwhartz; "The Max Patkin Story, A Film Preview"...John Borkowski and Jim Baker

All issues are 20 pages each, except #4, which is 24 pages. Each issue costs $2.75. This includes postage. Make checks or money orders payable to The Baseball Analyst, and mail to 945 Kentucky Street/Lawrence, KS 66044.