BASEBALL ANALYST

JOURNAL OF SABERMETRICS
February, 1989 Vol. #40

We have five articles this issue. Have a good season.

Editor and Publisher: Bill James Address: Box 327, Oskaloosa, KS 66066

CLUTCH SITUATIONS

by Tom Hanrahan

In recent years, there has been a proliferation of statistics which have attempted to describe how batters do in the clutch. The Game Winning RBI has become an official stat, and on televised broadcasts we hear no end of batting average with a) runners on base, b) runners in scoring position, c) runners in scoring position with two outs, d) you name some more. Are these numbers telling us something useful, or not?

Here is how I attempted to answer that question: I first took Pete Palmer's Potential Runs For 24 Base-Out Situations (The Hidden Game of Baseball, p. 153). I then made a model of expected probabilities for a typical hitter for one trip to the plate (see the notes at the end of this article). Then, for each base-out situation, I looked up the new potential runs that would occur for all of the possible batter results, and then calculated the DISPERSION between potential runs with good results (walks, hits) and bad results (outs). An article by Paul Pudaite in the Oct 88 Analyst suggested this method for determining clutch situations. That article outlines the necessary calculations, including all of the appropriate Greek symbols that my printer doesn't have, so I won't reproduce the calculations here. I think it's important to note that we are only discussing "run-dependent" clutch performance in this article. There is also "game-dependent" performance, which would require a chart for every combination of bases occupied / outs / inning / runs ahead or behind, and "pennant-dependent" performance, for which Bobby Thomson, Carlton Fisk and now Kirk Gibson are famous.

An example: Suppose our batter model was the ultimate Ron Kittle / Dave Kingman - either he hits a home run, or he strikes out. The first situation is bases empty, two outs. The expected number of potential runs for this inning (from The Hidden Game's table) right now is .095 - we expect, on the average, slightly less than 1/10 of a run to score from this point until the third out is made. After a home run, we get 1 run home, plus we still have no one on base and two outs, so the expected total runs is now 1.095. After an out, the expected runs is zero, since the inning is over. The range from good to bad is 1.095 minus .000 = 1.095 runs. If we had the bases loaded with 2 outs, the (grand slam) home run would be worth 4.095 runs more than an out. Thus, in this situation, the difference between a tater and a KO is almost 4 times as great as in the original scenario. The following table gives the dispersions for each base-out situation in terms of standard deviation of expected runs.

DISPERSION OF POTENTIAL RUNS BY BASE-OUT SITUATION

Runners on bases			# of outs- 0	1	2	weighted
FIRST	SECOND	THIRD				average
-	-	-	.315	.252	.189	.264
X	-	-	.593	.495	.403	\
-	X	-	.504	.509	.534	-.503
-	-	X	.440	.524	.568	/
X	X	-	.841	.799	.743	\
X	-	X	.689	.800	.772	-.784
-	X	X	.626	.762	.933	/
X	X	X	.935	1.111	1.138	1.100

CLUTCH SITUATIONS

These numbers were calculated using the weighted probabilities of an "typical" batter - see the notes for details. The right-most column gives averages, weighted according to likelihood of game situations, grouped by the number of runners on base. The larger the number, the greater potential difference in run scoring opportunity there is in that situation.

OBSERVATIONS

The difference in the two extremes, bases full versus empty with 2 out, is 1.138 / .189 = about 6 times the amount of dispersion in potential runs scoring. Of course, you probably figured that sacks full and two out was the spot you wanted your best man up to bat; but sacks full with nobody out really isn't that far behind, according to the chart. In fact, there are no situations in which having two men on is more volatile (equals important) than bases loaded, regardless of the number of outs; two on is always more important than one on, and one on more important than bases empty, REGARDLESS OF THE NUMBER OF OUTS. If you look at the right hand column, you'll see an almost perfectly linear (1-2-3-4) relationship between the dispersion of potential runs, and men on base plus the man at the plate. Hindsight being 20/20, this makes sense, because with none on the hitter is batting only for himself, while with the sacks full he represents the fate of 4 men.

CONCLUSIONS

The POSITIONING of the runners (3rd versus 2nd or 1st) is not nearly as important as the NUMBER of runners, and the number of OUTS is a minor factor.

I was surprised at the importance of many early-inning situations. How many NL managers would let their pitcher come to the plate with men on first and second and nobody out (potential dispersion of .841 runs), but would feel compelled to pinch-hit with the same bases occupied and two away (dispersion only .743 runs)?

Lineup selection: Because the leadoff man always begins the game by batting with nobody on, maybe his spot is just slightly less important than the 3rd or 4th spot. I don't know if I would use this to justify moving Mr. Boggs out of the number 1 hole, however.

If you're loooking for a stat to define clutch performance, I would say that the situations defined in "batting average with runners in scoring position with 2 outs" are not, at-bat for at-bat, all that much more important than just plain old "batting average with runners on base". And, since the sample size is much smaller, the former stat is far less reliable. I would offer that the best stat to measure "run-dependent" clutch performance would be to weight batting, on-base, and slugging averages by runners on base.

NOTES
Typical batter model: walks - .095 (9.5 % of all plate appearances), singles - .165, doubles - .040, triples - .007, homers - .023. Outs = .670, of which .120 are potential DP, .05 send runner 3rd to home, an additional .05 also send 2nd to 3rd, .05 also send 1st to 2nd, all others leave baserunners unchanged. Baserunners advance on the average 1.45 bases on a single and 2.25 bases on a double, depending on the # of outs.

HOW I SCREWED UP

By Gary Fletcher

This is a sad story. I have had two articles published in the Analyst based on computer simulations of baseball offense. I wrote the program code myself. It took a lot of work and I think I did a very good job. However, I made a mistake. The mistake was not part of the program, but had I not made it, I would have uncovered a programming error. That error, now discovered, has somewhat put the lie to both articles.

You see, due to memory limitations, I didn't want to be bothered with certain offensive incidents. Specifically, I didn't want to use HBP, IBB and SH at all. Well, SH do occur naturally within the simulation, but they were not "seeded". HBP I just ignored and IBB seemed to me to be a hopeless barrier. To consider intentional walks in a different way from unintentional walks would have required an unusual amount of code designed to recognize IBB situations - situations which are, after all recognized in different ways by different teams. I decided not to use them, to treat them as ordinary walks.

But I couldn't see building a simulation that didn't consider the double play. I intended to check the accuracy of the program by comparing the results to the runs created estimate of the simulated stats. Unfortunately, the runs created version that considers GDP also considers HBP, SH, etc. What to do?

I tried just adding GDP to the A factor of the formula. What was, in the stolen base version,

$$((H+W-CS)\times(TB+(.55\times SB)) \ / \ (AB+W)$$

became, instead,

$$((H+W-CS-GDP)\times(TB+(.55\times SB)) \ / \ (AB+W)$$

which doesn't work too good. I fiddled with it, adjusting the value of GDP until I got something that seemed to work. I just multiplied GDP by .25 and, checking this on a calculator with a few team stats from one of the Abstracts, was satisfied that the thing worked.

Well, here's the first mistake. When checking, I used the Abstract division boxes, which, until 1987, listed TBB and IBB as BB and IBB. For some reason, I thought that BB meant unintentional walks only, so I added IBB back into TBB. The GDP adjusted runs created formula generally overestimates by about 7% on average, but with the extra walks factored in, seemed to be right on the mark.

Part of it was bad luck, I think. Probably I checked the formula against too few teams. If I had been more diligent, I would likely have uncovered my error. Bad luck, or laziness? Anyway, my error.

Working on a new project, I uncovered the error and fired off an apologetic article. The next day, to my embarrassment, I found out that the article itself was in error. Although I had tested the formula version which used .25xGDP, the version which was keyed in to the simulation program was different. I had left it with GDP simply added to the A factor, with no adjustments.

Now, what does this mean? As far as I can tell, the program is now working with about a 5-7% error, instead of the 3% I had claimed. The second article, on leadoff types, is totally invalidated. The assertion that players with high OB and SLG are underrated is made somewhat questionable, although I think not too much. On the other hand, that baserunning - taking extra bases on hits and outs - accounts for 50% of all runs scored still stands, virtually unaffected.

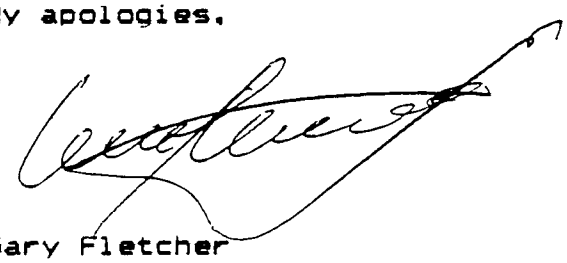
All this deserves reprimand and I am sorry. But, having calmed down somewhat, I now realize that what was really bad is that I failed to explain, in the first article, that I had used an adjusted runs created formula to check the simulation accuracy. Had I explained the adjustment, I am confident the editor would have fired the article back to me promptly, with a suggestion to revise both program and article. My only defense is that I keyed in the mistake rather early on in the game. I didn't work at the simulation continuously, not at all. Probably the whole thing took about a year before I considered it ready to roll(to give an insight into how I work, I would bet that there is no more than about fifty hours of actual work involved in the actual programming).

I forgot. I'm tempted to say it's unforgiveable. I hope not. Before I can be forgiven, I must apologize, and I do.

Sorry, everybody.

I don't think that simulations are without merit, not even mine if I should go back and revise it. As I said in a letter to one Analyst contributor, I had already lost my enthusiasm for the simulation method. Bill James comment on not doing big studies applies, somewhat, and this episode isn't going to rekindle the fire obviously.

My apologies,



Gary Fletcher

VALIDATING SIMULATION PROGRAMS -- SOME BENCHMARKS

Kenneth McLain

In Baseball Analyst #36 Gary Fletcher discusses his simulation program and asks and answers various questions about what he includes in the simulation, etc. One of his questions is "How do you know if the simulation is accurate?" His answer is to compare the average runs scored from running the simulation many times to the value given by the runs created formula. While this is a reasonable check the runs created formula itself is an approximation and is not infrequently off by 30-40 runs for a team-season. How do you separate possible simulation errors from runs created inaccuracies?

I have analyzed four special cases using basic probability theory and give the results below. <u>These results are exact</u>! The results from a simulation program for these special cases should approach these results in the limit of many innings. If not, then something is wrong either with the simulation program or with the random number generator.

These four cases are very simple cases and should be valuable benchmarks. Because there are many things they don't include they should be used in addition to and not in lieu of Fletcher's suggested check against runs created.

For these four cases I obtain equations for the expected number of runs scored per inning in terms of the team batting average. Implicit in the derivation of these equations is the assumption that the lineup is homogeneous -- all batters have team-average statistics. I give only the resulting equation for each case and not the derivations (straightfoward but somewhat "messy" application of probability theory).

In each of the four cases p is the probability of getting a hit for a given at-bat, i.e. p is the batting average (it is assumed that each batter either gets a hit or makes an out). R is the expected (average) number of runs per inning.

CASE #1: ALL HITS ARE HOME RUNS

$$R = 3p/(1-p)$$

CASE #2: ALL HITS ARE EITHER DOUBLES OR TRIPLES

$$R = 3p/(1-p) - 1 + (1-p)^3$$

CASE #3: ALL HITS ARE SINGLES AND ALL RUNNERS ON 1st SCORE ON TWO HITS

$$R = 3p/(1-p) - 2 + (2+3p)(1-p)^3$$

CASE #4: ALL HITS ARE SINGLES AND ALL RUNNERS ADVANCE ONE BASE ON ALL HITS

$$R = 3p/(1-p) - 3 + 3(1+2p+2p^2)(1-p)^3$$

If the results from a simulation of any of these cases do not agree with the results from the above equations (within statistical variation), then something is wrong with the simulation program or with the random number generator.

Obviously there are programming errors which will not affect the results for these four cases and hence will pass through this test undetected. Two examples are errors in accounting for stolen bases and in accounting for advancing a base on an out.

Again I emphasize that I am advocating using these test cases in addition to and not in lieu of comparing results with runs created.

Fletcher discusses the value of taking an extra base on hits and concludes that if extra bases are not taken runs scored will be reduced by approximately 50%. My CASES #3 and #4 give the same comparison when all hits are singles. For p=0.300 the results are R(3)=0.280 and R(4)=0.117. These results appear to be consistent with Fletcher's.

THE HOME FIELD ADVANTAGE REVISITED

by Jim Heg

After the Boston Red Sox posted an abysmal 1987 road record of 28-54, manager John MacNamara announced during the off season that his players would not be allowed to take their golf clubs on the road during the 1988 season, to improve their concentration. During April 1988, the Crimson Hose posted a 6-2 record away from Fenway, and Mac thereupon rescinded his ban on road golf. Of course, the Sox proceeded to mount a road tally of 30-43 the rest of the year. Mac was knifed along the way, but the Sox won the division with the worst road record in history for an AL East Champion.

Few if any of the record splits in baseball are as important as the home field advantage. Yet few are more elusive. Although both home winning percentage and road winning percentage are strongly correlated to overall team winning percentage, the correlation between the latter and the advantage, or difference between home and road performance, is very weak. Lacking reliable access to a home computer, I was unable to run a complete data sample. However, I did look at all AL East clubs, going back from 1987 to the latter of either 1946 or the team founding year, for a total sample of 240 team/years (those of you who are interested can check other teams/years). I found that the correlation coefficient between winning percentage (p) and home winning percentage (h) was .888, while the correlation coefficient between road winning percentage (r) and p was .871. However, the correlation between p and (h-r), or home field advantage, was only a measily .081. I believe that few areas of baseball contain more mysteries than the home field advantage, and below is a potpourri of musings on this subject.

The 1984 Bill James Study

In the 1984 Abstract (p.252), our distinguished editor reported that the home field advantage tended to increase with the winning percentage of the team, and to do so out to the largest observed winning percentages. (I must confess, I always liked the parts of the Abstract best that our editor apologized about the most profusely). He invited the readers to suggest a simple function which would reflect this, and yet bend back before reaching the maximum theoretical winning percentage of 1.000, where (h-r) must equal zero. No doubt many mathemeticians took him up on this, but I have yet to see anything published. My own favorite candidate to solve this question is a function of the form:

$$h = 1 - ((1-p)\#k)$$

(or (1-h) = ((1-p)#k)) where # in this instance stands for "to the power" and k is the exponent. The road winning percentage is given by the mirror equation r = 1-((1-p)#(1/k)). When k is greater than 1, the team will win more games at home than on the road, and vice versa. One can experiment with various values of k, to find one which satisfies the characteristics of the actual data:

p	h-r, k=1.10	h-r, k=1.13	h-r, k=1.15	h-r, k=1.20
.800	.0613	.0785	.0896	.1165
.750	.0660	.0844	.0965	.1255
.700	.0687	.0881	.1006	.1309
.650	.0699	.0896	.1014	.1332
.600	.0697	.0894	.1022	.1358
.500	.0660	.0846	.0967	.1259
.400	.0584	.0749	.0839	.1116
.300	.0476	.0610	.0698	.0911
.200	.0340	.0437	.0500	.0652

The value k = 1.13 appears to satisfy most of the Bill James specifications, causing the home field advantage to reach a peak value of .090 in the range p= 650, and to begin to bend back inward in the range of p= .700.

To check the theory against actual data, I ran the regressions:

$$\log(1-h) = (k)*\log(1-p)$$

and

$$\log(1-r) = (x=1/k)*\log(1-p)$$

without intercept for all major league teams during 1984-87. (As Sparky Anderson undoubtedly could have told you until August of 1988, when he lost his genius status, the natural log of a number to an exponent is the exponent times the log of the number). Each of these regressions had 103 degrees of freedom. The estimated value for k was 1.130846 (t= 91.6, R2= .6756). This comes very close to what the theoretical equation would suggest just by eyeballing the graph for various k values. The estimated value for x= 1/k was .888451 (t= 87.9, R2= .7594). Taking the reciprocal 1/x= k yields k= 1.125554. very close to our original estimate for k. So, it looks like we might have a good model linking the W/L percentage to the home field advantage.

However, this again is where the elusive nature of the beast comes to the fore. Suppose we construct an estimate d for the home field advantage by taking the difference between the estimates of home and road winning percentage developed using the regression equations above. The correlation coefficient between this estimate and the actual home field advantage for the 104 major league team/years 1984-87 was -.17078. That is, the higher the estimate,

the lower the actual home field advantage tended to be. This is polymorphous perversity at its worst.

Part of the explanation lies in the fact that, for the 1984-87 sample, while both actual h and actual r were highly correlated (in the range of 0.9) with their respective values estimated using the above equations, the direct correlation between h and r was only .4502. Also, when you think about it, deviations of actual h and r from the mean, for a given p, will always have opposite signs. Hence, taking the difference between two estimates for h and r will cause errors to balloon. However, maybe something else is going on. What if the recent year sample is not consistent with the James results? About a year after doing the above study, I decided to check.

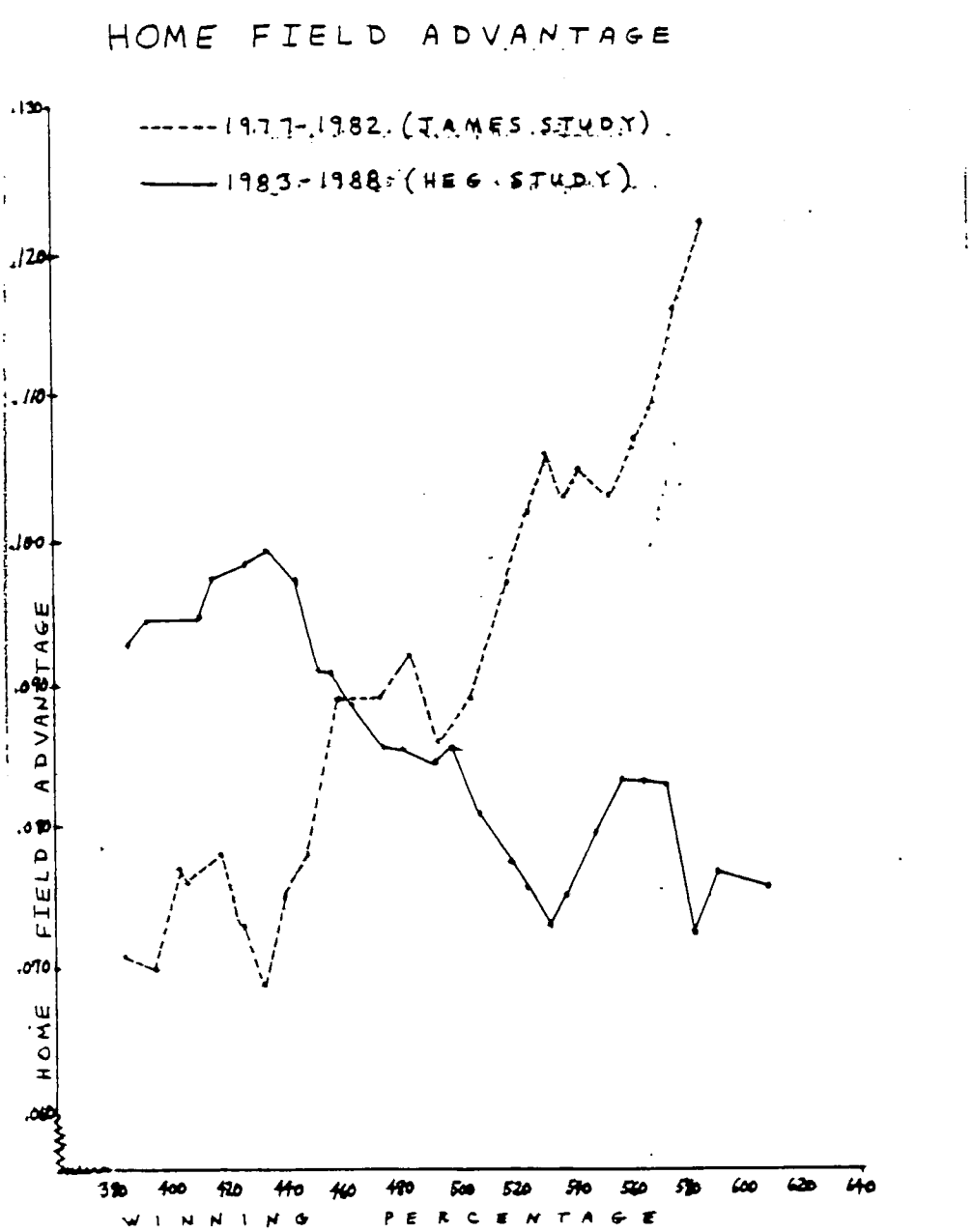
Home Field Advantage Revisited

I decided to redo the James study, using exactly the same methodology, but for the years 1983-88 rather than 1977-82 (excluding 1981). As you may recall, Bill James separated the data into groups of .010 won-lost percentage, and then, in a rolling fashion, looked at 24 megagroups of .100. I did the same thing, only I used 26 megagroups because the 1986 Mets posted a better W/L percentage than any major league team in the 1977-82 interval. The results are as follows:

	James Study 1977-82		Heg Study 1983-88	
Interval	Int W/L Pct	HFAdv	Int W/L Pct	HFAdv
.570-.669	N/A	N/A	.609	.076
.560-.659	N/A	N/A	.591	.077
.550-.649	.587	.122	.584	.072
.540-.639	.577	.116	.574	.083
.530-.629	.568	.109	.565	.083
.520-.619	.564	.107	.557	.083
.510-.609	.554	.103	.548	.079
.500-.599	.544	.105	.538	.075
.490-.589	.539	.103	.532	.074
.480-.579	.532	.106	.524	.076
.470-.569	.526	.102	.519	.077
.460-.559	.518	.097	.508	.081
.450-.549	.505	.089	.499	.086
.440-.539	.494	.086	.493	.084
.430-.529	.484	.092	.482	.086
.420-.519	.474	.089	.475	.086
.410-.509	.458	.089	.465	.088
.400-.499	.448	.078	.457	.091
.390-.489	.440	.075	.453	.091

Interval	James Study 1977-82 Int W/L Pct	HFAdv	Heg Study 1983-88 Int W/L Pct	HFAdv
.380-.479	.433	.069	.445	.097
.370-.469	.426	.073	.434	.099
.360-.459	.418	.078	.426	.098
.350-.449	.406	.076	.415	.097
.340-.439	.404	.077	.411	.095
.330-.429	.395	.070	.393	.094
.320-.419	.384	.071	.386	.092

I was astonished to see that the new sample completely reverses the previous finding that the home field advantage increases fairly uniformly with the won-lost percentage. The results of the 1984 study were about as clear and unambiguous as any sabermetric study that I have seen. Without any doubt, the results reflected reality as it was during 1977-82. So, what happened?



Frankly, I do not know. I do not believe that the relationship between W/L percentage and home field advantage is in fact random. Possibly, the reversal reflects some underlying structural change, much as reversal of the yield curve between short and long term interest rates in financial markets reflects a change in the structure of expectations. One preliminary conjecture: during the 1983-88 period, team positions in the standings have been much more volatile than in the 1977-82 period (I offer no proof here, but trust me). If we assume (one would have to check) that a team's road performance tends to be significantly more variable over time than its home performance, then the decreased (increased) home field advantage (in 1983-88 compared to 1977-82) of the best (worst) teams would be consistent with the increased (decreased) difficulty experienced by those teams in holding (improving) their performance level from one year to the next. This is just a thought, but whenever I get a computer, I will try and check it out.

Pythagoras at Home

Another peculiarity involves the relationship between the home/road split and the pythagorean estimation (not prediction !!!) of won-lost percentage from runs scored and allowed. The pythagorean estimation of home park winning percentage almost always falls short of the actual winning percentage. That is to say, the pythagorean equation produces a biased estimator of home park W/L percent (h). I looked at the American League from 1946-1987:

Average for Period	h W/L%	Std Dev h	Max h	Min h	Pyth Est	Diff
1946-1960	0.537	0.023	0.584	0.491	0.508	+.029
1961-1968	0.536	0.015	0.563	0.517	0.517	+.019
1969-1976	0.537	0.017	0.564	0.514	0.516	+.021
1977-1987	0.542	0.020	0.574	0.505	0.523	+.019
1946-1987	0.538	0.020	0.584	0.491	0.515	+.023

The AL home field advantage in 1988 was .543, but I don't know the Pythagorean projection yet. It's hard to believe, but there actually was a year (1953) when the American League as a whole registered a negative home field advantage. I imagine that was quite a unique event, although I do not have an easy way to check other league/years going back into baseball antiquity.

The above breakdown is based on the periods during which the number of teams in the American League was uniform. As is evident, the data are remarkably stable, although there is some tendency for the home field advantage to rise in recent years. The Pythagorean estimation for h consistently accounts for only about half of the

actual difference between league h and .500, and the estimator shows an average bias of .023 on the low side for the AL as a whole.

The reason for this bias becomes evident when one recalls that home teams do not bat in the bottom of the ninth inning if they are ahead. This introduces an asymmetry between home and road scoring. A team playing at home will skip some 40 odd innings of batting per season because it has already won the game. We can assume that the average team at home will score roughly 5% fewer runs (1/9)*(1/2) at home than it would if it went up to bat in the ninth inning regardless of the score. Indeed, adjusting the American League runs scored data upward by 5% for home games and then running the Pythagorean estimates of h again reduces the average difference between actual and estimated h to -.001 for 1946-87.

Of course, the Pythagorean bias washes out when road games are included because now the opposition's scoring is reduced by the same asymmetry. A team which wins a lot more games than it loses (or vice versa) could have this bias show up in the overall data. In practice, it would be very hard to tell this effect apart from other effects (such as the tendency for a team which greatly exceeds (undershoots) its Pythagorean projection to win (lose) a lot of games anyway)) which may be going on at the same time

At one time, I toyed with the notion that this home park asymmetry could account for the Bill James finding that the exponent 1.83 gives a more accurate Pythagorean formula than the far more elegant 2.00. However, there is no evidence that I can find of any association between home field advantage and errors in the Pythagorean projection. I now think the solution has something to do with the proportion of tight and/or low scoring games played, which in turn may relate to the overall level of scoring offense in baseball over time. That, however, is a subject for another study...

Conclusion

There remain a large number of interesting questions about the home field advantage which, given time and resources, one would like to investigate. For example, is the variance of team home W/L percentage significantly less over time than the variance in road W/L percentage? If so, will a team which significantly improves one year by improving its road performance be more likely to decline the next year than a team which improves via a better home performance? Is the variation over time of team offense (or defense) greater on the road than at home? If so, could erratic pitching on the road cause 75% of the ulcers among baseball managers? Could this be what is meant by "pitching is 75% of baseball"? The possibilities are endless...

On Why Teams Don't Repeat

1. Introduction

The assumption that luck is a factor in baseball performance allows us to deduce one reason pennant-winning teams tend not to repeat: they usually won their pennant with the help of a good deal of luck, and luck tends not to hold from year to year.

While random chance is a factor in helping any team to the record it posts, it is more likely, in retrospect, to have helped teams which posted good records than teams which posted bad records: and, conversely, teams which performed badly are more likely to have been hurt by bad luck than teams which finished .500 or better.

We can phrase this effect as follows:

> The majority of extreme achievements (good or bad) have been substantially aided by luck; that is, the talent that produced the achievement is not as good (or bad) as the achievement itself.

Rephrasing that as an example, of teams that win more than 100 games (an extreme achievement), most are teams that were not talented enough to win 100 games, but did so by luck. The same applies to players; of players who hit 40 home runs in a season, most were players who were not legitimate 40-home-run hitters, but just had a lucky year, and of players who have had ERAs of less than 2.00 in a season, most achieved that performance by a fair deal of luck: their talent was not alone enough to carry them to that mark.

In this essay, we'll deduce and quantify the above effect, thus giving a major reason why teams and players don't repeat: they just weren't good enough to achieve what they did in the first place.

2. Why

For this study, we assume that for every team (or player), there exists a talent level for that team which represents its probability of success. For example, we assume that a team with 100-game talent will, independently for every game, win that game with probability 100/162, and a player with 35-home-run talent will hit a home run in a particular at-bat with probability, say, 35/600. While these assumptions are not completely accurate — a 100-game team will have a better than 100/162 chance of winning against a poor team and less than 100/162 chance of winning against a good team — they are close enough and reasonable enough to not affect the conclusions that will follow. Also, when we refer to a lucky team, we are talking about one which has exceeded its talent (say, a 100-game talent that wins 102 games); when we say a team is unlucky, we mean it failed to realize its talent (a 100-game talent winning only 96 games).

Now, suppose a team wins exactly 100 games in a season. What is the expected talent level of that team?

Because of the effects of luck, that team is probably not exactly a 100-game talent (almost certainly not, considering that in theory, team talent can be something other than a whole number of games, such as 99.94 or 100.12). We're therefore looking at a team that is either (1) a less-than-100 game talent that got lucky, or (2) a more-than-100-game talent that got unlucky. But it is a fact that there are many, many, more less-than-100-game talents than there are more-than-100-game talents. There are probably several times as many teams capable of playing 95-game baseball than there are 105-game teams, for instance. So of teams that wind up with records of 100-62, there will be many more who were 95-game talents who got lucky than there will be 105-game talents who got unlucky, and the average talent of 100-62 will be less than 100 games.

The same reasoning applies to hitters; a .333 hitter. say, is either a .300-.332 talent who got lucky, or a .334+ talent who got unlucky. But the distribution of batting talent follows the tail-end of a normal curve: there are so many more talents worse than .333 than better that it is virtually certain that our .333 hitter would normally hit worse than .333.

Another way to look at it is that good luck moves teams towards the top, while bad luck moves teams towards the bottom. So, if you sample teams at the top, you'll find for the most part that they were lucky, because the teams that weren't moved to the bottom.

You can try an experiment with real data. Take a season's worth of actual runs / runs created data. Sort the teams by runs created, which, in a very rough way, represents the team's talent. Assign each team a "+" if it exceeded its RC estimate, and a "-" if it scored fewer runs than the prediction. The +'s and -'s will probably be distributed pretty evenly. Then, sort the teams by actual runs scored, which represents the achievement. Since the +'s move up and the -'s move down, you should notice the +'s are concentrated among the teams in the top half of the league. Again, the reason: luck moves mediocre teams up to the top among the genuinely good teams, while (bad) luck moves talented teams down to the ranks of the less talented.

3. The calculation

We have seen, so far, that the typical 100-game-winning team is actually a team of lesser talent than 100 games. But in order to conclude that this effect is the reason teams don't repeat, we need to determine what the expected talent of such a team is. If the average 100-62 team is actually of 90-72 talent, then we have a plausible explanation for its failure to repeat. If, however, such a team has average talent of 98-64, that doesn't explain why it fails to win even those 98 games the following year. To determine how responsible the effect is for failure to repeat, we need to get some kind of estimate of the magnitude of the effect, although we have already shown why it must exist.

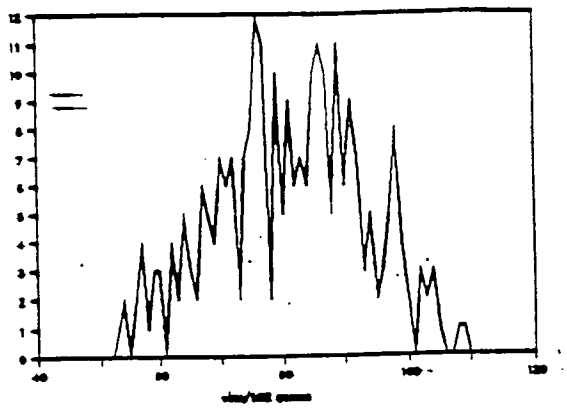
Graph #1 (next page) shows the distribution of team wins in the American League over the years 1961-1984 (1981 excluded). What may be hard to see from the graph, because of the irregularity between 75-80 wins, is that the distribution is roughly bell-shaped, but with a peak at around 85 wins rather than the peak at 81 we would find if the distribution were strictly normal. The peak occurs in the mid-80's because of a tendency by teams to stabilize at a level where they have a chance at 90-95 games and a pennant (see 1985 Abstract, Blue Jays comment). The exact number of seasons per win level are listed in the table on the last page. All team-seasons of more or less than 162 games (including 1972 seasons) were expressed in wins per 162 games.

The above distribution is that of actual wins, or achievement; it follows that the distribution of team talent is different. That's because the effect of luck is to spread out the achievements away from the talents. A 35-home-run hitter may hit 40 or 45 in a season if he has a lucky year, even if there isn't anyone in the league with strict 40 homer talent.

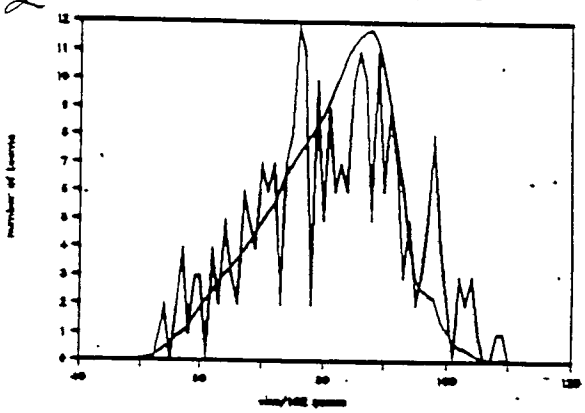
We can estimate the distribution of talent by finding a talent function that would theoretically produce the above team-win distribution when the seasons were actually played out (mathematical details are in the notes). I did this by taking a first guess, calculating, and then adjusting my guess until what I got looked like it approximated the above team-win distribution. What I finally settled on looks like graph 2 (superimposed on the actual team win distribution above).

A few things are immediately apparent from the graph: primarily, the talent distribution is clustered around the centre much more than the actual-win distribution. There are many teams who actually won more than 100 games in a season (far right of actual-win curve), but almost none who actually had the talent to do it. .Also, there are many more teams whose

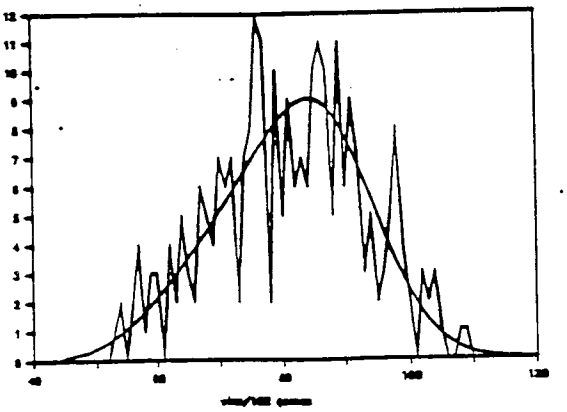
1



2



3



How I got most of this stuff

This should actually go at the end of the article, but I don't have room.

The first thing I had to do was come up with a talent distribution, t(x), where t(x) is the number of teams out of 264 with talent x games. For example, if we set t(80)=8.4, we're guessing that 8.4 out of 264 teams have talent of exactly 80 games.

Now, let g(x,n) = out of 264 teams, # of teams of talent x winning exactly n games. By binomial theorem,

$$g(x,n) = t(x) \cdot \binom{162}{n} \cdot (x/162)^n \cdot \left(\frac{162-x}{162}\right)^{162-n}$$

And so the total number of teams winning n games, which we'll call a(n), is given by:

$$a(n) = \sum_{x=0}^{162} g(x,n).$$

I chose t(x) in such a way that the above numbers a(n) matched the actual number of teams winning n games. (Compare columns B and C of chart.)

Then, to find the average talent of n-game-winning-teams, we just average g(x,n) weighted by x:

$$\text{average talent of teams that win } n \text{ games} = \frac{0 \cdot g(0,n) + 1 \cdot g(1,n) + \ldots + 161 \cdot g(161,n) + 162 \cdot g(162,n)}{g(0,n) + g(1,n) + \ldots + g(161,n) + g(162,n)}$$

talent is around 81–90 games than there are teams who actually won that many; again, random chance spreads the achievements away from the talent. The talent curve here is not perfect, though; there are probably theoretical reasons why the graph should look smoother and more bell-shaped. The conclusions of this study, however, should not be affected by this failing.

Again, the reason I settled on this talent distribution was that it predicted the actual distribution fairly well. Graph 3 is the actual-win graph superimposed on the predicted-win (predicted by the talent distribution, that is) graph. It seems to fit fairly well, considering that under our assumptions, the predicted-win curve must be round and bell-shaped. The numerous peaks and valleys in the actual-win curve are almost certainly caused by chance.

If there is significant error in the curves, it probably occurs in the far left (50–65 wins) and far right (97+ wins) portion. Because there have been so few teams to finish with very low or very high winning percentages, the actual-win curve is very jagged at these points, and no theoretical curve can match it very well. You might move my talent curve up or down at either end if you feel my theoretical-win curve lies too low or high with respect to the actual. Such changes, though, won't substantially change the results presented here, I think.

Again, numeric values are presented in the table. Just to read off a few figures, of 264 American League teams between 1961 and 1984, 5 finished with seasons of exactly 97 wins (after normalizing to 162 games and rounding to the nearest game). The talent curve says that of these same teams, only 2.4 had talent of exactly 97 wins, and the theoretical distribution predicts that 3.9 teams "should have" finished with exactly 97 wins. Also, 3 teams finished with records of 60–102, while only an estimated 1.9 teams had exactly that talent. And so on.

The talent distribution shown above and in the table is the main result of this study; all results and conclusions to follow are based on this talent distribution.

4. The results

Having obtained an estimate of the talent distribution, we can derive the average talent of a team with a given record. Some of the results (full results in the table):

Teams winning 85 games have average talent of about 85 games.
Teams winning 90 games have average talent of about 88 games.
Teams winning 100 games have average talent of about 93 games.
Teams winning 108 games have average talent of about 98 games.

The results substantially support the premise that teams with extreme records are lucky — the average 100-game-winning team is actually only good enough to have won 93, while the average 108-game-winning team is only, on average, a 98-game team. What this really means -- I haven't stated it explicitly yet — is that if the same 108-game-winning team, with the same players, were to play the same season over again, it would, on average, win only 98 games, 10 less than it actually did. It means that the team exceeded its talent by 10 games not because of leadership, nor heart, nor managerial skill, but only because of random luck.

Of course, when we say that the average 108-game-winning team is actually a 98-game-winning talent, we're not saying that's true for <u>every</u> team that wins 108 games. Some 108-game winners may actually be 108-game talents, some may be 95-game talents, and some may even be 85-game talents that were <u>extremely</u> lucky. All we're saying is that if you took a large number of teams that won 108 games, their <u>average</u> talent would be about 98 games. The actual calculation of how many 108-game winners are actually 108-game, 98-game, 85-game, etc. winners appears below (full data in table, again):

40-86	87-89	90-92	93-95	96-98	99-101	102-104	105-107	108-110	111+
0.9%	3.7%	11.0%	16.2%	24.3%	22.7%	14.6%	4.9%	1.8%	0%

Note that the chance a 108-game-winning team is actually a legitimate 108-game-talent is less than 2%, while a rather large 32% of such teams weren't even good enough to win 96 games. This result, again, strongly suggests that teams don't repeat because they were just lucky the first time.

5. Making a conclusion

We have established that the average, say, 100-game-winning team is around a 93 game talent, although it could actually be more talented than 100 games or less talented than 90 games. But for a particular team, how can you tell which it is? Were the '84 Tigers really a 104 game talent, or were they a 100-game talent or a 93-game talent?

Well, when a team wins more games than its talent indicates, there are at least five reasons why:

1. It produced more offense than its talent should have; that is, individual players got lucky and hit more home runs than they "should have", or more doubles, or more walks, or had a higher steal percentage, etc. This is the "career year" effect.

2. It produced more runs from its offensive components than it should have; that is, it exceeded its runs created estimate.

3. Same as #1, but for pitching: the team's opponents produced fewer offensive accomplishments than they should have. That is, the pitchers of the team in question had "career years".

4. Same as #2, but for pitching: the team's opponents scored fewer runs than their runs-created estimate predicts.

5. The team produced more wins than it should have given its number of runs scored and runs allowed; that is, it exceeded its pythagorean projection.

Numbers 2, 4, and 5 should theoretically be easy to check: just line up the particular team and see if it exceeded its projections, and by how much. You should be able to line up all teams in any period of years who won more than 100 games, and find that on average, (a) they exceeded their runs created estimates, (b) their opposition scored less than their runs created estimates, and (c) they exceeded their pythagorean projection. I did this for (c), and found that the AL 100-game-winners since 1961 beat their estimates by an average 3.5 games. (You have to use the 1.83 coefficient in the pythagorean formula, because it's otherwise inaccurate for very good or bad teams.)
For the '84 Tigers, (a) they actually undershot their runs produced estimate by 14; (b) their pitching staff allowed 24 more runs than their estimate; but (c) they overshot their pythagorean projection (exponent 1.83) by five-and-a-half games. So far, then, we estimate their "luck" as having been about a game and a half (5.5 from pythagoras minus a games for the 38 runs of bad luck).

For point 1 and 3, whether the players had career years, you have to sit down and figure out. I would work it by taking each player's age and career record prior to 1984, and estimating what a reasonable expected 1984 performance would look like (in fact, since this is 1988, we have post-1984 data to use in projecting estimated '84 performance) and comparing it to the player's actual 1984 performance. I won't go through that, because I'm not very good at that kind of thing, but you might look at the bench, Willie Hernandez, etc. Note that the effect is not limited to huge career year effects — if everyone on the team gets lucky by even just two home runs, you've got maybe three extra wins right there.

So, get your best estimate for #1 and 3, get your numbers for #2, 4, and 5, and add them up. That's your estimate of how lucky the team was. If you do that for a few different teams, you'll get an estimate of, when a team gets lucky, how the luck manifests itself. Does the luck show up, on average, as, say, 40% in runs created, 20% by Pythagoras, and 30% by the players, or what? Actually, it's probably roughly proportional to the standard deviations of the deviances from expected, so you could figure out which deviates most, expected pythagoras minus actual, expected runs minus actual, or expected performance minus actual. I suspect player performance would be the biggest, but that's just a guess.

6. Principles

The effect we have been discussing here — for the purposes of this paragraph, we'll call it the Extremes Effect — is similar to two other effects discussed by Bill James: the Whirlpool effect, and the Plexiglass principle.

The Whirlpool Principle (1983 Abstract, p. 220) states that "all teams are drawn forcefully towards the centre"; winning teams will tend to decline towards .500 in subsequent seasons, and losing teams will tend to improve towards .500.

The Whirlpool Principle is caused by two factors: One, the Extremes effect, which states that any team is closer in talent to .500 than its record indicates; and second, the talent of teams tends to approach .500 as time goes on. The reason for the latter is obvious; good teams tend to have good players, who, if they stay in the lineup, will eventually decline in talent, and who, if they don't stay in the lineup, will have to be replaced by worse players (If Wade Boggs gets hurt, there's nobody at his level of talent to replace him). Bad teams, on the other hand, will get rid of mediocre players and replace them with (better) young talent.

We can see how much of the Whirlpool principle is caused by the first (extremes) effect and how much by the second, as follows: take all AL teams from 1961-84 (excluding 1980-81) and look at how they did in the following year (teams were grouped in 5-win groups starting at 53-57 wins — the first row below is the average of the group):

Wins in 1st year:	56	61	65	70	76	80	85	90	94	99	104	109
Predicted talent:	62	65	69	73	78	81	85	88	90	92	95	98
Wins in next year:	69	70	72	74	76	83	83	88	90	88	93	103

These results pretty much agree with the Effect; the weak teams improved beyond their expected talent, while the strong teams declined past theirs. The pattern is not quite as strong for the good teams as for the bad, but it still holds (except for the 109 group, which consists of only three teams). If we are to trust the theoretical values, it appears that most of the whirlpool effect is caused by the extremes effect at high or low levels of achievement.

The Plexiglass Principle states that any team that has a large gain (decline) one season will tend to have a decline (gain) the next. There are three reasons for this. Two are the ones described above in the explanation of the whirlpool principle: that the team was probably lucky and talent changes tend to move towards .500. The third is that the average team that has a large gain (decline) over only one season has often done so out of luck, since personnel changes tend normally not to produce a large season-to-season fluctuation: a team's talent normally doesn't change drastically in only one year. Teams that show a large single-season fluctuation are thus particularly lucky, and since that luck doesn't necessarily carry on to the next season, many teams bounce back. (It would seem, then, that if there is an obvious reason for the large fluctuation, such as a free-agent gain or loss, or two, the bounce-back should be smaller. I haven't done the necessary research to check that, though.)

7. A few notes

— The basic effect here applies to individual players, as well as teams. Mark McGwire hit 49 homers last year. That's an extreme event, and there aren't very many 49-home-run hitters in existence, so, knowing nothing else about Mark McGwire, we would assume that his achievement was helped by a substantial amount of luck. I say "knowing nothing else" because for an individual, you have a much better idea of his level of talent than you do for a team. For Mark McGwire, we have previous major and minor league batting records, scouting reports of his talent, etc, that suggest that before 1987, nobody thought McGwire was an extraordinary power hitter at all. We might therefore conclude that his 49-homer performance was the result of luck based on the prior evidence of his talent, which suggests a figure in the range of 30 home runs, perhaps. For a team, on the other hand, since it is probably in its current state for only a year, what with trades, retirements, injuries, etc., we haven't had time to form independent notions of how good the team is, and we must usually resort to such analytical methods as described in section 4 to make a conclusion about the team's actual talent.

— The effect applies in reverse to bad teams, or bad players. Just as a 100-game winning team is not as good as its record indicates, neither is a 60-game winning team as bad. Note that the numerical results in the table may not be applicable because expansion does strange things with the distribution of teams at the low end. I think they're reasonably accurate, nonetheless. Anyway, since the principle applies at both ends, we could rephrase the effect as "On average, a team's (player's) talent is closer to .500 than its record indicates, and the difference is due to luck."

— The effect provides a suitable explanation of the sophomore jinx. Since very seldom does anyone talk about the sophomore jinx with respect to a rookie of modest first-year stats, the "disappointing" second seasons match up with extreme rookie accomplishments which were probably caused, in part, by luck.

— The following table shows the chance of a team repeating or improving its record next year, assuming its talent stays the same as this year. Note that because a team's talent will almost always change, these numbers understate the actual chance for the bad teams, and overstate the actual chance for the good teams. It might therefore be better to think of this as the chance a team would have a better record if it could play the <u>current</u> season over again.

The top number is the number of actual wins in the season; the bottom is the probability of reaching or exceeding that number.

50	55	60	65	70	75	80	85	90	95	100	105	110
.88	.81	.73	.68	.64	.62	.58	.52	.41	.29	.21	.14	.08