Baseball and Markov Chains: Power Hitting and Power Series

John P. D'Angelo

As children, my friends and I often played a game called All-Star Baseball. Each of us would manage a team of major league players. The *players* were circular cards. Each card was divided into labeled sectors of various sizes, corresponding to possible outcomes such as strikeout, walk, single, home run, and so on. The sizes varied considerably from player to player. To have a player bat, we placed the card on a spinner, filliped the spinner, and read the result. We all had a clear intuitive understanding of how the situation reflected baseball and how it didn't.

We kept intricate stats. We experimented with ways to make pitching matter. We introduced random fielding errors using a buzzer. Later I invented a version using dice instead of a spinner. Throwing four dice at a time was required to accurately model the statistical profiles of actual players. I created new players. One (call him Kingman) had only two possible outcomes, a strikeout and a home run. Another (call him Bowa) had only two outcomes, a single and a strikeout. I made the single on Bowa four times as likely as the home run on Kingman. I would play full games in which one team had nine players like Bowa and the other had nine players like Kingman.

Years later I realized that we had been regarding the game of baseball as a Markov chain. Before discussing Markov chains and how to use them to model baseball, I pose three exercises for the reader. I put the first two of them in the book [DW] (page 335), and I have assigned them on several occasions, in order to give an amusing application of summing power series. The following simple lemma on power series, and its generalization to more variables, can be used to evaluate explicitly many of the expected value formulas arising in a more involved treatment of these ideas. The reader who wishes to focus on baseball may skip the lemma, but she should at least attempt to estimate the answer to part 3 of Exercise 1.2. The answer appears in the conclusions.

Lemma 1.1. *Let f be a polynomial of degree d in one variable, and consider the power series*

$$s(x) = \sum_{k=0}^{\infty} f(k)x^k. \tag{1}$$

The series in (1) converges for $|x| < 1$, and there is a polynomial $b(x)$ of degree at most d such that

$$s(x) = \frac{b(x)}{(1-x)^{d+1}}. \tag{2}$$

In particular, s is a rational function.

Lemma 1.1 can be proved by differentiating the geometric series d times or by using the method of generating functions. We include it because it provides an elegant method for solving the key parts of Exercise 1.2, namely the formulas for $K(q)$ and $B(p)$. The calculation of $K(q)$ amounts to finding the expectation of a random variable with the *negative binomial distribution.* See page 95 of [HPS] for a derivation using probability generating functions.

Exercise 1.1. Prove Lemma 1.1. Find $s(x)$ explicitly when $f(k) = k\binom{k+2}{2}$ and when $f(k) = k^3$.

Exercise 1.2. Consider a player (B) who hits a single with probability p and otherwise strikes out. Make the assumption that three singles score a run. Assume that all at-bats are independent. Let

John D'Angelo is professor of mathematics at the University of Illinois. His email address is jpda@math.uiuc.edu.

$B(p)$ denote the expected number of runs scored per half inning (until three outs occur), assuming only (B) bats. Consider a second player (K) who hits a home run with probability q and otherwise strikes out. Let $K(q)$ be the expected number of runs scored per half inning.

- Prove that $K(q) = \frac{3q}{1-q}$ and that $B(p) = \frac{p^3(3p^2-10p+10)}{1-p}$.
- Sketch their graphs for $0 \le p \le 1$.
- Set $q = \frac{p}{4}$. For what values of p is $B(p) \ge K(\frac{p}{4})$?
- Set $q = \frac{2}{5}p$. For what values of p is $B(p) = K(\frac{2p}{5})$?
- How does $B(p)$ change if four singles are required to score a run?

Exercise 1.3. Consider a player who hits a single with probability p, a home run with probability q, and otherwise strikes out. Assume at-bats are independent and three singles score a run. For each nonnegative integer k, find the probability that this batter scores exactly k runs in a half inning. (Be careful of runners left on base.)

This article is a short introduction to the ideas connecting baseball and Markov chains. Many baseball fans have invented fantasy baseball games based on the Markov idea, and what I discuss in this article is by now a well-established notion. I do not intend to discuss the vast literature, but I do mention a few entries to it.

Lindsey [L] viewed baseball in this fashion as early as the 1950s. Cover and Keilers [CK] provided a Markov version for evaluating batters. Pankin ([P] and also `http://www.pankin.com/markov/theory.htm`) has written several precise articles about baseball and Markov chains. See also [N] and its references for serious discussion about the evaluation of baseball players using these methods. Katz [K] applies Markov chain analysis to the pitch count.

One of my favorite references on the subject of mathematical baseball is [TP], to which I refer for references published up to the early 1980s. Thorn and Palmer also edited *Total Baseball* [TB], more than two thousand pages of discussion and statistics. I am truly thankful for my research in several complex variables; without it I would have spent thousands rather than merely hundreds of hours with this volume. The many books by Bill James (item [J] in our reference list gives two early versions) helped introduce a generation of baseball fans to related ideas, especially the notion of *runs created*. Many additional aspects of baseball appeal to mathematicians. See, for example, the recent book [R].

The main idea in this article is the notion of *Markov runs*. We imagine, as described more precisely below, that a player is the only batter on his team. He bats randomly based on his statistics. How many runs will his team score? We also imagine a pitcher who pitches randomly based on his statistics. How many runs will his team allow? Jeff Sagarin, once upon a time a math major at MIT, publishes Markov-runs-per-game baseball statistics (and much more) online and updates them daily during the season. See `http://www.usatoday.com/sports/sagarin/nlb09.htm`.

Most mathematically inclined baseball fans have played with their own formulations of these ideas, and some baseball managers and executives have embraced them. On the other hand, many people in the baseball world sneer at "computer baseball", perhaps because they tremble with trepidation at anything scientific. I will never forget my disgust when a former Cubs manager, who was broadcasting a game on the radio, dismissed a physicist's study of how a baseball travels in varying weather conditions with "He don't know baseball."

This article is intended to be a short introduction whose primary purpose is to illustrate some of the connections between mathematics and baseball. Let us pause and briefly mention some interesting topics I omit.

One of the most interesting such topics concerns the notion of a *clutch hitter*. Such a player allegedly hits better when the situation matters more, but numerical studies seem to indicate that the concept is an illusion. On the other hand, the discussion in this article presumes that each plate appearance is an independent event. This assumption cannot be completely true. How accurate is it? A related question is whether players *get hot*. See, for example, [A].

What about fielding? Errors are rather unusual occurrences in professional baseball, and fielding percentages convey little important information. More important is how many plays a defensive player (fielder) makes. Would player (A) have made the same play as player (B) did? Except for completely routine plays, there is no way to know. Recent studies of fielding have attempted to measure how much ground a fielder covers. Many subtleties arise, and I do not know of any literature providing a good Markov chain model. Hence this article will ignore fielding.

Another omission from our discussion concerns the impact of pitching. In principle one can account for pitching via the Markov chain model, but, typically, a player will have faced a specific pitcher too few times for the statistical profile to be meaningful. Instead one could use a weighted average of the statistics of the pitcher (versus all batters) and the batter (versus all pitchers). Careful studies would then answer empirically the eternal conundrum: pitching is what percentage of baseball? When asked this question, the catcher Yogi Berra answered by saying "Ninety percent of this game is half mental."

I acknowledge the following people (all with Ph.D.s) for their insightful observations about baseball statistics over the years: Phil Baldwin (physics), John Marden (statistics), Bob Northington (statistics), Bruce Reznick (mathematics), and Kirk Sanders (classics). I also acknowledge NSF Grant DMS 07-53978 for research support. Finally I wish to thank several referees for useful comments on the exposition and for asking me to expand the reference list.

Markov Chains

Let us now recall the concept of a Markov chain. Let $X_1, X_2, \ldots,$ be a sequence of random variables taking values in a finite set. We think of this situation as a random system that can be in a finite number of possible states, labeled $e_1, \ldots, e_n$. The sequence of random variables is called a *Markov chain* if the following holds:

For each pair of indices i, j there is a fixed probability M_{ij} such that each time the system is in state e_i it will be in state e_j the next time with probability M_{ij}.

In other words, the conditional probability that $X_{k+1} = e_j$ given $X_k = e_i$ is, for each k, independent of the values of X_m for $m < k$.

We will regard the states e_i as basis vectors for n-dimensional real Euclidean space. Consider a convex combination $a(\lambda) = \sum_{j=1}^{n} \lambda_j e_j$ of these states. (The coefficients λ_k are nonnegative, and they sum to 1.) The n-tuple $a(\lambda)$ is called a *probability vector*. The coefficient λ_k represents the probability that the system is in state e_k at some fixed unit of time.

The system evolves in discrete intervals, often but not necessarily regarded as units of time. In baseball the unit will be a *plate appearance*. A plate appearance is similar to an at-bat, but includes certain situations (walks, hit batters, sacrifices) that baseball does not regard as an official at-bat.

The Markov matrix M, whose entries are M_{ij}, governs the evolution of the system. If x denotes the state at one moment (the value of the random variable X_k), then Mx denotes the state one moment later (the value of the random variable X_{k+1}). The matrix is time independent; in other words, the chance of going from state x to state Mx does not depend on any of the previous states. The number M_{ij} is called the *transition probability* from state e_i to state e_j. Each M_{ij} is nonnegative and, because the state must go somewhere, we have $\sum_{j=1}^{n} M_{ij} = 1$.

The reader should consult a book such as [KSK] for a more precise treatment of Markov chains and references to their many applications. Many elementary linear algebra books also discuss Markov chains and their applications.

Baseball

How do we regard baseball as a Markov chain? The number of outs and the location of the baserunners will determine twenty-five states. We will keep track of runs scored, but runs will not be part of the setup. What are the states?

We regard all situations with three outs as the same, and this situation is one of the states. Once we reach this state, we cannot leave it. For clarity we mention that in the simplified model we discuss, it makes no difference how many men are on base or where they are located after the third out is made. Runs scored will be the only thing that matters.

Otherwise the number of outs can be 0, 1, or 2, and there are eight possibilities for the runners. The bases can be empty; there can be one runner on first, second, or third base; there can be two runners on (in three different ways); or the bases can be *loaded*. These twenty-four situations define the other states.

Player H comes to the plate. Ignoring such situations as stolen bases or pick-offs that could occur during the plate appearance, we ask what happens after it. We imagine that the player bats randomly, according to his statistical profile. His plate appearances are independent. In other words, his chance of getting a specific kind of hit or making a specific kind of out is independent of how many runners are on and of how many outs there are. Based on this profile, we compute the transition probabilities for each pair of states, obtaining a Markov chain.

We could use stats from the previous year, the current year, or the career. But we must use the same profile each time. For convenience here, we will write the player's profile as follows, deviating slightly from the standard listing. Here AB denotes at-bats, S denotes singles, D denotes doubles, T denotes triples, HR denotes home runs, BB denotes walks plus hit batters, AVE denotes batting average, SA denotes slugging average, OB denotes on-base average, and $OPS = OB + SA$. We will discuss this rather new statistic later. We note for us that the number of plate appearances is $AB + BB$.

In the simplified scenario suggested above, we regard all outs as equivalent. Sluggers who often strike out are fond of reminding their managers that there is no practical difference between a strikeout and a pop-out. We will ignore double plays, sacrifice bunts, and sacrifice flies, also for simplicity. As a result our values of OB and OPS differ slightly from those in the official records.

Here is the profile; note that the last four elements can be computed from the first six, and hence they are not truly needed.

$$AB \quad S \quad D \quad T \quad HR \quad BB \quad AVE \quad SA \quad OB \quad OPS$$

For example, we give a fictional line for a star player X. We also include actual profiles for Albert Pujols and Chase Utley in 2009.

(X)	500	100	25	5	30	100	.320	.570	.433	1.003
(AP)	568	93	45	1	47	124	.327	.658	.443	1.101
(CU)	571	98	28	4	31	112	.282	.508	.398	.905

Player X has 500 at-bats but 600 plate appearances. In a given plate appearance player X has a $\frac{1}{6}$ chance to hit a single, a $\frac{1}{6}$ chance to reach first base by either a walk or hit batter, a $\frac{1}{120}$ chance to hit a triple, and so on.

We mention a small point here. In certain situations different actions by the batter produce the same effect. For example, if no one is on base, then a single and a walk have the same effect. On the other hand, assuming that a single advances a runner two bases, a single with a runner on second is far better than is a walk.

The sequence of random variables will be the list of situations arising in a half inning. It is easier to understand the baseball situation by focusing on the transition matrix. Given the batter H, we determine from his statistical profile a (twenty-five by twenty-five) matrix M_H. Under the assumptions discussed below, we claim that M_H defines a Markov chain.

We know the complete statistical profile of H. In other words, we know the probabilities that H makes each kind of hit or out, draws a walk, gets hit by a pitch, and so on. We assume a fixed result given each type of hit; for example, a single always advances runners two bases. We could, but we do not do so here, refine the model by allowing several different kinds of singles! Assuming that each plate appearance is independent (and in particular that what the batter does does not depend on the pitcher), we can assign probabilities that we pass from each state x to the state $M_H(x)$.

In this manner we encode the statistical profile of the player as a Markov matrix. For the player Kingman who either hits a home run with probability q or strikes out, we can easily determine the matrix M_K. We will not write out this 25 by 25 matrix. We simply note that, with probability $1 - q$, the number of outs increases by one and the base runners remain the same. With probability q, the number of outs remains the same, and the bases are cleared. Of course, the batter and the base runners all score a run in this case.

Now that we are regarding baseball as a Markov chain, there is little point in keeping the usual statistics. We know, for example, that a player who hits a home run every ten at-bats will on average hit one every ten at-bats in the Markov model. The key new statistic is how many runs will a team score if it uses this player for every plate appearance. We naturally call this statistic *Markov runs*. As we mentioned earlier, Sagarin publishes this stat, normalized using nine innings, for all players each day of the season.

It is possible, but inappropriate for this short article, to combine probability theory and linear algebra to compute the expected number of Markov runs by using the Markov matrix. Such results generalize the idealized situations discussed in Exercise 1.2. Rather than heading in this direction, we imagine finding expected Markov runs via computer simulation. We will then consider the relationship between Markov runs and other baseball statistics.

Given the statistical profile, the player bats randomly, using a computer simulation, according to the Markov chain model, until he makes three outs. The test is run thousands of times, and from it we determine the average number of runs scored per each nine innings. Below we will see that a team consisting of (nine clones of) Albert Pujols would score 9.38 runs per game; no other National League player would come within one run of him in 2009.

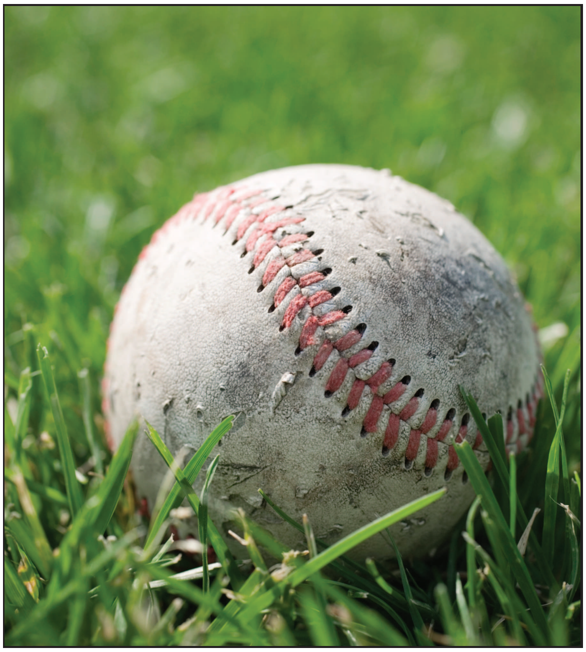
Some nice simplifications and approximations to these computations are known. Bill James [J] has found many formulas for runs created. In 1978 Pete Palmer [TP] ran a computer simulation of all games played since 1901. From this mammoth amount of data, he assigned *linear weights* to each batting event. Palmer performed these calculations separately for the years 1901–1920, 1921–1940, 1941–1960, and 1961–1978. The positive values of hits and the negative values of outs are not identical over the four periods, but they are all close to the following values:

A home run is worth 1.40 runs, a triple is worth 1.02, a double is worth .80, a single is worth .46 runs, and a walk is worth .33 runs. An out is worth −.25 runs. Given a profile, one can estimate Markov runs using these weights.

A second simplification now appears in official baseball statistics. One simply adds the player's slugging average to his on-base average, obtaining what is called OPS. It was discovered empirically that this number correlates fairly well with the expected number of runs scored using the Markov chain model. The list below of the top ten NL batters in 2009, ordered by Markov runs, suggests a correlation. It is likely that the correlation between Markov runs and OPS has been investigated carefully, but I do not know any references other than [J] and [TP], which are intended for baseball fans rather than for statisticians.

We make an important remark. For the actual values arising in baseball, OPS and Markov runs provide similar information. On the other hand, two players can have identical OPS but very different Markov runs. We give an example.

Example 3.1. We consider player one to have the profile of our contrived Bowa. His slugging average and his on-base average both equal p, and hence his OPS is $2p$. Now consider a player with the profile of Kingman, who hits a home run with probability q and otherwise strikes out. His slugging average is $4q$, and his on-base average is q. Hence his OPS is $5q$. Hence, if $q = \frac{2}{5}p$, then the two players have the same OPS. By Exercise 1.2, one sees that the Markov runs are not the same in general even when this equality holds. One can derive this conclusion more easily by noting what happens when p approaches 1. (Can analysis be worthwhile?) Obviously the Markov runs for Bowa approach infinity, whereas the Markov runs for Kingman are finite when $q = \frac{2}{5}$.



From a mathematician's perspective OPS is a bit strange. Let TB denote total bases, that is,

$$TB = S + 2D + 3T + 4HR. \tag{TB}$$

Roughly speaking, because we are ignoring sacrifices, as noted before, we compute OPS as follows:

$$\begin{aligned} OPS &= \frac{TB}{AB} + \frac{S + D + T + HR + BB}{AB + BB} \\ &= \frac{S + 2D + 3T + 4HR}{AB} + \frac{S + D + T + HR + BB}{AB + BB}. \end{aligned} \tag{OPS}$$

Formula (OPS) shows that we are adding two quantities with different denominators, at-bats and plate appearances. Perhaps a better number would be something such as

$$\frac{BB + 2S + 3D + 4T + 5HR}{AB + BB}. \tag{3}$$

Notice that (3) is essentially a linear weights formula.

Next we list, based on the Sagarin ratings for 2009, the top ten players in the National League, based on expected number of runs scored per nine innings computed via Markov chains. In this list, the first number after the player's name and team is this expected number of runs. The other number is the player's OPS, the sum of on-base average and slugging average.

1.	Pujols, Albert	St. Louis	9.38	1.101
2.	Votto, Joey	Cincinnati	8.26	.981
3.	Fielder, Prince	Milwaukee	8.24	1.014
4.	Lee, Derrek	Chicago	7.96	.972
5.	Ramirez, Hanley	Florida	7.81	.954
6.	Gonzalez, Adrian	San Diego	7.68	.958
7.	Helton, Todd	Colorado	7.61	.904
8.	Utley, Chase	Philadelphia	7.58	.905
9.	Braun, Ryan	Milwaukee	7.53	.937
10.	Dunn, Adam	Washington	7.24	.928

The method of *linear weights* warrants an entire fascinating chapter in [TP]. This method is essentially equivalent to the Markov chain method. In the Markov method, the batter himself bats every time. In the linear weights method, the value of each sort of hit is determined by empirical data involving all players. Both methods are based upon the assumptions of independence we have discussed.

Conclusions

One way to evaluate a batter is to determine his Markov runs. This very natural idea has been known intuitively for decades, and it lies at the basis of various strategy games. For the situations actually arising in baseball, the simpler statistic of OPS provides similar information. On the other hand, the formula for OPS is rather dissatisfying for a mathematician. Mathematicians who desire a simple statistic can use linear weights.

It is possible to include pitching in this discussion, where the key stat becomes Markov runs allowed. Space considerations prevent us from doing so here. Fielding is more difficult.

Readers who solved Exercise 1.2 can conclude that a player who bats .070 but hits only home runs is essentially equivalent (via Markov runs) to a player who bats .280 but hits only singles. Fix the ratio at $\frac{1}{4}$. For averages (for the singles hitter) lower than .27945 the home run hitter generates more runs, and for higher averages the singles hitter does so. I wonder how accurately baseball managers could guess this cut-off value.

Here is a novel idea for the true fan. Look up the Sagarin ratings. For each pitcher, find the Markov runs allowed, and then find a batter whose Markov runs equal this number. In this way you can determine tidbits such as whether there is any pitcher for whom the generic hitter he faces has the stats of Albert Pujols.

The world of baseball provides a striking application of Markov chains. Conversely, the use of Markov chains introduces fascinating new baseball statistics, such as *Markov runs* for batters and *Markov runs allowed* for pitchers. These statistics and simplified versions of them have evolved into standard methods for the evaluation of baseball players. As ever, mathematics provides artistic and insightful perspectives on another topic.

References

[A] Jim Albert, Streaky hitting in baseball, *Journal of Quantitative Analysis in Sports*, Berkeley Electronic Press, Volume 4, Issue 1, 2008.

[CK] Thomas M. Cover and Carroll W. Keilers, An offensive earned run average for baseball, *Operations Research* **25**(5) (1977).

[DW] John P. D'Angelo and Douglas B. West, *Mathematical Thinking: Problem Solving and Proofs*, Second edition, Prentice Hall, Upper Saddle River, New Jersey, 2000.

[HLP] Paul G. Hoel, Sidney C. Port, and Charles J. Stone, *Introduction to Probability Theory*, Houghton Mifflin, Boston, 1971.

[J] Bill James, *The Baseball Abstract*, Lawrence, Kansas, 1980, 1981.

[K] Stanley M. Katz, Study of 'The Count', *1986 Baseball Research Journal* (No. 15), pp. 67-72. An application of Markov chains to the ball-strike count.

[KSK] J. Kemeny, L. Snell, and A. Knapp, *Denumerable Markov Chains*, Van Nostrand Co., New York, 1966.

[L] G. R. Lindsey, An investigation of strategies in baseball, *Oper. Res.* **11** (1963), 477–501.

[N] Brad Null, Modeling baseball player ability with a nested Dirichlet distribution, *Journal of Quantitative Analysis in Sports*, Berkeley Electronic Press **5**(2), 2009.

[P] Mark D. Pankin, Baseball as a Markov Chain, *The Great American Baseball Stat Book* (First Edition, 1987), 520–524.

[R] Ken Ross, *A Mathematician at the Ballpark: Odds and Probabilities for Baseball Fans*, PI Press, 2004.

[TB] John Thorne and Pete Palmer, eds., *Total Baseball: The Ultimate Encyclopedia of Baseball*, Third Edition, Harper Collins Publishers, New York, 1993.

[TP] John Thorne and Pete Palmer, *The Hidden Game of Baseball*, Doubleday and Company, Inc., Garden City, New York, 1984.