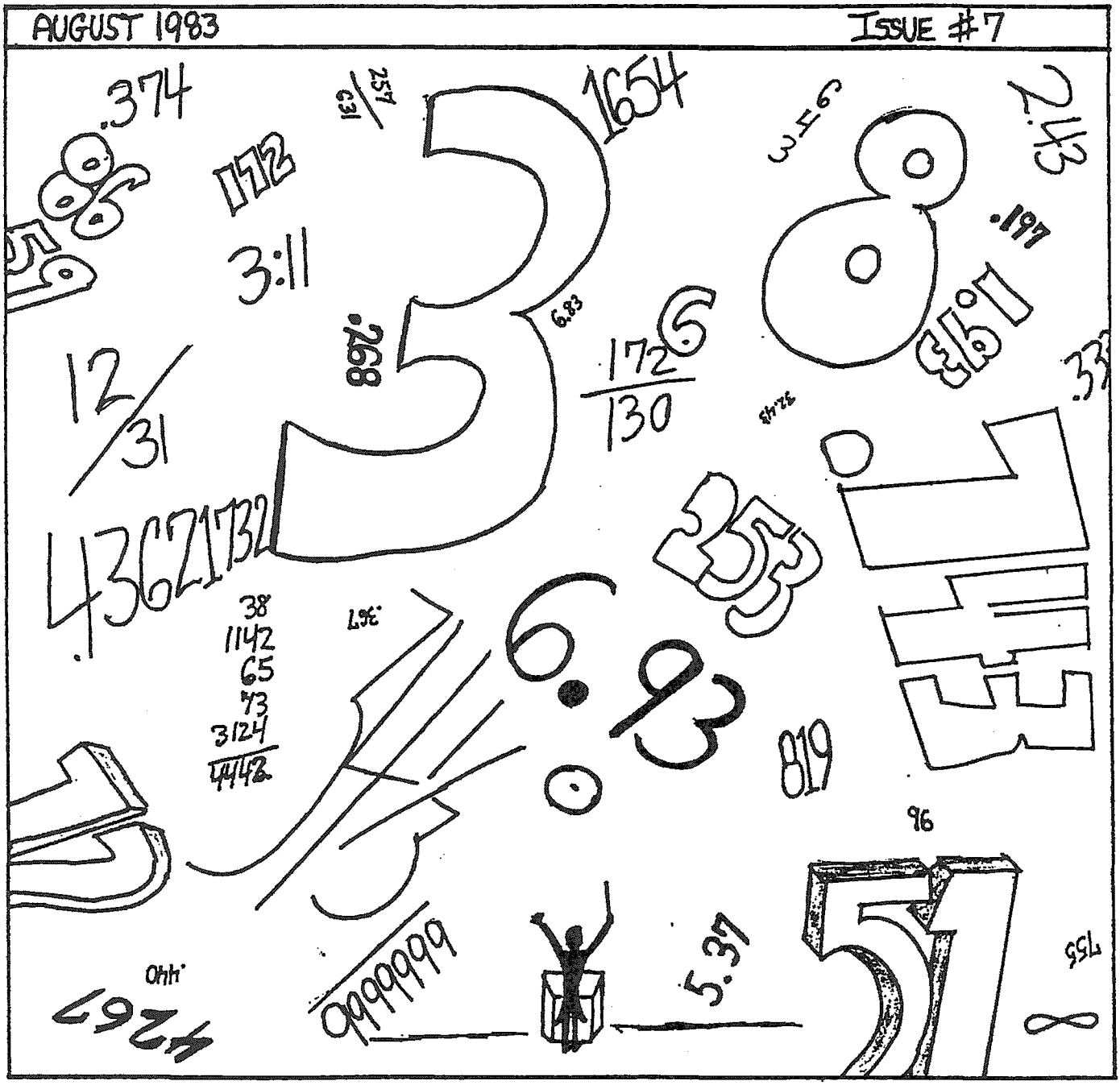
baseball-baseball-baseball-baseball-baseball-baseball-baseball-baseball-baseball

ANALYST

AUGUST 1983 ISSUE #7



SABERMETRICS:

Making Numbers Perform For You

RE-PRINT

NEW EDITOR'S NOTE

I have a dream, to start humbly, and it is that someday, sabermetrics will solve the mystery of baseball. Don't take this to heart; I know it wouldn't be much fun if we knew the answers to everything, but I feel that if total understanding could be achieved, sabermetrics would be the means. However, the field is just in its infancy, or, to put it on an evolutionary scale, it's just sliding its way onto land.

So here you are at the relative beginning of a new and exciting field. You missed psychology, nuclear physics, rock music, automotive design, and the birth of the film industry--anything you might do in those fields would either be icing on the cake or mere redundancy. But sabermetrics--ah ha! There's your chance! You are the original 100--the Freud's and Jungs; the Einstiens and Oppenheimers; the Buddy Hollys and Chuck Berrys; the Fords and Olds; the Edisons and D.W. Griffiths. You can make or break this thing. It is not often you get the opportunity to be in on something new. It is even more infrequently that you are given a forum for your ideas on that subject. This is the delivery room doctors, are you going to give the baby a hand, or will you drop it on the floor? In other words, we need more material for the Analyst.

My name is Jim Baker, new editor for the Baseball Analyst. If my pleas for submissions sould like those of our founder/publisher, Bill James, it is because we have both been faced with the same problem--a serious lack of reader participation. We have enough readers to keep the issues coming, but not enough contributors. Don't make us beg--let's see some of those studies you've been doing. By keeping them in your notebook you gain nothing, but by sending them to the Analyst you advance sabermetrics and help keep its journal in print. Soviet writers have a phrase called "writing for the desk drawer" because they have no outlet for their material. Well this is America--and here's your outlet.....

World Headquarters
945 Kentucky Street
Lawrence, KS 66044

$12.00 per annum

All Submissions ready for reproduction, please

baseball ANALYST

August 1983

Issue no. 7

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

IN THIS ISSUE:

LETTERS

Dear Baseball Analyst,

What sort of cult is this Sabermetrics thing anyway? You have taken my son Tommy, age 16, away from us, his loving parents. All day and all of the night he sits in his room with a calculator and ball scores doing "sabermetrics." He says he doesn't love us anymore because "we don't love him enough." If you really loved me," he goes, "there would be a visible impact on my behavior as opposed to other kids my age whose parents love them either more or less. According to my calculations, I don't see myself as being any more well-adjusted than they are. This love you profess should,

Cont. on page 8

RUN PRODUCTION BY BATTING ORDER POSITION--Part II

Dick O'Brien

My corresponding study in the August 1982 issue of The Baseball Analyst attempted to show what percentage of runs batted in were distributed among the nine regular batting order positions. That study has now been expanded to show both RBI and runs scored for a five year period (1978-82).

This larger data base pretty well confirms what was shown in the first study, but in the following tables, the percentages shown represent total team % rather than the total regular's % as were indicated in the first study. Using these tables as a guideline, the % can give us a pretty good idea of how individual players are contributing to total team production even though fewer and fewer lineups remain the same with the passing of each season.

Runs Scored

American League

Pos	Runs	Pct	
1	5487	.1180	
2	4961	.1067	
3	5092	.1095	.3342
4	5147	.1107	
5	4841	.1041	
6	4524	.0973	.3121
7	4436	.0954	
8	3743	.0805	
9	3492	.0751	.2510
Total team runs	46512		
By regulars	41723	.8971	
By subs and ph	4789		

National League

Pos	Runs	Pct	
1	4866	.1340	
2	4227	.1150	
3	4462	.1214	.3704
4	4326	.1177	
5	3922	.1067	
6	3565	.0970	.3214
7	3183	.0866	
8	2753	.0749	
9	1106	.0301	.1916
Total team runs	36762		
By regulars	32410	.8816	
By subs and ph	4352		

Runs Batted In

American League

Pos	RBI	Pct	
1	3240	.0741	
2	3954	.0882	
3	5686	.1301	.2924
4	6307	.1443	
5	5333	.1220	
6	4825	.1104	.3767
7	4364	.0998	
8	3347	.0766	
9	2921	.0668	.2432
Total team RBI	43706		
By regulars	39877	.9124	
By subs and ph	3829		

National League

Pos	RBI	Pct	
1	2733	.0796	
2	2959	.0859	
3	4503	.1312	.2967
4	5242	.1527	
5	4493	.1309	
6	3733	.1087	.3923
7	3358	.0978	
8	2637	.0768	
9	1185	.0345	.2091
Total team RBI	34329		
By regulars	30833	.8982	
By subs and ph	3496		

No position in the batting order varied more than 6% over the five year period with the following exception: in 1979, the seventh position outproduced the sixth in the AL in both runs and RBI. This was the only time when one positon exceeded its normal production. The designated hitter effect is clearly evident in the bottom third of the AL order: a 31% increase over the NL in runs and a 16% increase in RBI--almost all of which is accounted for in the ninth slot.

ON THE PROBABILITY OF HITTING .400

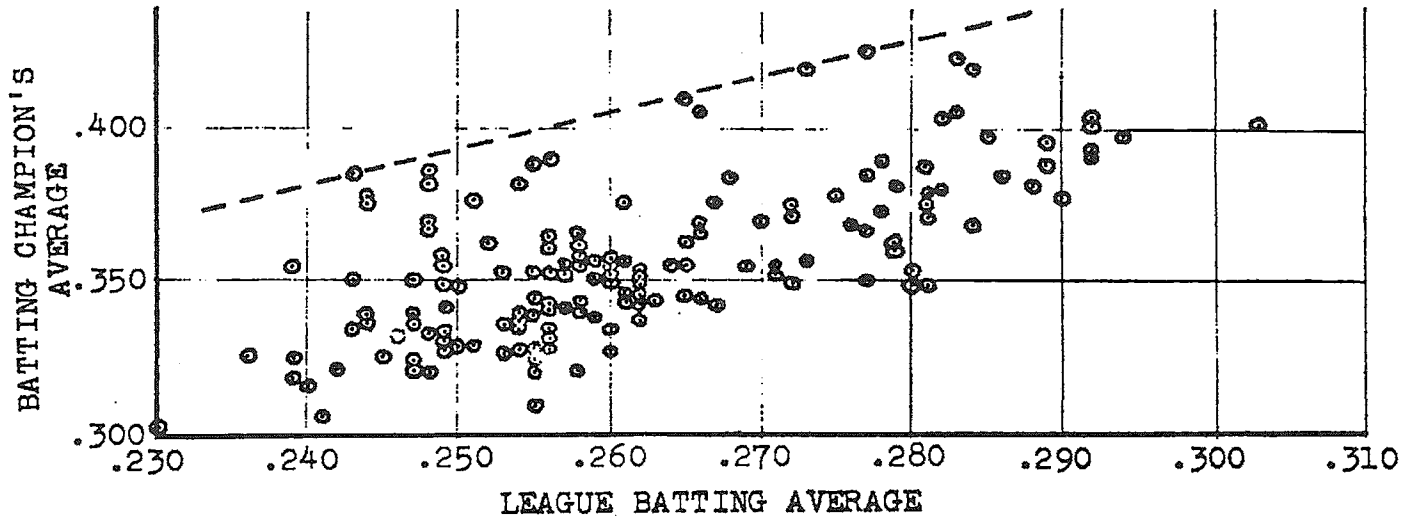
by Dallas Adams

In 1977 Rod Carew, while ultimately falling short, came close to hitting for a .400 average. The question naturally arises as to the probability of anyone hitting .400. The commonly held view nowadays is that night ball, transcontinental travel fatigue, the widespread use of top quality relief pitchers, big ballparks, large size fielders gloves, and other factors all act to a hitter's detriment and make a .400 average a near impossibility.

But, surely, the above items will affect all batters, not only the potential .400 hitters; and, therefore, the net effect of all these factors will be reflected in the composite league batting average. If the league average is low, the chance of there being a .400 hitter is also low; a high league average means a higher chance of a .400 hitter.

Consider the experimental data: Figure 1 shows, for each major league season from 1901-1976, the average for each league's batting champion plotted against the league batting average. Of particular interest is the dashed line which marks the rather well-defined upper boundary of the data points. This line represents the ultimate level of batting performance in 76 years of major league baseball. Note that this boundary crosses the .400 level of individual performance at a league average of .255, this can be considered the effective minimum league level from which a .400 hitter can emerge.

FIGURE 1
BATTING CHAMPION'S AVERAGE AS A FUNCTION OF LEAGUE BATTING AVERAGE (1901-1976)

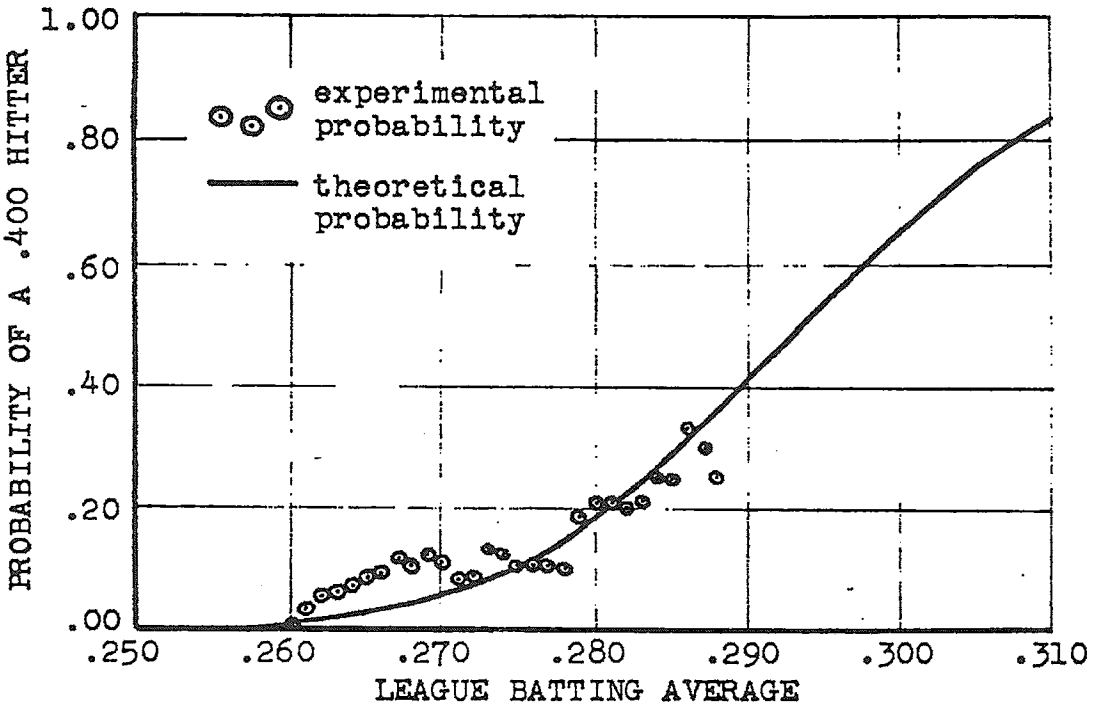


For any given league batting average, the experimental probability of an individual .400 hitter could, if there were sufficient data, be obtained directly off Figure 1 by counting. For example, at a league average of .265 there was one season with a .400 hitter and three seasons without; a probability of 25% for a .400 hitter when the league batting average is .265. Unfortunately this simple approach is inadequate because of sparseness of data; eleven .400 hitters spread over a range of .230 to .303 in league batting average. It is necessary, therefore, to group the data.

For this study a moving average covering .009 points in league batting average was employed. This means that the experimental data for each specific league batting average was augmented by all the data within ±.004 points. Thus for a .265 league average, by way of example, the 24 data

points in the range .261 through .265 are used, rather than only the four data points at exactly .265. Those ranges above .288 contained ten or fewer data points and were considered insufficiently populated to be included in the calculations. Despite the smoothing effect of the moving average technique, there remains some jumping about of the resultant experimentally determined probabilities but the general trend is apparent, as shown by the individual points on Figure 2. From these experimental results it can be seen that there is approximately a 6% chance for a .400 hitter if league batting averages remain at 1977 levels in the mid-.260's.

FIGURE 2
PROBABILITY OF A .400 HITTER AS A FUNCTION OF LEAGUE BATTING AVERAGE



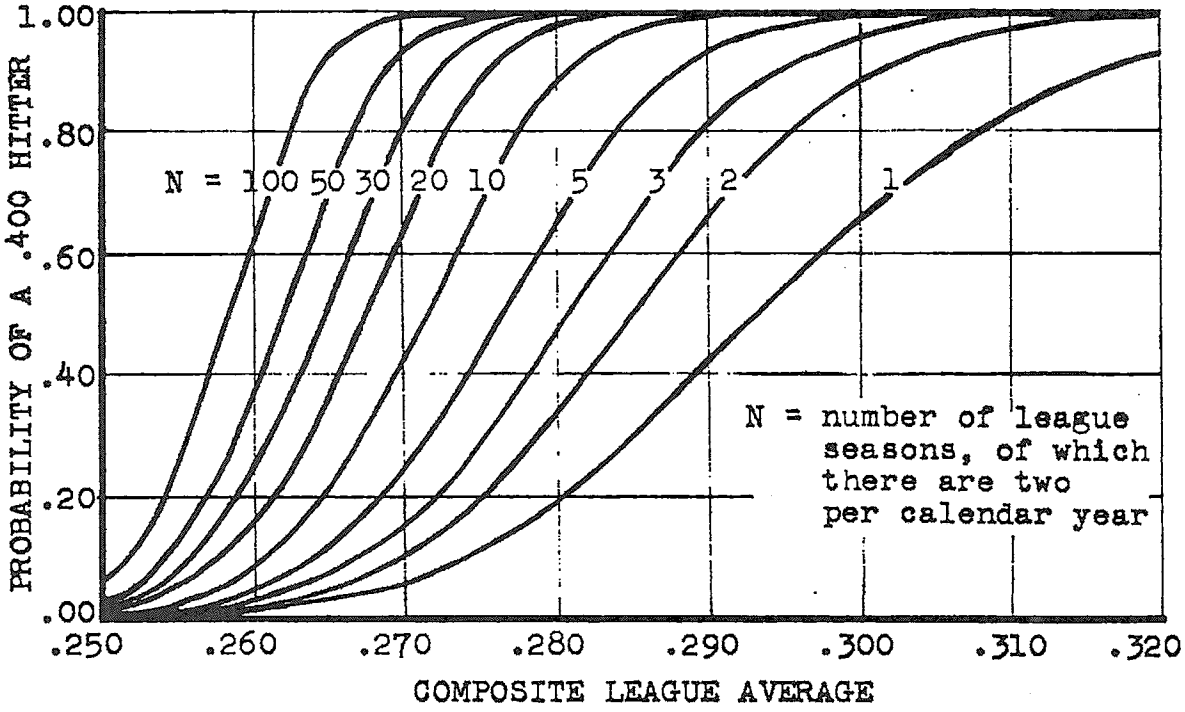
From a more theoretical point of view: consider, for example, a league batting average of .265; .400 is 51% higher than .265. Thus the question: what is the probability of a player compiling a personal batting average which is at least 51% higher than his league's .265 average? At this juncture it is necessary to introduce the "Relative Batting Average" concept of Shoebotham (1976 Baseball Research Journal, published by the Society For American Baseball Research).

In its simplest form, a relative batting average is a player's average divided by his league's average. If one calculates the relative batting averages for all major league batting champions from 1901 through 1976, the results approximate a Normal Distribution (the familiar "bell-shaped curve") with a mean (average) value of 1.363 and a standard deviation (a measure of the dispersion of the data about the mean) of 0.074. Now, the useful thing about a normal distribution of known mean and standard deviation is that the probability of occurance for any arbitrary value, above or below the mean, can be calculated. For a league average of .265, we want to calculate the probability of a player making a relative batting average of 1.51; the computations give a 2.4% probability for this.

Similar computations have been made for a wide range of league batting averages and the resulting theoretical probabilities are shown by the solid line on Figure 2. The theoretical and experimental results are in good agreement.

Figure 2 gives the probability of a .400 hitter occurring in a given season. But what about the chances of a .400 hitter appearing at least once in a given number of seasons? The correct procedure would be to consider, for the league batting average of each individual season, the probability of there not being a .400 hitter that season; then multiply together these probabilities and subtract the result from 1. Because of this, it would be extremely difficult to create a probability diagram showing, for the general case, the probability of a .400 hitter in a given period of years. A rough approximation may be made, however, by assuming that the composite league batting average over the whole period in question does not differ appreciably from the league average of any individual season within the period. Making this assumption, it is then possible to construct Figure 3. In employing Figure 3, remember that the results given by it are an approximation. Also note that the curves of Figure 3 are for different numbers of "league-seasons", of which there are two (one American and one National) per calendar year.

FIGURE 3
PROBABILITY OF A .400 HITTER OCCURING AT LEAST ONCE IN A GIVEN PERIOD, AS A FUNCTION OF THE COMPOSITE LEAGUE BATTING AVERAGE OVER THAT PERIOD
(NOTE: THIS GRAPH GIVES ONLY A ROUGH APPROXIMATION OF THE TRUE PROBABILITY)



From Figure 3 it is understandable why there has been no .400 hitter in the major leagues since 1941. The composite major league batting average from 1942 through 1976 was .255; Figure 3 shows that a .255 league average maintained over a period of 70 league-seasons (35 calendar years) results in roughly a 18% chance of a .400 hitter appearing at least once during that period.

A TREND ANALYSIS OF BATTING AVERAGES

by Gary T. Brown

One of the million myths continually uttered by baseball followers is that good pitching always beats good hitting. Every October, after watching Schmidt, Murray, and Brett tear up the league, it is suddenly said that pitching wins pennants.

If that is true, then what happened in 1930 when National League teams batted an average of .303 and there were more baserunners than breadlines? In fact, from 1920-1930, neither the National nor the American League ever averaged under .279 per team. But that didn't continue. Since 1935, .270 is the best either league has been able to manage. Is it because of pitching? Or did pitchers get help? And if so, where are batting averages headed now?

Baseball, like most anything else, is a phenomenon enveloped in trends, much like a national government falls in and out of recessions. When things get out of hand, something is done to right the course. When a lowly team batting average of .255 was good enough to lead the American

Letters, cont. from page 3

ideally, appear on paper somewhere in the stats...." or something like that. All I can say is, we want back our son!

Mrs. B. Kwandrie
Fairmont, Minn.

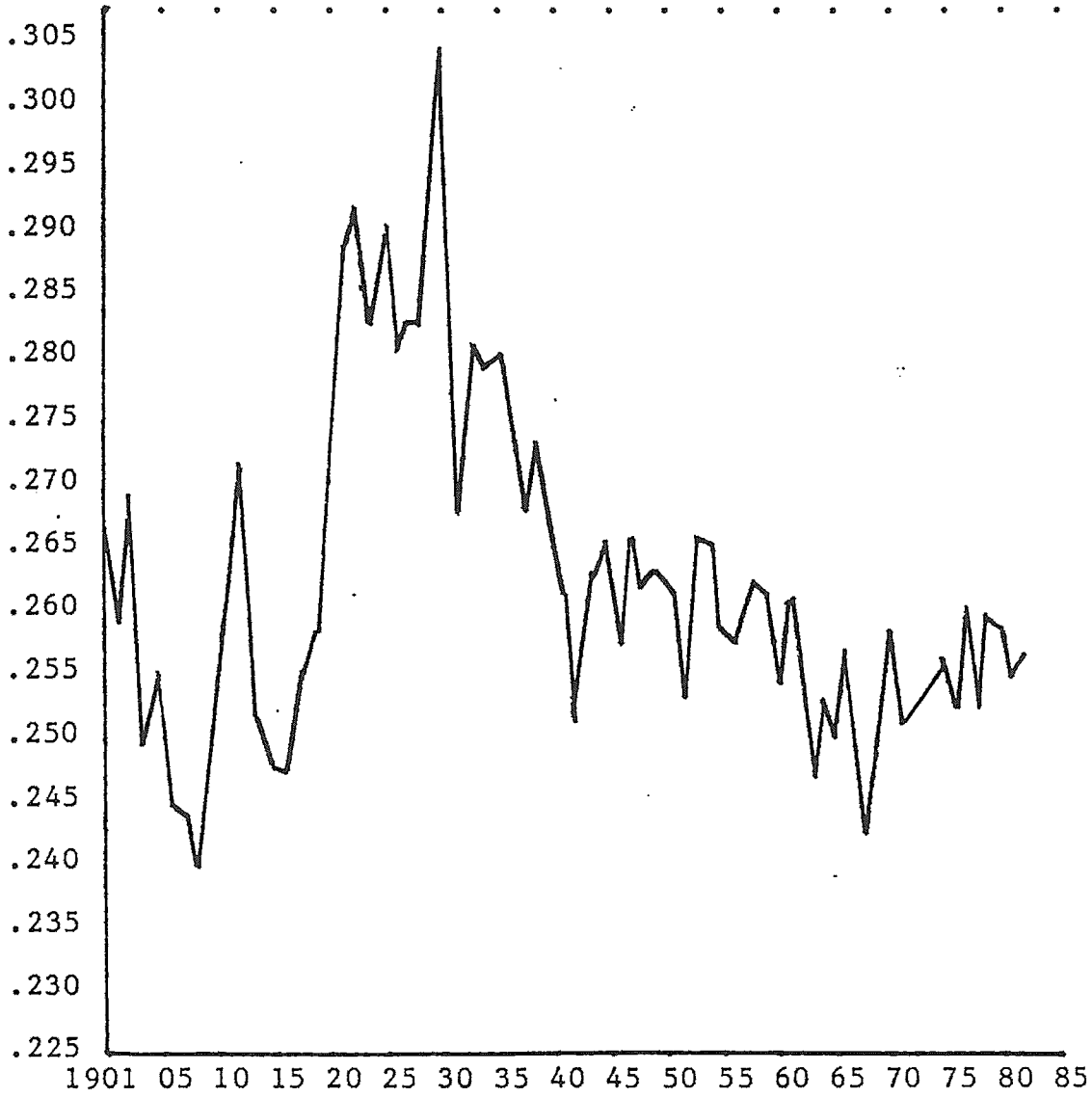
Editor's Reply: An American Tragedy to be sure. Another American tragedy is that letter above was the only one recieved for this issue! Baseball Analyst subscribers are not exercising their privilege of free speech and using these pages for their created purpose--as a forum for ideas and thought exchanges. Certainly you have some reaction to what goes on here! In future issues, we plan to run a letters page for the purpose of advancing sabermetrics Say you have an idea for a study but don't have the time to see it through, throw it out in the letters column and perhaps another reader will pick up on it. Propose joint projects; ways to improve an existing study; compliment contributors; speak your mind. Of course, we'd rather have full blown studies and articles, but a letter is a good start. We'll say it again--this is your magazine.

League in 1972, the rules committee decided it had seen enough and introduced the designated hitter, and presto, averages rose a record 20 points per team in 1973. Yet batters have never prospered like they did way back when. It is only once in a great while now when .400 is threatened, whereas in the market-crash decade, that rare plateau was shattered eight times.

Let's take a look at league batting averages from 1901-1982 and see what happened.

AMERICAN LEAGUE AVERAGES

NATIONAL LEAGUE AVERAGES



Four distinctive patterns, or trends, jump out at you right off the bat, and these can be dated as follows:

1901-1919
1920-1945
1946-1968
1969-Present

For each of these trends there are several important factors that determined them and changed the way baseball was played at the time.

The first period, commonly known as The Dead Ball Era, has become a legend in itself. The same scuffed up ball was kept in play longer, the spitter was legal, and parks were not enclosed. St. Louis led the majors with a meager 39 home runs, but finished fourth in the senior circuit while the White Sox took the first American League crown aided by 280 stolen bases. By contrast, the 1961 Yankees ravaged the league with 240 roundtrippers, yet stole but 28 bases. Things do change.

League averages soared in the Twenties with a livelier ball and the construction of hitters' parks in St. Louis, Detroit, and Brooklyn. The Major League home run average jumped from 28 per team in 1919 to a whopping 98 in 1940. Meanwhile, stolen bases took a 130-59 slide in the same period of time. Clearly, it became preferable to trot home than to swipe it.

Pitching endured baseball's batting average boom, however, and by 1945, the end of the second "era," hitters had cooled to a .259 clip. But home runs were still on a gradual rise, peaking at 157 per team in the National League in 1955, and at 155 per tem in the American in 1964. All the while, batting averages were dropping to a dismal .236 per team in 1968, for a number of reasons.

Perhaps most importantly, night baseball began dominating the game by the mid-Sixties when over half the schedule was played under the lights, when Abner Doubleday became Abner Double-twi-night-twin-bill, and pitchers took advantage of decreased visibility. And, much to the batter's chagrin, the strike zone was enlarged in 1963 in order to

"speed up the game." That was accomplished, as 1,681 fewer runs were scored that year and 297 fewer balls left the park. And speaking of parks, from 1960 to 1972, eleven new ones were built in the National League alone and had fences farther away and dimensions more symmetrical than anyone was used to. No more Polo Grounds, no more Ebbets Field, no more Braves Field, and no more Baker Bowl. The only thing that remained was the home run swing, which was missing contact more often than not (1,206 more K's in '63), partly because of the intimidating distances, and partly because of the improvement in pitching itself.

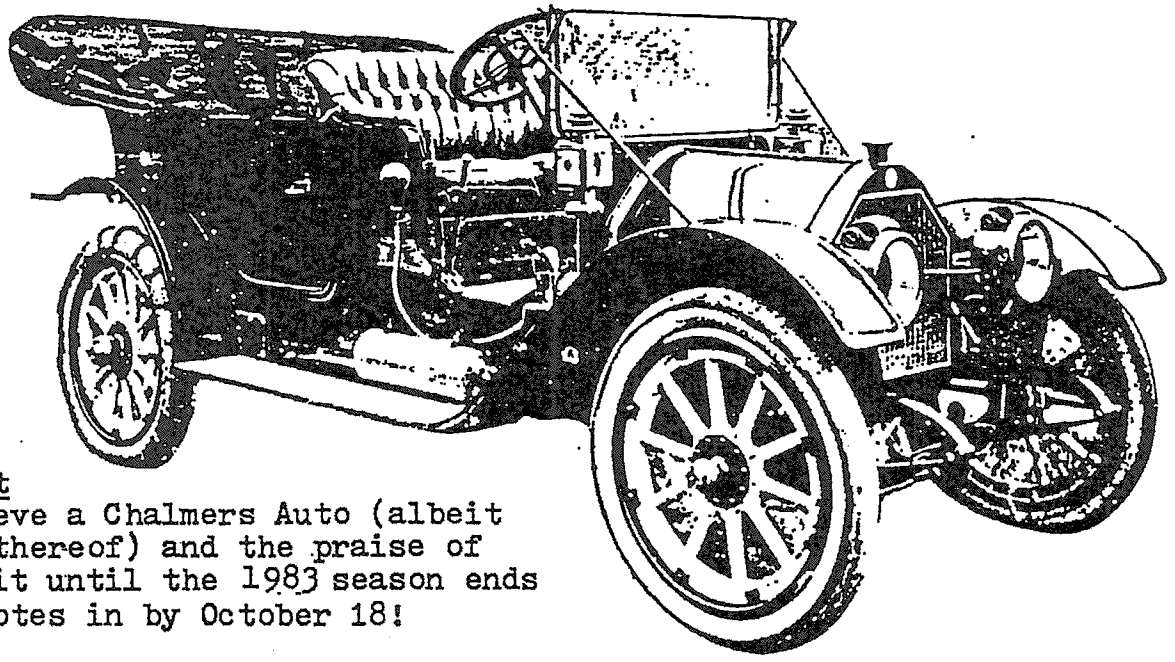
Pitchers began refining their talents at a much earlier age, and pitches that were once unheard of, palm balls, sliders, knucklers, and fork balls became standard weapons mastered by kids no older than Jackie Hernandez' career home run count. This not only crossed up the batter, but allowed for a greater reliance on relief pitching, which has now become a specialty in itself.

It wasn't until 1969 that batters made a significant reemergence, and in that year, baseball expanded, bringing more inferior pitchers into circulation, and the strike zone was reevaluated back to normal, which helped boost batting averages up by 38 points from 1968-1980 in the American League, and by 17 points in the National. This latest trend, then, leads one to believe that averages will continue to rise. For how long is another matter, because historically, there has been a series of manmade checks and balances (enough to make Adam Smith cringe) to "control" increases and decreases in batting average. Night baseball, increased

concentration on defense as well as better defensive equipment, and a larger repertoire of pitches work against the hitter, but those checks are supposedly balanced by a livelier ball, the designated hitter rule, and artificial turf (which we'll get to later), things designed to be in the hitter's favor.

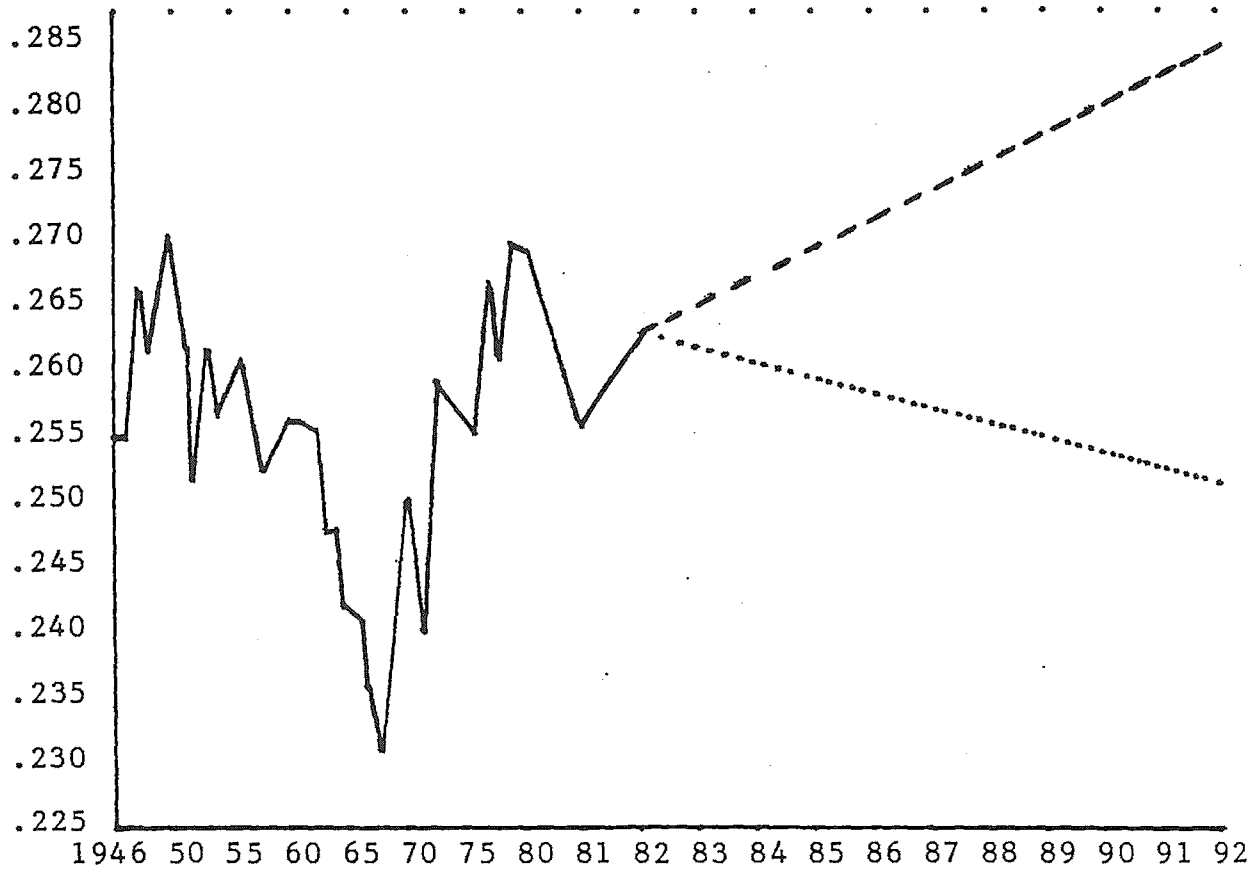
So how can we predict future batting averages? One thing we can do is use the batting average graphs presented earlier to draw up a trend line analysis that will project batting averages in each league for a designated number of years (we'll use ten for now, as trends don't last forever). A trend line analysis, sometimes called a linear regression, plots a "best fitting" straight line through several known data points. From this line we can determine the projected batting average for any given year. Let's do two projections, one starting from 1946 (dotted line in graphs below), which will include two trends in baseball, and one from 1969 (dashed line), baseball's latest trend period.

THE CHALMER'S AWARD BORN ANEW!

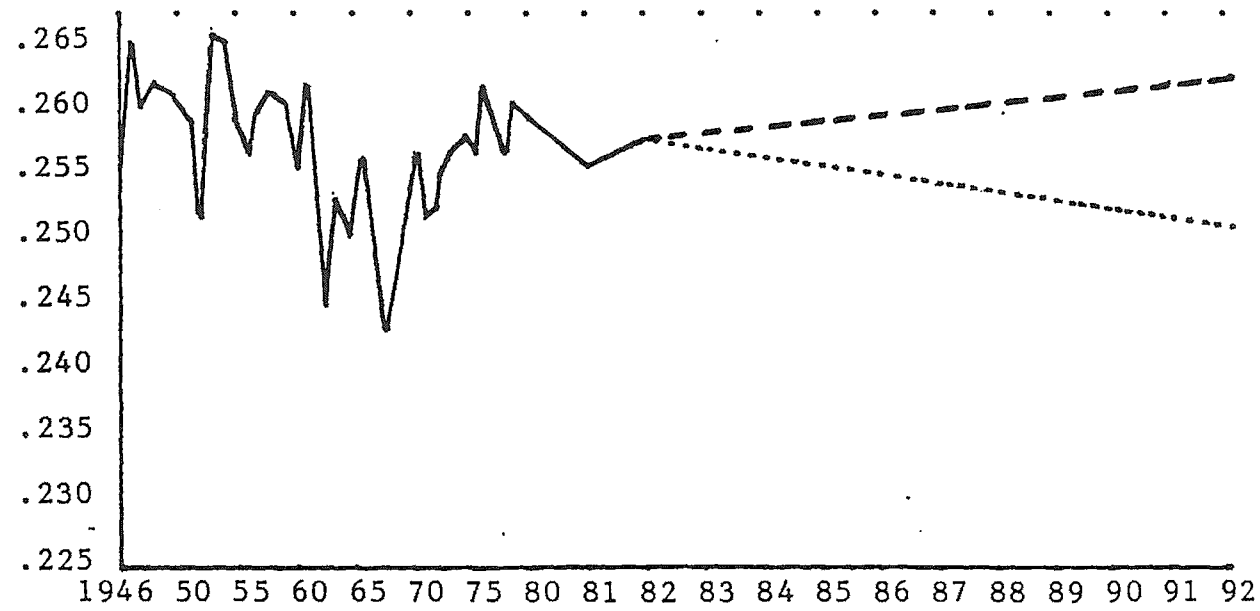


That's right! The famous award of yore is being resumed by the Baseball Analyst and its readers are going to pick the recipients. A simple postcard with your choice for the AL and NL MVP will suffice. Send to:
Chalmers Award
925 Kentucky Street
Lawrence, KS 66044
ATT: Baseball Analyst
The winner will recieve a Chalmers Auto (albeit a miniature facsimile thereof) and the praise of his peers. Please wait until the 1983 season ends and try to get your votes in by October 18!

AMERICAN LEAGUE BATTING AVERAGE PROJECTIONS



NATIONAL LEAGUE BATTING AVERAGE PROJECTIONS



Projections from the combination of the last two trends show averages declining gradually in both leagues, while the linear regressions from 1969 are, of course, on the up and up. The trend line analysis displays the slope of the overall increase or decrease within a given timespan, though its projections do not take into account any future checks and balances baseball may produce. Because the game is always changing, predicting batting averages becomes a matter of predicting factors that will affect them.

In order to attempt to do that, let's take a look at what factors are currently prevalent in today's rising averages. In the chart below, league batting average leading teams are given with their finishes in home runs, doubles, and stolen bases. What this does is help determine how these clubs are scoring runs, for even though good pitching may beat good hitting, runs still win ballgames.

LEAGUE BATTING AVERAGE LEADERS SINCE 1969

American League						National League					
Year	Team	Avg.	HR	2B	SB	Year	Team	Avg.	HR	2B	SB
1969	Minn	.268	4	1	4	1969	Pitt	.277	7	4	5
1970	Bos	.262	1	1	10	1970	Pitt	.270	7	4	8
1971	Balt	.261	4	5	8	1971	StL	.275	9	2	1
1972	KC	.255	10	2	4	1972	Pitt	.274	6	1	9
1973	Minn	.270	7	1	7	1973	Atl	.266	1	4	7
1974	Tex	.272	10	11	5	1974	Pitt	.274	4	2	11
1975	Bos	.275	4	1	10	1975	StL	.273	11	4	5
1976	Minn	.274	8	4	5	1976	Cin	.280	1	1	1
1977	Minn	.282	10	2	6	1977	Phil	.279	2	5	4
1978	Mil	.276	1	3	7	1978	Chi	.264	11	10	7
1979	Bos	.283	1	1	14	1979	StL	.278	9	1	7
1980	KC	.286	9	3	1	1980	StL	.275	8	1	9
1981	Bos	.275	5	4	14	1981	Phil	.273	3	4	3
1982	KC	.285	9	1	4	1982	Pitt	.273	3	1	3

Of the 28 leaders. 15 finished first or second in doubles, which not only tends to support the theory that a big power swing won't up the ol' average, but more importantly, it points to a major phenomenon that came around in '66 at that domed place--artificial turf. Of the 15 clubs that finished high in their respective leagues in doubles, eight have turf in their home parks. And of the eleven turf parks in the majors today, seven favor the hitter. Ironically, Houston, the perpetrator of all this, is not one of them, but is the best pitcher's park in baseball. Other turf parks, such as Candlestick, and Olympic Stadium, are for the most part cold and windy, favoring the pitcher. Yet in each of these three pitcher's parks, triples are up over 20%. At least six of the other parks show an increase in doubles. Only a handful of the turf parks are conducive to home runs, and two, Seattle and Minnesota, are crackerboxes.

The introduction of synthetic turf, therefore, has induced a change in the game, which shows up when the current batting average trend is analyzed. Averages are on the rise because teams are tailoring their personnel to fit their parks, and there just aren't that many good home run parks anymore. Good contact, line drive hitters are in demand. Note in the above chart that only five league batting average leaders won the home run crown as well, and three of those were Boston, Boston, and Atlanta, or Fenway, Fenway, and Fulton.

Accompanying the batting average rise has been the return of emphasizing team speed (particularly for turf

teams), and the stolen base. National League teams went from averaging 68 stolen bases in 1969 to 153 in 1980. American League teams went from 86 in '69 to 140 in '76. And in 1982, turf teams averaged 133 stolen bases while grass teams managed only 114.

It is a safe bet that increases in batting average will continue for the rest of the decade. The American League lacks dominating pitchers, some say because of the DH, and the National League will soon lose Carlton, Jenkins, Seaver, Ryan, and Rogers to age. It is also a safe bet that more and more clubs will change to artificial turf, for it was introduced for economic reasons in the first place, and economics always take precedence over tradition. Imagine synthetic turf in Fenway.

This is not to say that pitching is kaput. It isn't. Trends have always ended somewhere along the line, and this one is no different. Baseball would not allow either hitting or pitching to dominate the game for very long. It never has. Look for league averages to flirt with the .275 mark in the near future (the AL is already close), but don't expect 1930 to reappear. Pitching has improved at a faster rate than hitting since that time, and batters are just now beginning to catch up.

ASSIGNING RELATIVE VALUES TO RELIEF WINS, LOSSES AND SAVES

by John Schwartz

Two methods of rating relief pitchers appear periodically during the baseball season in The Sporting News. Their "Fireman of the Year" is simply the reliever in each league who records the highest total of relief wins and saves. The Rolaids Relief Award assigns a value of +2 for each relief win and save, and a -1 to each loss in relief.

I have examined the frequency with which relief wins, losses and saves are recorded by a large sample of leading relievers. The results suggest different point values than those used by either The Sporting News or the gas company. However, my main purpose here is not to establish a new rating system, but merely to assign relative values. Using the 1982 edition of the McMillan Encyclopedia, the first step involved compiling lists of the the top 130 pitchers in saves and relief wins. There were 130 pitchers with 48 or more saves, and 137 with 31 or more wins. (Ties for 130th place account for the additional pitchers.)

The ratio of 48 to 31 is about 3 to 2. Thus, based on the tables of leaders, the same number of pitchers have recorded 48 saves as have recorded 31 relief wins, indicating that saves are recorded about 50% more frequently than relief victories. The next step involved calculating the relief winning percentages of the leaders in relief victories, and pinpointing the median. This value (69th of the 137) was .5676 (Murry Dickson's, based in his 42-32 record). The median percentage is very close to the fraction 4/7 (.571). Thus, among these relievers, there are about 3 relief losses for every 4 relief wins, on the average.

Since saves are recorded 50% more often than relief wins in this sample, there are 6 saves for every four wins in relief. So the ratio of relief W-L-S is approximately 4-3-6. Dividing each of these numbers into 12, their least common multiple, gives a set of point values of W: +3; L: -4; S: +2.

These are, of course, empirically derived values, and, as such, subject to external influences, such as past changes in the definition of "save." They might be affected by number of relievers tabulated or by future changes in the way relievers are used.

SYSTEM COMPARISONS--Relative Values

(Records as of 8/7/83)

	W	L	S	Rolaids	TSN	Relative Value
American League						
Quiz	5	1	28	65	33	57
Stanley	7	7	21	49	28	35
Lopez	7	4	16	42	23	37
Gossage	9	3	12	39	21	39
Caudill	2	7	21	39	23	20
National League						
Bedrosian	7	4	16	42	23	37
Reardon	5	5	16	37	21	27
Holland	6	0	12	36	18	42
Smith, L	4	7	17	35	21	18
Orosco	9	5	11	35	20	29

DISTRIBUTION OF RUNS

Pete Palmer

The distribution of runs was first studied by Lindsey in three fine works published around 1960. Two articles were in Operations Research, the journal of the Operations Research Society of America, volume 7, number 2 and volume 11, number 4. The other was in the Journal of the American Statistical Association, volume 56. These articles were combined and updated in a book called Optimal Strategies in Sports, published in 1977 by North-Holland of New York. More recently, Dallas Adams has done some fine work published in the Analyst, notably issues one, four and five. A book published by A. S. Barnes in 1970, Player Win Averages, by Mills and Mills, used the probability of winning the game from various inning, base, out and score situations to rate major leaguers from play-by-play data for the 1969 season.

My own work as uncovered two useful facts about the distribution of runs. First, the variance of the distribution is equal to twice the mean. A standard distribution found in any texbook is the Poisson, in which the variance is equal to the mean. This works very well if events occur singly, such as goals in a hockey game. Knowing the mean of the distribution, any term can be found.

$$P_n = e^{-\mu} \frac{\mu^n}{n!}$$

where n is the term desired, μ the mean, e is 2.7183, the base of natural logarithms, and n! is n factorial. However, for runs in baseball, the variance is twice the mean and Poisson doesn't work. So the second is a modification of the Poisson which works very well. I discovered it in a book by Moroney, called Facts from Figures, published by Penquin Books in 1951 and revised through at least 1963. Anyway, all you need is the mean and variance for the desired distribution, indicated by μ and σ^2.

$$\mu = p/c \quad \text{and} \quad \mu + \mu/c = \sigma^2$$

where p and c are constants to be used later. For the case where the variance is twice the mean, c is 1 and p is μ. The distribution is defined as

$$P_0 = (c/(c+1))^p$$

$$P_1 = P_0 \cdot p/(c+1)$$

$$P_2 = P_1 \cdot (p+1)/(2\cdot(c+1))$$

$$P_3 = P_2 \cdot (p+2)/(3\cdot(c+1))$$

$$P_n = P_0 \cdot (p(p+1)(p+2)\ldots(p+n-1))/(n!(c+1)^n)$$

In cases where the variance is not twice the mean, you simply use different values for p and c. However, the original assumption works pretty well. Looking at the figures in Table 2 of Dallas' article in issue one of the Analyst, he has broken down run distribution into eleven categories by means.

cat	mean	variance	ratio	cat	mean	variance	ratio
1	2.91	5.96	2.05	7	4.35	8.89	2.04
2	3.18	6.01	1.89	8	4.60	9.73	2.12
3	3.37	6.45	1.91	9	4.87	10.41	2.14
4	3.61	7.28	2.02	10	5.16	11.14	2.16
5	3.86	7.79	2.02	11	5.29	11.03	2.09
6	4.12	8.57	2.08				

This information is useful when analyzing run samples. For example, let's find the expected difference between runs scored by a team in successive years if nothing had changed. A typical season total might be 700 runs. The variance in each sample would be 1400, meaning the standard deviation (the square root of the variance) would be about 37. However, when comparing two samples, the total variance is equal to the sum of the individual ones, or 2800. This means that the standard deviation between seasons would be 53 runs. That is about two-thirds of the time, you would expect the difference to be equal to or less than that if the team had not changed, and that five percent of the time, the difference would exceed two sigma, or 106 runs. Thus if a team scored 106 more runs one season than the other, you could be 95 percent sure that the difference was not due to chance. I believe these differences are larger than one might have thought they were without doing the analysis.

When comparing player batting averages from one season to the next, the binomial distribution can be used. Here the variance can be easily stated as pq/n, where p is the probability of success (i.e. batting average), q the probability of failure (equal to 1 - p) and n the number of samples. For a .300 hitter with 600 at-bats, the variance is .3 · .7/600, the square root of which is .019. Again, when comparing two samples, the variance must be doubled, the standard deviation therefore multiplied by SQRT(2), which gives .026. So this means that if a player moves fifty-two points in his average from one season to the next, there is still a five percent probability that nothing has changed. Again, I believe that these variations are larger than one might have expected.

A computer search of all American League games from 1980 through 1982 shows an average number of runs per inning of .487. This includes partial innings in which the winning run was scored. There were 54320 innings and 26476 runs. The variance of the distribution was .999, or 2.05 times the mean. Using the modifed Poisson distribution, c and p were found.

$$c = \mu / (\sigma^2 - \mu) = .951 \qquad p = \mu \cdot c = .463$$

Solving the equations for the various terms shows a prediction of more one run innings than actually occurred. Solving for the case where the variance was twice the mean did not change the calculated values very much.

runs	innings	frequency	predicted frequency $\sigma^2 = 2\mu$	$\sigma^2 = 2.05\mu$
0	39572	.728	.714	.717
1	8056	.148	.174	.170
2	3781	.070	.065	.064
3	1668	.031	.027	.027
4	728	.013	.012	.012
5	291	.005	.005	.005
6	131	.002	.002	.002
7	60	.001	.001	.001
8	23	.0004	.0006	.0006
9	4	.00007	.0003	.0003
10	5	.00008	.0001	.0001
11	1	.00002	.00007	.00007