By the Numbers

Volume 14, Number 3 The Newsletter of the SABR Statistical Analysis Committee August, 2004

Summary

Academic Research: A Managerial Firing Model

Charlie Pavitt

The author reviews a recent attempt to model managerial firings.

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

Scott M. Berry, Hired to be Fired, Chance, Spring 2004, Volume 17 Number 2, pp. 55-59

I am happy to note that Scott M. Berry continues his "A Statistician Reads the Sports Pages" column in most of the quarterly issues of *Chance*. This time, he has attempted to discover the factors that best predict the odds of a head coach/managerial firing in each of MLB, NBA, NFL, and NHL, based on teams that were in existence through the 25-year stretch from 1979 to 2003. A couple of admitted problems; he considered all managerial changes to be "firings" unless he remembered otherwise (e.g., the retirements of Joe Gibbs and Tom Kelly) but apparently did not attempt a rigorous check of his memories, and he considered all coaches to have finished the season even if fired partway through.

Beginning with a baseline presumption that the average coach would end up at .500, Berry developed a series of logistic models, using winning percentage for the given year, change in winning percentage between that year and the previous year, difference in winning percentage between that year and the last season before his hiring, and years on the job as predictors. Some of his more interesting findings: Overall, firings are less likely in the NFL (19 percent in a given year) than in the others (28 percent in MLB, 29 percent in NBA, 38 percent in NHL); Berry believes that among the sports, coaches have the biggest influence in football and the least in hockey and that this difference explains this difference in likelihood of firing. Based on just last-year record and tenure, the odds of a first-year MLB manager with a .500 record being fired is 21 percent (NHL 29, NBA 12, NFL only 4). In all the sports, the probability of being fired with a .500 record after a given year goes up from the first year, reaching its maximum during the third year in MLB, fourth year in the NFL and NHL, and fifth year in the NBA. It then drops thereafter. In other words, new hirings are given a grace period of, depending upon the sport, two to four years.

Every 100 percentage points of won-loss changes the risk of firing by 10 percent in MLB and 7 percent in the others; Berry attributes the difference between sports to the relative lack of variation across teams in won-loss during a given year in baseball, so that a 100-point change probably means more in the standings in baseball. A change in winning percentage from the previous year has an additional effect on this risk for MLB, NHL, and NBA, whereas change from the previous coach's last year has an additional effect in NFL only. The full model for MLB

In this issue

predicts firings about 5 percent better than a null model prediction of 28 percent (the average proportion each year) and about 8 percent better in the others sports. For a little fun, Berry listed the most extreme probabilities for each sport. For baseball, the most likely firing among managers in his data was Chuck Tanner (75.5 percent) after 1988, his third year with the Braves, with the team at 54-106, 15 games worse than his previous year and 12 games worse than the previous manager's last season. The most unlikely was Lou Piniella (1.6 percent) after 2001, his seventh season with the Mariners, with the team's historic 116-46.

Thanks

Thanks to all of those who responded to my request for help with locating sabermetric articles from the first two editions of the Journal of Sports Economics, and a SABR Salute to Don Coffin for lending me his copies of those editions.

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

Correction

Shane Holmes

Please note the following corrections to my May 2003 article, "Toying with 'The Favorite Toy.'"

a) Bill James has a caveat I failed to include. The chance of an event occurring cannot exceed 0.97 ^ yrs remaining. This caveat affects page 18. Foxx (who should have had an asterisk next to his name) becomes 77.4%, Griffey 78.9%, and A-Rod 76.0%.

b) On page 15 (grey box), I write, "Another caveat: A player's performance level must exceed or be equal to 75% of his most recent season performance. This protects against certain intrusions in a great player's career, such as a season-long injury or a labor-related work stoppage." But on page 17, I contradict this statement. The former is true, though it hardly solves the problems that arise.

c) Page 19 lists Erstad twice. The .0826 figure is incorrect.

d) I also noticed that I'd given an incorrect e-mail address. My permanent email address is actually shane@northwestern.edu.

Are Deviations from Pythagoras Really Just Luck?

Thomas Thress

Why do some teams finish above or below their Pythagorean projection? Is it just luck, or is there some other, underlying reason, such as a good bullpen or manager? In this study, the author notes that if there is a reason other than random chance, teams that over- or undershoot their projection at home should also do so on the road, and vice-versa. He then examines the evidence to see if that's the case.

At various times, I have read a number of theories of why teams might win more or fewer games than would be predicted by their Pythagorean winning percentage. For example, I have seen several people suggest that wins above Pythagorean winning percentage may be a valid means of evaluating how good a manager is. I have also seen it hypothesized that teams with good bullpens will tend to outperform their Pythagorean winning percentage. At other times, I have heard over/underperformance relative to Pythagoras attributed to how "efficient" one's offense is or to how balanced one's starting pitching is.

In the last issue of "By the Numbers," I presented a formula which would produce unbiased Pythagorean estimates for home teams and road teams ("Bias against the Home Team in Pythagorean Theorem", By the Numbers, May, 2004). In doing this, I was able to come up with separate Pythagorean projections at home and on the road for all American and National League teams since 1900, a total of 2,054 teams.

If any of the various hypotheses for why teams might be inclined to over/underperform their Pythagorean projection – be it managers, bullpens, offensive efficiency, or anything else – have any validity, one would expect that a team which outperformed its Pythagorean projection at home would also be likely to outperform its Pythagorean projection on the road, and vice-versa. After all, teams generally have the same manager, the same bullpen, the same offense, and the same starting rotation for both home and road games.

Data Used in the Analysis

In this paper, then, I compared teams' performance relative to their Pythagorean projection at home with that on the road, to see if there was a tendency for teams to over/underperform in the same way in both places. If there were such a tendency, this would support the general hypothesis that there was something intrinsic to baseball teams that made them more likely to over- or underperform their Pythagorean projection. On the other hand, if teams' performance relative to their Pythagorean projection at home was generally unrelated to this same performance on the road, then one could, perhaps, conclude that deviations from Pythagorean projections are not a function of anything intrinsic to particular baseball teams – i.e., that over/underperformance relative to Pythagorean projection was essentially a matter of luck.

For this analysis, I compared how teams did relative to Pythagorean projections at home and on the road. These projections were made using a PythagoPat projection with home-team adjustment as described in my previous paper, the results of which were summarized in Table 3 of that paper (By the Numbers, May 2004, page 22). This analysis included results for 2,054 teams from 1900 – 2003. For each of these teams, I calculated the difference between actual and predicted wins at home and on the road. These home and road differences were then compared for each of these teams.

I compared home versus road differences in three ways, of roughly increasing mathematical sophistication.

Tendency of Home and Road Differences to Have the Same Sign

If there were no relation in performance relative to Pythagorean projection at home and on the road, we would expect approximately 50% of all teams to have home difference and road differences which were the same sign. In other words, we'd expect half of all teams to either outperform Pythagorean predictions both at home and on the road or underperform these predictions both at home and on the road.

In fact, 52.1% of all teams from 1900 to 2004 either outperformed these projections both at home and on the road, or underperformed both at home and on the road. This would suggest that, while there may be some small tendency for a team to over/underperform in the same way at home and on the road, this tendency appears to be very small.

Correlation between Home and Road Differences

Next, I calculated the simple correlation between differences from expectations at home versus differences from expectations on the road for all 2,054 teams. The correlation between home and road differences was calculated to be equal to 0.0314, i.e., a correlation of 3.14%. As above, while the sign here is consistent with the hypothesis that teams may have some tendency to over/underperform in the same way at home and on the road, this number is very close to zero, suggesting that this tendency is very, very small.

Using Home Differences to Predict Road Differences and Vice-Versa

Finally, I fit a simple linear equation, which attempted to explain deviations from expectations at home as a function of deviations on the road. In this equation, the coefficient on road deviations was found to have a value of 0.032. Hence, for example, a team which outperformed its Pythagorean prediction at home by 5 games (just under 4% of the teams in the sample outperformed their home projection by 5 or more games) would be expected to outperform its Pythagorean prediction on the road by 0.16 games.

The t-statistic on this coefficient was estimated to be 1.43, meaning that we could say with approximately 85% confidence that this coefficient was, in fact, greater than zero. On the other hand, the R-squared for this equation was 0.0009, meaning that deviation from prediction on the road was only able to explain 0.09% of the variance in deviation from prediction at home.

A parallel equation that attempted to explain road differences as a function of home differences found almost identical results – coefficient of 0.031, t-statistic of 1.43, R-squared of 0.0010.

Finally, just to test the stability of these results, I fit the same equations using data only since 1998. Using data from 1998 – 2003, the coefficients on both of these equations were estimated to be equal to -0.05. That is, a team that outperformed its Pythagorean prediction at home had a very slight (and insignificant) tendency to underperform its Pythagorean prediction on the road (and vice-versa).

Anecdotal Evidence of Extreme Teams and Conclusions

Overall, there appears to be little, if any, evidence to suggest that teams which outperform their Pythagorean projection at home are any more likely to do so on the road, or vice-versa. One final way in which this can be confirmed is by looking at extreme teams.

From 1900 – 2004, a total of 80 teams (3.9% of all teams) outperformed their Pythagorean projection at home by at least 5 games. These teams outperformed their Pythagorean projection at home by an average of 6.05 games, but outperformed their Pythagorean projection on the road by an average of only 0.25 games per team. At the same time, a total of 74 teams underperformed their Pythagorean projection at home by at least 5 games, with an average underperformance of 6.26 games. These teams underperformed their Pythagorean projection on the road by an average of only 0.26 games per team.

Over this same time period, a total of 65 teams outperformed their Pythagorean projection on the road by at least 5 games. These teams outperformed their Pythagorean projection on the road by an average of 6.07 games, but underperformed their Pythagorean projection at home by an average of 0.02 games per team. The 77 teams which underperformed on the road by at least 5 games (average road underperformance of 6.12 games), meanwhile, underperformed their road projection by an average of only 0.27 games per team.

Ultimately, one is led fairly convincingly to the conclusion that over/underperformance relative to Pythagorean projection is not due to anything intrinsic about a team, but appears to be overwhelmingly random.

Thomas Thress, tomthress@aol.com ♦

Log5 – Derivations and Tests

Ray Ciccolella

Here, the author describes three different aspects of the log5 formula: first, he explains why *it works, by deriving it from the Pythagorean formula; second, he tests its accuracy against the results of actual team-to-team matchups; and, third, he compares its accuracy to a competing formula.*

In the May 2004 issue of *By the Numbers*, a derivation of the log5 formula was presented. In the May 2003 issue, Steven Skiena's model, based on his research of Jai Alai statistics, was provided as a possible alternative to the log5 method. Empirical tests for team versus team results were not given in either case, although the log5 formula has been tested for batter versus pitcher match-ups and found to be accurate (see the February 1999 and August 2001 issues of BTN).

About three years ago I developed a formula for predicting team versus team winning percentages, derived from Bill James' Pythagorean Formula. I discovered, after reading the May 2004 issue, that while my derivation and formula are different than the log5 method, the two formulas are actually equivalent. Recently I also developed an alternative explanation of the log5 formula based on a lottery concept.

Additionally, I found an adjustment that needs to be made to the log5 formula to make it more accurate and unbiased. The adjustment corrects for the implicit but incorrect assumption that the average opponent a team faces has a winning percentage of 0.500. The lottery concept highlights this implicit assumption.

Lastly, I tested the adjusted log5 formula empirically and found it to be accurate. I also tested Skiena's formula and found it was somewhat less accurate than the log5 approach, even after optimizing the variable exponent in his formula.

In the next few of sections I'll show my derivation of the log5 formula, demonstrate its equivalency to the original log5 formula, and present the adjustment described above. These three sections are mostly algebra and proofs. My alternative explanation of the log5 formula follows these sections, and while it has some equations, it is more intuitive than mathematical.

I conclude the paper with an evaluation of Skiena's Jai Alai formula.

Pythagorean Formula Derivation

In this section, I show how the log5 formula can be derived from the Pythagorean formula. Readers not interested in the details of the algebra can head directly to the section "Alternative Explanation."

Assumptions

1. Bill James' Pythagorean Formula holds true;
2. All teams allow on average the same number of runs per game;
3. Against each opponent a team will allow, on average, that opponent's average number of runs scored per game.

Definition of Terms

R_A = Runs Scored per game by Team A; R_B = Run Scored per game by Team B
OR_A = Team A Opponent Runs per game; OR_B = Team B Opponent Runs per game
L_A = League Average Runs Allowed
WP_A = Winning Percentage for Team A; WP_B = Winning Percentage for Team B

Derivation

We start with the standard Pythagorean formula:

$$WP_A = \frac{R_A^{\,2}}{R_A^{\,2} + OR_A^{\,2}} \text{ and } WP_B = \frac{R_B^{\,2}}{R_B^{\,2} + OR_B^{\,2}} \text{ (equations 1)}$$

Since $OR_A = OR_B = L_A$ (assumption 2), then substituting L_A for OR_A and OR_B yields

$$WP_A = \frac{R_A^{\,2}}{R_A^{\,2} + L_A^{\,2}} \text{ and } WP_B = \frac{R_B^{\,2}}{R_B^{\,2} + L_B^{\,2}} \text{ (equations 2)}$$

Cross-multiplying and then simplifying the terms creates

$$R_A^{\,2} = WP_A R_A^{\,2} + WP_A L_A^{\,2} \text{ and } R_B^{\,2} = WP_B R_B^{\,2} + WP_B L_B^{\,2}$$

$$WP_A L_A^{\,2} = R_A^{\,2} - WP_A R_A^{\,2} \text{ and } WP_B L_B^{\,2} = R_B^{\,2} - WP_B R_B^{\,2}$$

$$WP_A L_A^{\,2} = R_A^{\,2}\left(1 - WP_A\right) \text{ and } WP_B L_B^{\,2} = R_B^{\,2}\left(1 - WP_B\right)$$

$$R_A^{\,2} = \frac{WP_A L_A^{\,2}}{1 - WP_A} \text{ and } R_B^{\,2} = \frac{WP_B L_B^{\,2}}{1 - WP_B} \text{ (equations 3)}$$

When team A plays team B $OR_A = R_B$ (assumption 3) then WP_A is given by

$$WP_A = \frac{R_A^{\,2}}{R_A^{\,2} + R_B^{\,2}} \text{ (equation 4)}$$

Now substitute into (4) for R_A^2 and R_B^2 the equalities from (3) to create

$$WP_A = \frac{\dfrac{WP_A L_A^{\,2}}{1 - WP_A}}{\left(\dfrac{WP_A L_A^{\,2}}{1 - WP_A} + \dfrac{WP_B L_A^{\,2}}{1 - WP_B}\right)} \text{ (equation 5)}$$

Cancelling L_A^2 from all 3 terms yields the final version of my derivation

$$WP_A = \frac{\dfrac{WP_A}{1 - WP_A}}{\left(\dfrac{WP_A}{1 - WP_A} + \dfrac{WP_B}{1 - WP_B}\right)} \text{ (equation 6)}$$

Equivalency to the log5 Formula

With some algebraic manipulation we can see that (6) is equivalent to the log5 formula for predicting team versus team winning percentages.

Let A = the winning percentage of Team A; Let B = the winning percentage of Team B.

$$Log5 = \frac{A(1-B)}{A(1-B)+B(1-A)}$$

Multiply the terms:

$$Log5 = \frac{A(1-B)}{(A-AB)+(B-AB)}$$

Then add the values in the denominator to get a re-arranged version of the log5 Formula

$$Log5 = \frac{A(1-B)}{(A+B-2AB)} \quad \text{(equation 7)}$$

My derived formula was

$$WP_A = \frac{\frac{A}{1-A}}{\left(\frac{A}{1-A}+\frac{B}{1-B}\right)} \quad \text{(equation 6)}$$

First add the terms in denominator. The common denominator is (1-A)*(1-B). Multiply each numerator as appropriate for the new denominator to get:

$$WP_A = \frac{\frac{A}{1-A}}{\left(\frac{A(1-B)}{(1-A)(1-B)}+\frac{B(1-A)}{(1-A)(1-B)}\right)}$$

Which simplifies to:

$$WP_A = \frac{\frac{A}{1-A}}{\left(\frac{A(1-B)+B(1-A)}{(1-A)(1-B)}\right)}$$

And further to:

$$WP_A = \frac{\frac{A}{1-A}}{\left(\frac{A - BA + B - BA}{(1-A)(1-B)}\right)}$$

Next multiply top and bottom by (1-A)(1-B):

$$WP_A = \frac{A(1-B)}{A - AB + B - AB}$$

Add the terms in the denominator:

$$WP_A = \frac{A(1-B)}{A + B - 2AB} \quad \text{(equation 8)}$$

The result (8) equals the re-arranged version (7) of the log5 formula shown above.

Alternative Explanation

I've developed another explanation of the log5 formula using the analogy of a lottery. Imagine that a team has the same winning percentage against a .500 team as it does overall. Next assume each team gets a certain number of chips to represent their chances of winning with each team getting a unique color. To represent a game both teams' chips will be thrown into a hat and a chip will be drawn out of the hat with the winner being the team that had its chip selected. How many chips should the 0.400 team receive? What about the 0.600 team?

If we give the 0.500 team an arbitrary 500 blue chips then the 0.400 team needs 333.33 red chips. That way, the chance of drawing a red chip is .400:

```
   0.400 = X ÷ (X+500)
.4X +200 = X
     .6X = 200
       X = 333.33
```

Doing the same math, the 0.600 team needs 750 green chips to the .500 team's blue chips, so the chance of drawing a green chip is .600:

```
   0.600 = Y ÷ (Y+500)
       Y = .6Y + 300
      4Y = 300
       Y = 750
```

Then, when the 0.400 team (333.33 red chips) plays the 0.600 team (750 green chips) its winning percentage will be 0.308:

```
333.33 ÷ (333.33 + 750) = 333.33/1083.33 = 0.308
```

Using the log5 formula we get the same result.

```
.400 * (1-.600) ÷ [.400 *(1-.600) + .6 * (1-.400)] = .16 ÷ .52 = .308
```

The Adjustment

We need, however, to adjust winning percentages before doing an empirical test of the log5 formula since unless a team has a winning percentage of 0.500, their opponents' winning percentage is not 0.500. Without this adjustment, the log5 formula will be biased; the winning percentages of teams that are over 0.500 will be predicted to be a little bit higher than actual and the opposite will be true for the under 0.500 teams.

The adjustment is based on the number of teams in the group and the difference between 0.500 and the team's actual winning percentage.

I used the following formula to adjust winning percentages:

$$WP_{New} = WP_{Old} + \frac{.500 - WP_{Old}}{N}$$

I found that N =12 minimized the error in my test.

For example, a team with an actual winning percentage of 0.400 was adjusted to 0.408, since

$$.408 = .400 + \frac{.500 - .400}{12}$$

Empirical Test

I tested this model using data from both leagues from 1902 (first season I could get head to head results) through 1996 (last season without inter-league play). For each league and season I grouped each team based on their winning percentage rounded to the nearest 0.050. The groups were:

Group	Winning Percentage Range	Base Group Name
1	Less than 0.376	0.350
2	0.376 to 0.425	0.400
3	0.426 to 0.475	0.450
4	0.476 to 0.525	0.500
5	0.526 to 0.575	0.550
6	0.576 to 0.625	0.600
7	0.626 and above	0.650

I then calculated the won-loss record for each group versus group combination. The results of this process for the 1980 American League season are shown in Tables 1 and 2.

Table 1– Head to Head Results, 1980 American League

Team	NY	Bal	Mil	Bos	Det	Clv	Tor	KC	Oak	Mnn	Tex	Chi	Cal	Sea	W	L	Pct	Group
NY		6	8	10	8	8	10	4	8	8	7	7	10	9	103	59	0.636	0.650
Bal	7		7	8	10	6	11	6	7	10	6	6	10	6	100	62	0.617	0.600
Mil	5	6		7	6	10	5	6	7	7	5	7	6	9	86	76	0.531	0.550
Bos	3	5	6		8	7	7	5	9	6	5	6	9	7	83	77	0.519	0.500
Det	5	3	7	5		10	9	2	6	6	4	10	7	10	84	78	0.519	0.500
Clv	5	7	3	6	3		8	5	6	9	6	7	6	8	79	81	0.494	0.500
Tor	3	2	8	6	4	5		3	4	5	5	7	9	6	67	95	0.414	0.400
KC	8	6	6	7	10	7	9		6	5	10	8	8	7	97	65	0.599	0.600
Oak	4	5	5	3	6	6	8	7		7	7	7	10	8	83	79	0.512	0.500
Mnn	4	2	5	6	6	3	7	8	6		9	8	6	7	77	84	0.478	0.500
Tex	5	6	7	7	8	6	7	3	6	3		7	2	9	76	85	0.472	0.450
Chi	5	6	5	4	2	5	5	5	6	5	6		10	6	70	90	0.438	0.450
Cal	2	2	6	3	5	4	3	5	3	7	11	3		11	65	95	0.406	0.400
Sea	3	6	3	5	2	4	6	6	5	6	4	7	2		59	103	0.364	0.350

Read across the table for wins and down the table for losses: New York went 10-3 against Boston.

I then took the head to head results and reconfigured them as group versus group results. See Table 2. Again, read across for wins and down for losses. In the 1980 American League the "400 group" (California and Toronto) was a combined 132 - 190 and went 12-38 versus the "600 group" (Baltimore and Kansas City).

Table 2 – Actual Group vs. Group, 1980 AL

Group	.350	.400	.450	.500	.550	.600	.650	Total
.350	-	8	11	22	3	12	3	59
.400	17	12	26	46	14	12	5	132
.450	15	24	13	52	12	20	10	146
.500	40	77	69	124	26	49	21	406
.550	9	11	12	37	-	12	5	86
.600	13	38	30	76	13	12	15	197
.650	9	20	14	42	8	10	-	103
Total	103	190	175	399	76	127	59	

Repeating this procedure for all seasons from 1902 to 1996, combined, results in Table 3. From now on, the the groups are now classified based on the actual winning percentage.

Table 3 – Actual Group vs. Group, 1902 to 1996 Combined

Group	.336	.402	.452	.500	.548	.598	.655	Total
.336	601	965	1174	1592	1516	1190	530	7568
.402	1280	1223	2402	2834	2781	1686	632	12838
.452	1822	2822	4703	5013	5022	2517	1024	22923
.500	2858	4123	5995	6641	6128	3411	1103	30259
.548	3472	4839	7055	7336	6352	3327	1233	33614
.598	3199	3416	4328	4945	3971	2090	925	22874
.655	1736	1718	2161	1929	1932	1125	264	10865
Total	14968	19106	27818	30290	27702	15346	5711	140941

I created the predicted group versus group winning percentages by adjusting the actual overall group winning percentages as described above and plugging these values into the log5 formula. Using the "402" group and "598" group as an example:

First the Adjustment:

0.402 + (.5-.402) ÷ 12 = 0.410; 0.598 + (.5-.598) ÷ 12 = .590

Then the Calculation:

.410*(1-.590) ÷ [.410*(1-.590) + (.590*(1-.410)] = .1681÷ (.1681+ .3481) = 0.3256

So the log5 predicted percentage was .3256. The actual winning percentage was 0.3305 (a W-L record of 1,686–3,416). The error is calculated as (prediction minus actual), which is negative 0.0049 or (0.8) games per 162. (In the tables that follow, numbers in parentheses, such as the (0.8), denote negatives.)

Table 4 shows the log5 errors for each group.

Why the Adjustment was Necessary

Table 5 shows the same errors as table 4, but *without* the log5 adjustment to the winning percentages.

As expected, the errors are higher than we see in Table 4, except for the .500 group. We would expect the .500 group to not need an adjustment since this group's opponents do have a winning percentage of 0.500 but the errors even for this group are not random. The unadjusted log5 formula overpredicts how well the "500" group will play against the groups with winning percentages under 0.500 while underpredicting how well the "500" group will do against the stronger groups.

The log5 formula, without the adjustment, predicts too many wins for all the groups with a winning percentage over 0.500 with the opposite error for the groups under 0.500. In addition, the absolute size of the total error increases for the groups that are furthest away from 0.500.

Table 4 – Net Errors per 162 Games Log5 Method

	.336	.402	.452	.500	.548	.598	.655	Total
.336		1.0	(0.2)	(1.3)	1.0	0.1	(0.6)	0.1
.402	(1.0)		(1.0)	0.5	0.5	(0.8)	1.6	(0.3)
.452	0.2	1.0		0.1	(0.6)	(0.0)	(0.6)	0.1
.500	1.3	(0.5)	(0.1)		0.1	0.2	(1.1)	(0.0)
.548	(1.0)	(0.5)	0.6	(0.1)		(0.4)	1.5	0.1
.598	(0.1)	0.8	0.0	(0.2)	0.4		(1.0)	(0.0)
.655	0.6	(1.6)	0.6	1.1	(1.5)	1.0		0.2
Total	(0.1)	0.3	(0.1)	0.0	(0.1)	0.0	(0.2)	

Table 5 – Net Errors per 162 Games, Log5 Method without adjusting winning percentage

	.336	.402	.452	.500	.548	.598	.655	Total
.336		(0.1)	(1.9)	(3.5)	(1.6)	(2.9)	(3.9)	(13.8)
.402	0.1		(1.7)	(0.8)	(1.4)	(3.2)	(1.3)	(8.4)
.452	1.9	1.7		(0.5)	(1.9)	(1.9)	(3.1)	(3.8)
.500	3.5	0.8	0.5		(0.6)	(1.1)	(3.2)	0.0
.548	1.6	1.4	1.9	0.6		(1.2)	-	4.3
.598	2.9	3.2	1.9	1.1	1.2		(1.9)	8.4
.655	3.9	1.3	3.1	3.2	-	1.9		13.3
Total	13.8	8.4	3.8	(0.0)	(4.3)	(8.4)	(13.3)	

Summary

I believe that this empirical test demonstrates that the log5 model, after adjustment, accurately predicts head to head winning percentages in all pairings. The largest error is 1.6 games per 162 and 62% of the pairings had an error rate of less than 1 game per 162. These errors are less than the typical error of +/- 3 games generated by the Pythagorean Formula for predicted a team's winning percentage based on its runs scored and allowed.

I believe the formula, once adjusted, is also unbiased, as the negative and positive errors appear to be random and the total net error for any one group is no more than 0.3 games per 162. There is no apparent tendency for the errors to all be negative for one group and positive for another. I also believe that this empirical test demonstrates that to remove bias and increase accuracy the log5 formula must be used with adjusted winning percentages. I used a particular adjustment in my test; there are probably other adjustment formulas that would work just as well if not better.

Jai Alai Formula

I also checked Skiena's "Jai Alai" formula using the same process and adjusted winning percentages. His formula is:

$$WP_{A \cdot vs \cdot B} = \frac{1 + \left(WP_A - WP_B\right)^{\alpha}}{2}$$

The suggested alpha was 0.40 but I found that value did not optimize the error rates; a value of 0.62 yielded the best fit. Table 6 shows the results for Skiena's formula with alpha at 0.62 and adjusted winning percentages.

The "Jai Alai" model yields good results but the log5 model is more accurate overall and appears to be less biased. The log5 method had a lower error rate in 85.7% of the head to head comparisons and a lower overall error rate for each group.

The bias appears to be slightly higher as well, as the negative and positive errors appear to be less random. For the log5 formula with adjusted winning percentages the sign (positive or negative) of the error was the same as the previous one 40% of the time while the Jai Alai formula had 80% of the signs the same as the previous sign. I also tested the Jai Alai formula with unadjusted winning percentages but the error rate was higher while the error minimizing alpha changed to 0.66.

Table 6 – Net Errors per 162 Games, Jai Alai Method (Skiena) with adjusted winning percentage

	.336	.402	.452	.500	.548	.598	.655	Total
.336		(2.9)	(2.7)	(2.0)	2.4	3.6	5.3	3.7
.402	2.9		(5.4)	(3.2)	(1.4)	(0.5)	4.7	(3.0)
.452	2.7	5.4		(4.5)	(4.4)	(1.9)	0.3	(2.3)
.500	2.0	3.2	4.5		(4.5)	(3.4)	(2.2)	(0.5)
.548	(2.4)	1.4	4.4	4.5		(4.9)	(1.3)	1.7
.598	(3.6)	0.5	1.9	3.4	4.9		(5.1)	2.1
.655	(5.3)	(4.7)	(0.3)	2.2	1.3	5.1		(1.6)
Total	(3.7)	3.0	2.3	0.5	(1.7)	(2.1)	1.6	

Ray Ciccolella, rciccolella@austin.rr.com ♦

An MVP Voting Model – Part II

Tom Hanrahan

In a previous article, the author created an MVP voting model that described what factors appear have historically been important important to voters. Here, in Part II, he investigates what types of players tend to be under- or overpredicted by the model, and how the factors have changed in importance over time.

Part I of this article appeared in the August, 2003 issue of *By the Numbers*. In it, I explained what factors have been important to BBWAA members when they cast their annual MVP votes. Using only year-end statistics of the top 3 finishing batters and their teams for the years 1938-2002, I created a linear regression model (including dummy variables) that predicted (in hindsight) the correct winner in 82 of 129 races (63.6%), with an RMS error of predicted to actual MVP vote totals of 49 points. I ignored pitchers in the voting; so in 1986, Don Mattingly is considered to have "won" an MVP trophy over Jim Rice, even though in reality Roger Clemens owns the hardware.

Quick overview of conclusions for part I: The factors important in gaining MVP votes, in rough order, are

- Very Important: Great Triple Crown stats, leading the league in RBI for a playoff team, and playing SS for a playoff team;
- Important: Stealing lots of bases for a playoff team, and leading the league with lots of steals;
- Somewhat Important: Being a new guy on a surprise winner, playing C or 2B for a playoff team, leading in SLG or OBA, hitting over .300, winning a gold glove, and finishing 2nd in RBI for a playoff team;
- Helpful: Playing for a team that won many games (and didn't win the previous year), scoring and/or driving in at least 100 runs, being a veteran, and finishing 3rd in RBI for a playoff team.

In this article, I check out a few variables I did not consider before, and then look at how voting patterns have changed over the years. After creating a revised time-based model, I will compare the voting record of many of the all-time great hitters to what the model predicts for them: who got more credit than the raw stats they produced, and who didn't the voters like?

New Variables

Someone suggested that I look at whether there is any historical bias related to media coverage. Do players benefit in the voting just by being a Yankee, or on a large-market team in general?

Answer: not that I could find. There have been 34 Yankees who have finished in the top 3 in MVP voting between 1938 and 2002. They averaged one point less in MVP point totals than the model prediction. The same pattern is true for other "big" clubs: no significant bias for or against, over the years, for Giants, Red Sox, or Dodgers hitters.

Another item I looked at was race. Tom Timmerman (who deserves and gets a big thank you!) supplied a database of players, classified by origin: Caucasian, African, and Hispanic (while Ichiro got his very own category). Out of 324 total players (using only the years beginning in 1947, when the color barrier was broken), 161 were Caucasian, 122 African, and 41 Hispanic. The white players received an average of 9 points more than blacks and Hispanics, when compared to what the model predicted. That isn't a large amount, and in fact if I added 9 points to all of the white players' totals in the model, only one of the MVP races would have been predicted differently: Robin Yount's award over Ruben Sierra in 1989. But if we look merely at MVP winners, there have been far more trophies given to fair-skinned players when dark-skinned players "deserved" it by the model prediction, than the other way around. There were 17 occurrences since 1947 when a player of one race won an MVP award (once again, "won" here means finished first among non-pitchers) while a player of another race was predicted to win by the model. Fourteen of these involved white-vs-black players, and in only 3 of these 14 cases did the black player win: see Table 1.

What does this mean – is it blatant discrimination? I don't know. There could be other factors involved: maybe it is that blacks are over- or underrepresented in ways the model is in error. It could also be that voters subjectively favor those players who are more media-friendly, and quite possibly many blacks were not as good "interviews" to the predominantly white journalists.

I note that in many of the cases above, the white man played a more demanding defensive position. While the model is supposed to correct for this, an error in the model coefficient would change many of the predicted results. One of the problems of attempting to assign a value to something such as "playing shortstop for a playoff team" is that the amount of credit given by the voters is very inconsistent; in some years a huge amount of credit is given, and in others, none.

Table 1 – White vs. Black MVP Controversies

	Actual Winner	Predicted Winner
2000 NL	Kent	Barry Bonds *
1995 NL	Larkin *	Bichette
1991 AL	Ripken	Fielder *
1988 NL	Gibson	Strawberry *
1984 AL	Hrbek	Murray *
1973 NL	Rose	Bobby Bonds *
1971 AL	Bando	Frank Robinson *
1971 NL	Torre	Aaron *
1968 NL	Rose	McCovey *
1963 NL	Groat	Aaron *
1960 NL	Groat	Mays *
1959 NL	Banks *	Mathews
1955 NL	Campanella *	Snider
1954 AL	Berra	Doby *

* Black player

Creating a time-based model

In attempting to determine how the value of some variables might change over time, I first divided the data into two and three different time periods; more than this, and there may have been too little data in each sub-group. Of course, any effects that occurred over a small time period would be missed in this way, but I was looking for larger trends. I checked for significant differences in regression coefficients. When I found them, I determined whether the better measure would be to create different factors for the time periods, or changing the value of the variable over time using a function in the form

coefficient = A + year * B.

I found two factors that significantly changed over time: the weight of leading in RBI for a winning team, and the position bonus. In voters' eyes, RBI leaders have become more important over the years, and the effect of playing a key defensive position for a winning team has become smaller. Both of these effects have been noted by others previously. Bill James once stated that there was a stronger bias before 1965 to give the MVP award to an up-the-middle (C-2B-SS-CF) performer. This is what the model has found, with the exception that CFers are given no extra credit. It seems to me that the reason centerfielders used to won many MVPs in the 1950s is that there were many great CFers in the 1950s!

To derive the time-based model, I took the model I had previously created (in Part I), and substituted the time-based criteria for RBI and position bonus for the standard ones I had before. Of course, these new variables caused the other coefficients to change as well. Most were not significantly different, as can be seen in the table below. The only one that had a dramatic effect was batting average; it became more important overall, but this is mitigated by the lowering of the bonus for attaining a .300 AVG. In the new model, hitting .300 is worth an extra 25 points of batting average, so that the difference in batting .300 and .299 is the same as the difference between averages of .326 and .300.

Table 2 – Coefficients when using Timeline Model vs. Original Model

	Original model	Timeline model
AVG	.34	.47
HR	.62	.64
RBI	.57	.65
Team Wins	.35	.44
Team Won last yr	-9	-11
Yrs played MLB	2.2	2.1
Gold Glove	20	20
>= .300 AVG	17	12
>= 100 runs	8	7
>= 100 RBI	8	7
League Leading Bonuses		
Stolen Bases	SB * .50	SB * .54
Slugging	20	17
On-base average	24	25
Playoff team		
RBI league leader	50	1.04 * (year – 1930)
RBI 2nd in league	25	.52 * (year – 1930)
RBI 3rd in league	13	.26 * (year – 1930)
Stolen Bases	SB * .82	SB * .80
Played SS	50	.52 * (2090 – year)
Played 2B or C	25	.26 * (2090 – year)
First year with team that didn't win in the previous season	31	34

Table 2 shows the different coefficients used in the original and time-based models.

An explanation of the time-dependent variables: for RBI leaders on playoff teams, in the year 1978 the weight for the time-based model is the same as in the original, because

```
1.04 * (1978 – 1930) = 50 points.
```

By 2002, the perceived value of the this has increased by 50%. Back in 1938, voters apparently paid little attention to this metric. Similarly, in 1994, the weights for the position bonus on winners are the same for both models. Back in 1946, the time-based model weight was 50% higher.

As with any time-based coefficients like I have used above, it would be incorrect to attempt to extrapolate for years beyond the ones used here (1938 to 2002). I surely don't propose that in the year 2100 voters will be giving negative credit for playing shortstop.

Is the new time-based model better (more accurate) than the original? Slightly. The overall RMS error reduced from 49.0 to 48.6 points, the r^2 rose from .420 to .431, and there was a new increase of one race predicted correctly (4 new ones correct, but 3 others newly called wrong), for a total of 83 out of 129, or 64.3%.

The following are the 9 instances (out of 24) where the time-based model has incorrectly predicted the MVP winner among hitters in the past 12 years – with a brief comment about what circumstances I believe led to the writers choosing the player they did:

2003 NL Bonds over Sheffield -- recognition of Bonds' huge # of walks, thus forgiving his low HR & RBI totals
2003 AL A Rod over Garciaparra – no credit given for playing shortstop on a winner
2001 AL Suzuki over Boone -- Ichiro given huge bonus for new guy leading Seattle to record 116 wins
2000 AL Giambi over A Rod -- no credit given for playing shortstop on a winner
1999 AL I Rod over Ramirez -- Alomar splits Cleveland vote
1999 NL C Jones over Williams -- September heroics against Mets clinched award for Chipper
1996 AL Gonzalez over Belle -- writers hated Belle
1995 AL Vaughn over Belle -- writers hated Belle
1995 NL Larkin over Bichette – Coors field compensation; usually occurs for Rockie batters

For those of you REALLY paying attention, the Kent/Bonds 2000 NL award is *not* listed above, even though it was mentioned previously as an example of "white man bests black man." This is because the time-based model saw Kent edging out Bonds that year, while the original model did not.

Which Players Did Better or Worse Than Their Numbers Predicted?

To answer this question, I focused on players who appeared in the top 3 very frequently in MVP voting, and who not coincidentally are often thought of as the greatest hitters in the past 65 years: chronologically, Ted Williams, Stan Musial, Mickey Mantle, Willie Mays, Hank Aaron, and Barry Bonds. I compared predicted to actual points, and also predicted to actual finish in place value. For the exercise of comparing place of finish, I included pitchers, since this way in the end I could measure the actual MVPs a player won versus what the model said. When there were pitchers involved, I left their position as is. For example, I will use the National League of 1963. The actual MVP finish was 1st

Table 3 – Ted Williams MVP voting record in years he was in the top three

Year	Predicted Pts	Actual Pts	Difference	Predicted Rank	Actual Rank
1957	265	209	-56	1	2
1949	286	272	-14	1	1
1948	258	171	-87	2	3
1947	244	201	-43	1	2
1946	264	224	-41	1	1
1942	279	249	-35	1	2
1941	267	254	-13	1	2
1939	204	126	-78	3	3
Average			-46		

Koufax – 2nd Groat – 3rd Aaron. The model predicted finish among hitters was 1st Aaron – 2nd Mays – 3rd Groat. Aaron's actual finish was 3rd, but the model predicted him to finish 2nd; Koufax keeps his "predicted" trophy, as I am unwilling to vault Aaron over him just because the model saw him beating Dick Groat.

One player about whom it is often said that he was shafted in MVP voting was Ted Williams. The model confirms this. Table 3 shows how the model saw Williams faring in the voting (among the 8 times he actually finished among the top 3), and how he actually did. Remember, this does take into account that Ted's team only won 1 pennant (1946) in this time, and gives him no credit for the Sox' frequent close 2nd place finishes in the 1940s.

The lower-than-model-predicted voting totals for the Splendid Splinter are amazingly consistent. He averaged 46 points fewer than the model predicted. One item I have not touched on yet is whether voters account for the hitter's home field effects in their voting. Certainly it is possible that the voters did account for a Fenway effect in Williams' case. I tested this by comparing the actual to predicted voting for all other Red Sox in this database. On average, the 17 non-Ted-Williams Sox garnered a measly seven fewer points than the model prediction. Thus, there may have been be a small adjustment made by the writers for their home park's benefits, but it is not significant. It's more likely, in my judgment, that the writers may have been influenced by the fact that the Boston club consistently had many good hitters at that time but often failed to win, and they corporately concluded that Ted's statistical accomplishments were overdrawn. Or, maybe they just plain hated his guts. Regardless, he did in fact finish lower than a purely mathematical model would suggest every single year, and won only 2 instead 6 MVP awards as predicted by the model.

Table 4 – Stan Musial

Year	Diff	Predicted Rank	Actual Rank
1957	-13	2	2
1952	-81	4	5
1951	-20	2	2
1950	-66	2	2
1949	-9	2	2
1948	+33	1	1
1946	+46	1	1
1943	+32	1	1
Average	-10		

Table 5 – Mickey Mantle

Year	Diff	Predicted Rank	Actual Rank
1964	-25	1	2
1962	+77	1	1
1961	+26	2	2
1960	+47	2	2
1957	+18	2	1
1956	+50	1	1
1952	-13	4	3
Average	+26		

Table 6 – Willie Mays

Year	Diff	Predicted Rank	Actual Rank
1966	-59	4	3
1965	-17	1	1
1963	-71	4	5
1962	-26	1	2
1960	-89	1	3
1958	-41	1	2
1954	+41	1	1
Average	-36		

Table 7 – Hank Aaron

Year	Diff	Predicted Rank	Actual Rank
1971	-42	1	3
1969	+4	3	3
1963	-71	2	3
1959	-63	1	3
1958	-24	3	3
1957	-12	1	1
1956	-35	3	3
Average	-34		

Table 8 – Barry Bonds

Year	Diff	Predicted Rank	Actual Rank
2003	+37	2	1
2002	+55	1	1
2001	+78	1	1
2000	-1	2	2
1993	+58	1	1
1992	+30	1	1
1991	+12	1	2
Average	+40		

Results for the other great hitters I mentioned above are in tables 4, 5, 6, 7, and 8. Table 9 is a summary of these players, a composite of tables 3 through 8.

Table 9 – Summary of Tables 3 to 8

Player	Predicted 1st/2nd/3rd	Actual 1st/2nd/3rd
Bonds	6/2/0	6/2/0
Aaron	1/0/6	3/1/3
Mays	2/2/2	5/0/0
Mantle	3/3/1	3/3/0
Musial	3/4/0	3/4/0
Williams	2/4/2	6/1/1
Totals	17/15/11	26/11/4

In Part I of this article, I found no strong relationship between MVP points and previous awards won. However, in sampling these few superstars it is apparent that cumulatively they won many fewer trophies than a mathematical model would expect (a total of 17 actual MVP awards versus 26 predicted). It may be that there is not a large bias in general against past award winners, but that in close votes, there has been a tendency to "let someone else have a turn," and often the superstar ends up 2nd or 3rd. In particular, Mays, Aaron and Williams each fared much worse than their statistics would suggest. In contrast, Barry Bonds (the much despised) has in fact done quite well. I take this as evidence that the media has focused on Bonds' incredible number of walks drawn, and has forgiven his comparatively lower totals in the key counting stats of home runs and RBI.

Quick Summary

To win an MVP award:

- Leading the league in RBI for a playoff team will usually turn the trick by itself.
- Hit for a high batting average and hit lots of home runs.
- Playing shortstop (or second base or catcher) for a playoff team is real nice – since normally its hard to find hitters at those positions. But recently, there have been lots of good hitting shortstops around.
- Leading the league with 50 or more steals for a playoff team is a way to for non-sluggers to win.

Trends and Other Notes

- RBI weren't as big a thing in 1950 as they are today.
- There may racial biases in voting – or maybe they are just "perceived ornery player" biases.
- Superstars have won fewer trophies than a pure statistical model suggest.
- Barry Bonds' 6 awards, in spite of his surliness, and in spite of usually not even finishing near the RBI leader, is a simply astounding accomplishment.
- Some things just aren't captured in year-end statistics.

For Further Study

There are limitations to a linear regression model. Perhaps using a logarithmic approach like Rob Wood (February, 1999 issue of BTN) has done would work in some ways that this model did not. Mr. Wood use only league-leading variables, but the challenge would be to make it work for continuous variables such as batting average. Good luck to anyone who wishes to attempt this.

Using more than the top three finishers may well give better coefficients for a linear model. This is both because the model using only the top three will miss some players who finished lower in the voting but had good stats, and also because with more players, the model could better quantify the effect of two or more players on one team 'splitting the credit' for winning.

Predictions for 2004

This article was written after the end of the season, but before the MVP announcements. What does the model say for this season? Here is what will happen if previous trends hold:

AL

A very close race, with Vladimir Guerrero edging out the two Sox sluggers, Manny Ramirez (SLG leader and 3rd in RBI) and David Ortiz (2nd in RBI). Vlad's advantage according to the model comes from being the "new guy" leading a team (Anaheim) to a playoff berth that they didn't get last year. In reality, Ramirez and Ortiz may take some support away from each other, and so someone else may sneak into the top three. The model doesn't know about the Angels strong finish and Guerrero's pile of homers in the last week of the season, so he likely will win by a greater margin. The model sees two Orioles, Melvin Mora (100 RBI and the OBA leader) and Miguel Tejada (RBI leader, who would have walked away with the award had the O's won) as 4th and 5th among hitters, but the media are playing up Sheffield, despite other Yankees having fine years. Ichiro has the new hit record, but even though he had as good a year as when he actually won an MVP trophy, it ain't gonna happen for a last-place team. No Twin had a banner offensive year, but Johan Santana will get decent consideration. Mariano Rivera may as well. Michael Young had a chance to finish high if Texas won, but they didn't, and he doesn't.

Model prediction: Guerrero first with 277 total points, and Ramirez second with 262.

NL

There are three Cardinals hitters having monster years. While they probably will steal votes from each other, Scott Rolen is the favorite; finishing 2nd in RBI while playing gold-glove third base for a runaway winner is a sure thing most years. There is also Adrian Beltre, who could parlor awesome overall stats on a playoff team to a very high finish. And of course, there is Superm--, I mean, Barry Bonds. The model sees him finishing a close second to Rolen. The model will be wrong. The model gives Barry no credit for drawing over 220 walks and setting new OBA and OPS records. Sportswriters, however, have indeed recognized that it's pretty difficult to rack up high RBI totals when opposing managers absolutely refuse to pitch to you, and many of them may for the 7th (!) time decide to write Bonds' name at the top spot on their ballots. The Astros' late run for the wild card bodes well for Lance Berkman and Jeff Kent, both of whom might finish in the top 10. Mark Loretta could have made some noise if the Padres had won, but they didn't, and he won't.

Model prediction: Rolen first with 338 pts, Bonds second with 331, then Pujols, Edmonds and Beltre all in a knot.

Tom Hanrahan, Han60Man@aol.com ♦

Informal Peer Review

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

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

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

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

Member	E-mail	Expertise
Jim Box	im.box@duke.edu	Statistics
Keith Carlson	kcsqrd@charter.net	General
Dan Evans	devans@seattlemariners.com	General
Rob Fabrizzio	rfabrizzio@bigfoot.com	Statistics
Larry Grasso	l.grasso@juno.com	Statistics
Tom Hanrahan	HanrahanTJ@navair.navy.mil	Statistics
John Heer	jheer@walterhav.com	Proofreading
Dan Heisman	danheisman@comcast.net	General
David Kaplan	dkaplan@UDel.Edu	Statistics (regression)
Keith Karcher	karcherk@earthlink.net	Statistics
Chris Leach	chrisleach@yahoo.com	General
John Matthew IV	john.matthew@rogers.com	Apostrophes
Nicholas Miceli	nsmiceli@yahoo.com	Statistics
Duke Rankin	RankinD@montevallo.edu	Statistics
John Stryker	johns@mcfeely.interaccess.com	General
Tom Thress	TomThress@aol.com	Statistics (regression)
Joel Tscherne	Joel@tscherne.org	General
Dick Unruh	runruhjr@dtgnet.com	Proofreading
Steve Wang	scwang@fas.harvard.edu	Statistics

An Examination of Season Sweeps and Near-Sweeps

Bruce Cowgill

For many years of major-league baseball, teams played each other 22 times per season. However, in all those years, no team has ever swept a season series 22-0. Is that unexpected? What is the actual chance of a 22-0 sweep?

In a recent SABR-L web posting, David Paulson commented that the Dodgers' baseball broadcasters announced that the Dodgers had just achieved a season sweep of the Pirates. The announcers added that it was the Dodgers' first season sweep since moving to the west coast in 1958. I cannot confirm whether or not that is true, but I find it hard to believe with all the short season series on today's schedule (e.g., some teams only play three games against each other). What I do know is that prior to 1958, the Dodgers' had *never* swept a season series, even dating all the way back to 1904. Mr. Paulson, in his posting, correctly assumed that with today's schedule, a season sweep was no great feat, especially the Dodgers' six game sweep of the Pirates (yes, six games). This was not always the case.

In response to Mr. Paulson's inquiry about season sweep records, Marc Alan Jones reported that the most games won in a modern day sweep was 13 by the 1993 Atlanta Braves over the expansion Colorado Rockies. There have also been instances of 12-0 sweeps, most recently the 1999 Texas Rangers over the Minnesota Twins. Prior to 1961, and the arrival of increasingly shortened season meetings, no team ever swept a 22-game season series. However, four teams have been close with 21-1 records: the 1909 Chicago Cubs over the Boston Braves, the 1927 Yankees over the St. Louis Browns, the 1937 Pirates over the Reds, and the 1945 Cubs over the Reds.

While certainly better than 6-0, how does Atlanta's 13-0 mark compare with the Dodgers' "feat"? And how does Atlanta's sweep compare to the 21-1 near-sweeps? My question was from a statistical viewpoint, not one based on differences in era that might make it more or less possible (e.g., travel comforts, personal assistants, day vs. night games, etc.). I realize in the "interesting statistical feats" category, a 13-win sweep does not deserve even a footnote in comparison to feats such as DiMaggio's hitting streak, but regardless, 13-0 is a decent record, especially if one considers that only five teams have gone 12-0 since 1970. If that is the case, then 21-1 seems that much more remarkable despite missing the sweep (let's face it, there is not much difference in the court of humiliation between 22-0 and 21-1).

In short, my questions are:

- How likely were the Dodgers to sweep the Pirates?
- How likely was Atlanta to win 13 games against Colorado?
- How does Atlanta's 13-0 compare to the teams who achieved 21-1 records?
- How likely is a 22-game season sweep?
- Should a 22-game season sweep have occurred in history?

Calculating Team vs. Team Sweep Probabilities

I decided to take a deeper look at the likelihood of such events occurring. I investigated the teams involved in the near-sweeps mentioned above. Rather than taking a slightly easier road using each team's overall win-loss record, I decided to examine the impact of each team's record at home and away as the basis for my analysis. Table 1 shows each team's overall record, place finished in league (division rank in 2004 and 1993), and home and away splits. Also included is the winning percentage differential.

There are some interesting findings in Table 1. First, aside from the Dodgers-Pirates sweep, none of these sweeps or near-sweeps were achieved by first vs. last place teams. However, in each case, either a first place team or a last place team was involved. Second, although somewhat obvious, if the head to head record is removed for these teams, their overall records suffer (improve) quite a bit. That is, removing 21 wins (or 21 losses) has a substantial impact on a team's record. This is most evident in the case of the 1937 Pirates. If the Pirates had not nearly swept the Reds, their record would have been below .500 (65-67, .492)! (Conversely, the Reds' record would be 55-77, .417.) Third, three of the 12 teams had a better record on the road than at home: The 1993 Braves, 1909 Cubs, and 1937 Reds.

Table 1

Team	Overall Record	%	Finish	Home Record	%	Away Record	%
2004 Los Angeles Dodgers	64-43	.598	1st	35-19	.648	29-24	.547
2004 Pittsburgh Pirates *	49-57	.462	6th	26-26	.500	23-31	.426
	Δ	.136		Δ	.148	Δ	.121
1993 Atlanta Braves	104-58	.642	1st	51-30	.630	53-28	.654
1993 Colorado Rockies [1]	67-95	.414	6th	39-42	.481	28-53	.346
	Δ	.228		Δ	.149	Δ	.308
1909 Chicago Cubs [2]	104-49	.680	2nd	47-29	.618	57-20	.740
1909 Boston Braves [3]	45-108	.294	8th	27-47	.365	18-61	.228
	Δ	.386		Δ	.253	Δ	.512
1927 NY Yankees	110-44	.714	1st	57-19	.750	53-25	.679
1927 St. Louis Browns [4]	59-94	.386	7th	38-38	.500	21-56	.273
	Δ	.328		Δ	.250	Δ	.406
1937 Pittsburgh Pirates	86-68	.558	3rd	46-32	.590	40-36	.526
1937 Cincinnati Reds [5]	56-98	.364	8th	28-51	.354	28-47	.373
	Δ	.194		Δ	.236	Δ	.153
1945 Chicago Cubs	98-56	.636	1st	49-26	.653	49-30	.620
1945 Cincinnati Reds [6]	61-93	.396	7th	36-41	.468	25-52	.325
	Δ	.240		Δ	.185	Δ	.295

* Record as of August 5th following the sweep. Also note that unlike the other series listed, the Dodgers (West) and the Pirates (Central) are in separate divisions. At the time, the Pirates record was actually better than two teams in the West.
[1] The San Diego Padres record was actually worse at 61-101
[2] The Cubs finished behind the Pirates' 110 wins
[3] The Braves almost achieved the dubious honor of going 1-21 against two teams in one year, but they remained at 1-20 vs. Pittsburgh saved by a season with only 153 official games
[4] The Boston Red Sox record was worse at 51-103 but managed a few additional wins to finish with a 4-18 record vs. their former pitcher's new team
[5] Finished 3rd, 10 games back of the Giants
[6] The Philadelphia Phillies finished 15 games below the Reds in last place at 46-108

To calculate the probability of Team A winning at home, I used the following form of the log5 method:

$$P(TeamAwinsAtHomeVs.TeamB) = \frac{(AHomeWin\%)(BRoadWin\%)}{[(AHomeWin\%)(BRoadLoss\%)] \cdot [(AHomeLoss\%)(BRoadWin\%)]}$$

So, for the 1993 Atlanta Braves, the calculation would be as follows:

$$P(AtlWinsAtHomeVs.Col) = \frac{(.630)(1-.346)}{[(.630)(1-.346)] \cdot [(1-.630)(.346)]} = .763$$

The same formula is used for probability of winning away, substituting appropriately. Table 2 shows the single game probabilities of each "sweep" team (the first team listed) winning based on their home and away records from Table 1.

Intuitively, the probabilities are what one would expect from teams that have a high winning percentage when playing teams with very low winning percentages. Noteworthy is the 1909 Cubs' expected dominance over the Braves at home *and* away. Also, the 1927 Yankees' home probability is astounding (with today's schedule, if the Yankees played only the Browns, they would be 72-9 at home).

Since these are single game probabilities, the next step is to calculate the probability of winning a series of games. For the 1993 Atlanta Braves who won all 13 games, the calculation is fairly easy. The probability of winning 13 games in a row is the product of each game's probability. In 1993, the Braves played 6 home games and 7 away games against the Rockies (note that the order of the games does not matter in this calculation):

```
Pr(93 Braves winning 13 in a row vs. Rockies)
  = .763 x .763 x .763 x .763 x .763 x .763 x .671 x .671 x .671 x .671 x .671 x .671 x .671
  = .0121 or about 1 in 83
```

The formula for 22-0 follows the same logic, with an even split of 11 home games and 11 away games. The probability that the 1927 Yankees would sweep the Browns is 0.0039 or about 1 in 258.

The calculation for 21-1 is slightly more complicated, simply due to the fact there are 22 different ways for a team to have a record of 21-1 (as opposed to only one way for a team to have a record of 22-0). That is, a team could lose game 1 and win games 2 through 22 or win games 1 through 9, lose game 10, win games 11 through 22, etc. As it turned out, the Yankees came the closest of all four teams to a 22-game sweep. Losing the 22nd game! At home!

Table 2 – Log5 Probabilities for a Single Game

	p(winning at home)	p(winning on road)
2004 Dodgers vs. Pirates	.713	.547
1993 Braves vs. Rockies	.763	.671
1909 Cubs vs. Braves	.846	.832
1927 Yankees vs. Browns	.889	.679
1937 Pirates vs. Reds	.708	.669
1945 Cubs vs. Reds	.796	.650

Although the above long-hand calculations are not difficult, the binomial distribution function simplifies the process further and is available in programs like Excel. The binomial function can be found in any introductory statistics textbook. However, the home and away split does add extra steps to the normally straight-forward formula. Specifically, to determine the probability of 21-1 using the binomial function requires these steps:

Calculate Pr(win 11 out of 11 at home)
Calculate Pr(win 10 out of 11 away)
Calculate Pr(win 10 out of 11 at home)
Calculate Pr(win 11 out of 11 away)

The product of 1 and 2 covers the 11 ways of a 21-1 record with an away loss. The product of 3 and 4 covers the 11 ways of a 21-1 record with a home loss. The sum of these two products is the probability of a 21-1 record accounting for all 22 ways of a team having a near-sweep record of 21-1. Table 3 summarizes these calculations.

Table 3 – Streak Probabilities

	Pr (6-0)			
2004 Dodgers vs. Pirates	.0593	1 in 17		
	Pr (13-0)			
1993 Braves vs. Rockies	.0121	1 in 83		
	Pr (22-0)		**Pr (21-1)**	
1909 Cubs vs. Braves	.0210	1 in 48	.0887	1 in 11
1927 Yankees vs. Browns	.0039	1 in 258	.0255	1 in 39
1937 Pirates vs. Reds	.0003	1 in 3715	.0027	1 in 372
1945 Cubs vs. Reds	.0007	1 in 1406	.0062	1 in 161
Any two equally matched teams	.00000024	1 in 4,194,304	.0000052	1 in 190,650

As a point of reference, I included probabilities for a .500 team vs. another .500 team.

As Mr. Paulson assumed, the 2004 Dodgers' six-game sweep is not that impressive. Although relatively high, the probability is lower than I expected for a six-game sweep. This is more a function of the relatively small winning percentage differential between these two particular teams.

Because of the large winning percentage differential of the 1909 Cubs and Braves, it was actually more likely for the 1909 Cubs to have a 22-0 record than the 1993 Braves to have a 13-0 record. And, 21-1 was almost expected. With a 1 in 39 chance in 1927 and many other dominant years, one should not be surprised that there would be a Yankee season with a 21-1 record at some point. With each team playing seven 22-game series each year, 1 in 39 amounts to less than 1 in 6 team seasons (assuming the same dominance over those seasons).

The 1937 Pirates were the most unlikely of these teams to go 21-1. That's because they were not that strong a team, finishing third that year with a .558 record, resulting in the smallest winning percentage differential in our group. The 1 in 372 translates to 1 in 53 team seasons.

Should a 22-game sweep have occurred?

As I thought about it more and upon further examination of the above figures, I actually became less impressed. In fact, I am somewhat surprised that a 22-0 record never happened.

Again citing Marc Alan Jones, there have been 111 league seasons that included 22-game series: AL 1904-17 and 1920-60; and NL 1904-17 and 1920-1961. With eight teams in a league, there are 28 such series that occur each season. Multiplying 28 series by 111 seasons gives a total of 3,108 series. Accounting for the fact that both teams in a series have a chance for a sweep, there have been 6,216 opportunities for a sweep. None have happened. Zero. Baseball is 0 for 6,216 in 22-game sweeps! Given that the 1937 Pirates had a 1 in 3,715 chance as a 3rd place team with a relatively ordinary record at sub-.600, it presented me with the initial justification to examine my "less than impressed" intuition.

While the 1909 Cubs at 1 in 48 chance and the 1927 Yankees at 1 in 258 chance are probably not common, I suspect that there have to be numerous combinations of similar winning percentage differentials between two teams. In fact, in both these cases, there were other combinations of teams in those very years who were *more* likely to be involved in a season sweep: the 1909 Giants vs. Braves (twice as likely as the Cubs vs. Braves) and the 1927 Yankees vs. Red Sox (five times as likely as the Yankees vs. Browns). These other combinations result in higher winning percentage differentials (which leads to a higher sweep probability), and I doubt that these are the largest differentials over 111 seasons.

Should a 22-0 sweep have occurred? Over 3,000 series, many teams with 90 or more wins, many teams with less than 60 wins, dynasties, eternal celler dwellers???

Calculating Season Sweep Probabilities Among Multiple Teams: 1909 NL

Before calculating probabilities over 111 seasons, I needed to justify the task. Of the teams examined, I decided to start with the 1909 Cubs season because of their relatively high probability of sweeping the Braves (yes, 2.1% is relatively high).

For this phase, I chose to use each team's overall record and ignored home and away splits. The sole reason for this decision is to reduce the number and complexity of the calculations required. The analysis in the first part of this paper focused on Team A vs. Team B. To determine the probability of a 22-game sweep in one season requires a more formidable task of examining all 28 season series (team A vs. B, A vs. C, A vs. D,…, B vs. C, B vs. D, etc.), and so a simplified approach seemed reasonable.

Table 4 shows the 1909 National League final results. Using these figures, the next step is to calculate the probability of Team i winning over Team j for all 28 combinations: Pr(i,j). The probability formula stated earlier or the log5 method can be used. This produces a grid of 56 probabilities for all possible head to head matchups. As we learned earlier, the probability of a team winning 22 games out of 22 is Pr(i,j) raised to the 22nd power. In the initial calculations we cared only about the Cubs sweeping the Braves, but now we have to factor in the probability of the Braves sweeping the Cubs (however remote), so all 56 combinations need to be calculated.

Adjusting the log5 Step

Before I was too deep into my analysis, I asked a statistics professor to review my initial calculations. While examining my calculation logic, Dr. Krieger pointed out a potential imperfection in the way single game head-to-head probabilities are calculated.

The problem is that when calculating all combinations of head to head probabilities in a league, a team's average Win% across all the teams does not equal its original winning percentage. Table 5 lists the 1909 Win% for each team and the log5 probability of the Pirates winning a single game vs. each team. That is, since the Pirates played each team the same number of times, the Pirates' log5 average should equal its overall Win%. It does not. The average of these probabilities equals 0.737, which is higher than the Pirates original 0.724.

Table 6 compares actual Win% to the log5 averages for each team. The figures are close but they are not exact. In this example, all teams play all teams, so one would expect the average across all teams to be 0.500. It is. But, the more the Win% deviates from .500, the larger the discrepancy.

Table 4 – Winning Percentages, 1909 NL

Pirates	.724
Cubs	.680
Giants	.601
Reds	.503
Phillies	.484
Dodgers	.359
Cardinals	.355
Braves	.294

Having never calculated all series combinations before, I was unaware of this issue. Strangely, I was not able to find reference to the issue in thousands of log5 web discussions, past BTN issues, or other SABR publications. Most unfortunately, I was unable to find Bill James' original papers on the method (I believe dating back to 1981-1983), so he may have referenced it. In any event, please forgive my ignorance.

Fortunately, Phil Birnbaum was familiar with the issue having (coincidentally) just received a paper by Ray Ciccolella[1]. Having not read Mr. Ciccolella's finished work, I probably cannot do it justice, so I will only summarize the adjustments I made based on his research. In short, the log5 method works best when each team's Win% is "relative to the same opposition." This is not the case above; the .680 Cubs played the Braves but fortunately did not have to play themselves, and the .294 Braves played the Cubs but *un*fortunately did not have to play themselves. Because of this, the Cubs played weaker opposition than the Braves. Therefore, the Cubs' log5 probability will be overstated and the Braves will be understated.

Mr. Ciccolella proposes a "normalizing" procedure prior to the log5 calculation to adjust each team's Win% so that the resulting log5 calculations better match a team's original Win%. The formula is as follows:

$$WP_{New} = WP_{Old} + \frac{.500 - WP_{Old}}{N}$$

where N is set to minimize the error.

Table 5 – Pirates' Raw log5 Winning Percentages Against 1909 NL Teams

Team	Team win Pct	Pirates log5 against team
Pirates	.724	
Cubs	.680	.552
Giants	.601	.635
Reds	.503	.722
Phillies	.484	.737
Dodgers	.359	.824
Cardinals	.355	.827
Braves	.294	.863
Average		.737

Table 6 – 1909 NL Teams' Unadjusted Expected Win% Calculated Using Unadjusted log5

Team	Team win Pct	Team's expected log5 win pct
Pirates	.724	.737
Cubs	.680	.689
Giants	.601	.606
Reds	.503	.503
Phillies	.484	.483
Dodgers	.359	.352
Cardinals	.355	.348
Braves	.294	.282

Rather than use a fixed N for all seasons, I calculated the N that minimized each season's sum of squares error. For most of the seasons, I found N=10 minimized the error, but N ranged from 8 to 16.

[1] See elsewhere this issue – Ed.

As Table 7 shows, after adjustment, the resulting probabilities are nearly identical to the initial Win%. Specifically, Mr. Ciccolella's procedure reduces the log5 probability error by a factor of nearly 200.

Table 7 – Results of Using Adjusted log5 Instead of Unadjusted

1909 NL	Win%	Ciccolella Adjustment to Win% "Ray%"	Log5 Average % using Win%	Log5 Average % using "Ray%"
Pirates	.724	.710	.737	.724
Cubs	.680	.669	.689	.679
Giants	.601	.595	.606	.600
Reds	.503	.503	.503	.503
Phillies	.484	.485	.483	.484
Dodgers	.359	.368	.352	.360
Cardinals	.355	.364	.348	.356
Braves	.294	.307	.282	.294
Sum of Squares Error Index			52.90	0.282

Back to 1909

Now that I have adjusted the 1909 Win% for log5, I can recalculate all 56 sweep probabilities. The results are shown in Table 8.

Table 8 – Probability of a given team sweeping a given opponent, 1909 NL

		"Sweeping" Team							
		Pirates	Cubs	Giants	Reds	Phillies	Dodgers	Cardinals	Braves
"Swept" Team	Pirates		0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
	Cubs	0.0000018		0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
	Giants	0.0000326	0.0000060		0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
	Reds	0.0004970	0.0001318	0.0000098		0.0000001	0.0000000	0.0000000	0.0000000
	Phillies	0.0007771	0.0002198	0.0000183	0.0000005		0.0000000	0.0000000	0.0000000
	Dodgers	0.0091817	0.0038060	0.0006440	0.0000455	0.0000253		0.0000002	0.0000000
	Cardinals	0.0098260	0.0041191	0.0007120	0.0000518	0.0000290	0.0000003		0.0000000
	Braves	0.0258115	0.0127513	0.0030225	0.0003394	0.0002081	0.0000039	0.0000034	

Table 8 shows very small probabilities for most matchups, but a few stand out. The near sweep Cubs over the Braves probability is 0.0128. The Pirates over the Braves is 0.0258, double that of the Cubs' chances. The only other probabilities close to 0.01 are the Pirates over the Cardinals (0.0098) and the Pirates over the Dodgers (0.0092).

Next, I am going to reduce the grid back down to 28 by calculating the probability of *neither* team winning 22 out of 22 games. This is simply 1 $-Pr(i,j)^{22} - Pr(j,i)^{22}$. As one can expect, the chances of an underdog sweeping are very slim, thus Pr(j,i) is very small in all cases. Also as one would expect, most of these no-sweep probabilities are very high, with only two less than 0.99.

Some might dismiss these "less probable" combinations just because they are so remote. While these combinations may not be substantial on an individual basis, collectively they will increase the likelihood of occurrence. This is somewhat analogous to purchasing multiple lottery tickets: buying two lottery tickets will double your chances of winning, even though increasing your odds from 1 in 10 million to 1 in 5 million is no reason to quit your job (buying 1 million tickets still gives you a 9 in 10 chance of losing!). While some may view these as non-substantial changes saying "there is no difference to me between 1 in 5 million and 1 in 1 million," statistically there is quite a difference. The point is that small individual team sweep probabilities do add up and should not be ignored.

Back to the math. The product of these 28 head to head, no-sweep probabilities gives us the probability of *no* 22-game sweep occurring in *any* series in that season. For the 1909 NL, this is equal to 0.9298. To calculate the probability that *at least one* sweep occurs, subtract the result from 1. So, the probability of *at least one* 22-game sweep occurring in the 1909 NL is 1 - 0.9298 or 0.0702.

In the 1909 NL, there was a 1 in 14 chance of a 22-game sweep! That seemed fairly significant. If that is the case, the chances of a 22-game sweep *not* occurring over 55 plus years must be small, and my intuition would be validated.

Counting the AL's 55 seasons and the NL's 56 seasons, there are 111 league seasons. Using the 1909 NL probability of *no* season sweep 0.9298 raised to the 111^{th} power, we get an estimate of the probability of a 22-game sweep *not* occurring over 55+ years of play. Not surprising, this figure is very small. Subtracting it from 1 results in the probability of *at least one* 22-game sweep over these seasons. This probability is 0.9997! This is about as near to a statistical guarantee as one can get. So far my thinking was correct: a 22-game sweep should have occurred. Correct? Not exactly.

Calculating More Probabilities: Beyond 1909 NL

If the 1909 NL season is representative of the other 110 seasons, then one would expect a sweep to have occurred. However, I suspect that the 1909 NL is *not* like other seasons. In that season, the top team's Win% was 0.724 and the last place team was 0.294, a difference of over 0.400. That seems quite large to me. The 1998 Yankees 51-game lead over the expansion Devil Rays only amounted to a differential of .315.

To further place that season in context, I examined a league of total parity. That is, all the teams finished with a 0.500 record. In this scenario, the chances of a 22-game sweep occurring in one particular season are about 0.000013. Over 111 parity league seasons, the probability of at least one sweep occurring is only 0.0015 or 1 in 675. So, this suggests that in seasons where there is competitive balance, the chance of a sweep occurring is low. Conversely, less competitive balance results in higher probabilities of a sweep. In fact, even small deviations from parity result in significantly higher probabilities. For example, consider a league with a 0.600 team, a 0.400 team, and six 0.500 teams every year. This would result in about a 0.051 sweep probability over 111 seasons. This is 34 times the likelihood of a total parity league.

The difference between the top and bottom team is not the only difference that matters. Deviation across all eight teams contributes to increasing the likelihood of a sweep occurring. The standard deviation of Win% from the 1909 NL season is 0.159. I calculated the other near-sweep years' standard deviation along with their sweep probabilities. Table 9 compares these seasons.

Table 9 illustrates how changes in the Win% standard deviation impact the probability of a season sweep occurring. The 1927 AL's deviation may appear only slightly smaller than the 1909 NL, but statistically the difference is substantial. In 1927 AL, the Yankees did dominate but the other teams in the mix did not, at least not to the same extent as the 1909 NL teams. In 1909 NL, there were three teams over .600 and three teams below .400. In 1927 AL, only the Yankees were above .600 and only two teams were below .400. So, in this comparison, 1927 AL was more competitively balanced (I am sure you never heard that before) and resulted in a lower standard deviation. This lower deviation subsequently results in a season sweep probability of 0.0192, considerably less than 1909's.

Table 9 – Probability of Season Sweep for Various SDs of Win%

Year	Win% Std Dev	Pr(at least 1 season sweep)	Previous column expressed in odds	Pr(at least 1 sweep over 111 seasons)
Parity	0.000	0.000013	1 in 14,899	0.0015
1909 NL	0.159	0.0702	1 in 14	0.9997
1927 AL	0.123	0.0192	1 in 52	0.8838
1937 NL	0.099	0.0027	1 in 365	0.2623
1945 NL	0.115	0.0099	1 in 101	0.6676

However, like 1909 NL, 111 seasons of 1927 AL would result in a very high probability of a 22-game sweep. This is not quite the case for 1937 NL. While the 1937 NL had two teams above .600 and two teams below 0.400, the spread was tighter. In fact, the top team was only 0.625 and the last place team was 0.364. With 111 seasons like the 1937 NL, the probability of a sweep is much less than 50%.

So where are we? Some seasons produce a high likelihood of sweeps while others produce hardly any chance. Using these near-sweep seasons as our sample is probably not reasonable, simply because they were the ones that attracted our attention by producing matchups resulting in 21-1 records. Yet it is important to point out that these four league seasons alone had a 1 in 10 chance (0.0996) of producing at least one season sweep. And we still have to account for another 107 league seasons. One can reasonably assume that at least some of those other seasons will have high standard deviations. Any additional poor competitive balance seasons would contribute heavily to the chances of a sweep occurring, thus supporting my initial hypothesis that a 22-game sweep should have occurred.

With four seasons already accounted for, the next step was to input the remaining 107 seasons of data. Fortunately, a colleague was able to extract the data I needed in an efficient manner so I avoided the laborious task of having to input each team's record.

Examining the season records proved informative. The standard deviations of unadjusted Win% for all 111 league seasons ranged from 0.049 to 0.159 with an average of 0.101. As it turns out, the highest standard deviation season was the 1909 NL! (With my sample of four containing an outlier, my hypothesis was in jeopardy.) In that season, the difference between the first and last place team's Win% was .430 or 65.5 games back. This was not the largest differential. That occurred in 1906 NL, with the Cubs having a .439 differential over the Beaneaters (Braves) who were 66.5 games back. The smallest deviation (.049) occurred in the 1915 NL, where the Phillies differential over the Giants was only .138, a mere 21 games back.

Table 10 – Single Game Probabilities Using Adjusted and Unadjusted log5

	Method	Pr (winning at home)	Pr (winning away)
1909 Cubs vs. Braves	UnAdj	.846	.832
	Adj	.832	.813
1927 Yankees vs. Browns	UnAdj	.889	.679
	Adj	.875	.650
1937 Pirates vs. Reds	UnAdj	.708	.669
	Adj	.690	.650
1945 Cubs vs. Reds	UnAdj	.796	.650
	Adj	.780	.625

These extreme seasons produced a season sweep probability of 0.0757 for the 1906 NL and 0.00013 for the 1915 NL, a difference by a factor of 582. Unfortunately, I was disappointed to learn that the average single season sweep probability was under 1% at 0.0088. A low average single season sweep probability coupled with an outlier as part of my test sample, I knew that the likelihood of at least one 22-game sweep occurring over 111 seasons was no longer a statistical guarantee.

Table 11 – Streak Probabilities Using Adjusted and Unadjusted log5

	Method	Pr (22-0)		Pr (21-1)	
1909 Cubs vs. Braves	UnAdj	.0210	1 in 48	.0887	1 in 11
	Adj	.0136	1 in 74	.0644	1 in 15
1927 Yankees vs. Browns	UnAdj	.0039	1 in 258	.0255	1 in 39
	Adj	.0020	1 in 496	.0151	1 in 66
1937 Pirates vs. Reds	UnAdj	.0003	1 in 3,715	.0027	1 in 372
	Adj	.0002	1 in 6,770	.0016	1 in 623
1945 Cubs vs. Reds	UnAdj	.0007	1 in 1,406	.0062	1 in 161
	Adj	.0004	1 in 2,706	.0036	1 in 279

Over 111 seasons, the probability of *at least one* 22-game season sweep occurring turned out to be 0.6288, nearly a two in three chance. Although a far cry from a statistical sure thing, a 22-game sweep was more likely to occur than not. Therefore, I can still conclude that I am "somewhat" surprised that a 22-game sweep did not occur.

Revisiting Sweep Probabilities

Having not discovered the log5 issue until I was calculating all series combinations, my initial team vs. team sweep probabilities were based on the log5 method using unadjusted Win% as shown in Tables 1, 2 and 3. I recalculated these probabilities using adjusted Win% and maintained the home and away ratio of wins in order to make an "apples to apples" comparison. The results are shown below in Table 10 and 11. In Table 10, the win probability for each series decreases only about one or two percentage points. However, this relatively small decrease dramatically impacts each team's sweep probability as seen in Table 11.

Caveats

It is important to note that when performing these types of calculations several assumptions have to be made, some explicit than others. For example, we do not really know the probability of team A winning vs. team B. We use each team's record to estimate such a probability, but it is still just an estimate. In this research, this estimate is taken a step further as each team's record is adjusted or "normalized" to account for scheduling discrepancies between teams. The approach I took is a reasonable one, but others may also be just as reasonable.

Some teams match up better or worse against certain teams. This is evident in this research by examining the 1937 Pirates, a sub-.500 ballclub which happened to win 21 out of 22 games against one of their opponents. The 1927 Yankees beat up on the Browns losing only once, but managed to lose four games (I almost said "four times as many") against a Red Sox team that was seven games *behind* the Browns. There are examples like this every season, but we assume each team's record is reasonable to estimate head to head probabilities.

Another "leap" is that we assume that the head to head probability is constant. That is, the 1927 Yankees probability of winning at home vs. the Browns is 0.892 (from Table 2) for all 11 home games. In reality, there may be several factors that cause this figure to deviate. A team's travel schedule preceding a series may affect the chances of winning, especially in the days before airlines and modern amenities. Injuries or illness are other factors. What if Ruth sat out a game in 1927 against the Browns (he only played in 151 of 154 games that season)? Finally, the starting pitchers may have the largest impact on game-to-game probability deviation. Maybe the Yankees Win% against the Browns would be 0.892 over 11 games, but depending on the pitching matchup, it may be .900 one day, .700 the next, and .500 the next.

Even if we could agree on all the factors to include, I doubt we could agree on the proper adjustments for each factor affecting a team's chances. The recent papers on Dimaggio's hitting streak shed further light on many of the factors involved and the subsequent difficulty of such estimates. In reality, the true probability of such events will never be known. Yet, we should be able to accept such estimates knowing that their precision is based on a reasonable set of assumptions.

That said, I thought it might be worthwhile to point out how these estimates are affected by even small changes in the assumptions. Consider the following:

- Team A has a 0.600 probability of winning vs. Team X every game (constant probability);
- Team B's overall probability of winning vs. Team X is also 0.600 but each game alternates between 0.700 and 0.500.

The probability of winning back to back games is 0.360 for Team A and 0.350 for Team B. Not much difference between the two. However, consider a 22-game sweep: Team A's chances are 1 in 76,000 and Team B's chances are 1 in 104,000. That is much more significant. The point is that each change, while relatively small as an individual change, becomes amplified over repeated calculations.

Acknowledgements

Special thanks to Dr. Abba Krieger of the University of Pennsylvania for verifying my initial calculations and pointing out a subtle flaw in my logic besides the log5 issue. Thanks also to Alice Muehlhof who extracted the necessary season data in minutes, saving me many bungling hours of data entry and database manipulation.

Bruce Cowgill is Vice President, Quantitative Research for Paragon Research & Consulting, a marketing research firm, specializing in the pharmaceutical industry. He can proudly claim that BALCO has never been a client. brucec23@comcast.net ♦

Submissions

Phil Birnbaum, Editor

Submissions to *By the Numbers* are, of course, encouraged. Articles should be concise (though not necessarily short), and pertain to statistical analysis of baseball. Letters to the Editor, original research, opinions, summaries of existing research, criticism, and reviews of other work are all welcome.

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

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

I usually edit for spelling and grammar. (But if you want to make my life a bit easier: please, use two spaces after the period in a sentence. Everything else is pretty easy to fix.) If you can (and I understand it isn't always possible), try to format your article roughly the same way BTN does.

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

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

Get Your Own Copy

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

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

The easist way to join the committee is to visit http://members.sabr.org, click on "my SABR," then "committees and regionals," then "add new" committee. Add the Statistical Analysis Committee, and you're done. You will be informed when new issues are available for downloading from the internet.

If you would like more information, send an e-mail (preferably with your snail mail address for our records) to Neal Traven, at beisbol@alumni.pitt.edu. If you don't have internet access, we will send you BTN by mail; write to Neal at
4317 Dayton Ave. N. #201, Seattle, WA, 98103-7154.