An Intuitive Markov Chain Lesson From Baseball

Joel S. Sokol
School of Industrial and Systems Engineering
Georgia Institute of Technology
jsokol@isye.gatech.edu

Editor's note: This is a pdf copy of an html document which resides at http://ite.pubs.informs.org/Vo5No1/Sokol/

1. Introduction

One of the biggest challenges when teaching about Markov chains is getting students to think about a Markov chain in an intuitive way, rather than treating it as a purely mathematical construct. We have found that it is helpful to have students analyze a Markov chain application (i) that is easily explained, (ii) that they have a familiar understanding of, (iii) for which a large amount of real data is readily available, and (iv) that teaches them new insights about the application they thought was so familiar. Finding such examples can be difficult; introductory textbooks such as Durrett (1995), Winston (1997), Denardo (2002), and Hillier and Lieberman (2002) provide numerous examples that are easily explained, but the examples are generally written in "toy problem" form so that there is no need to work with real data; this makes it difficult to obtain believable new insights from the models. We feel that at taking students through least one in-depth example is useful, because it gives them a chance to experience a model with real-world complexity that is detailed enough to provide realistic insights. In this paper, we suggest an example from the world of sports − analyzing baseball with Markov chains.

The baseball model has all of the advantages we would like. It is easy to explain, and the model and the application are easily understood by most undergraduate students. Moreover, there is a staggering amount of detailed data that is readily available on the internet. Most importantly, the model allows students to gain new insight into a process that many of them start off thinking they fully understand.

In this paper, we describe a straightforward Markov chain model of baseball that has been used frequently in the literature for more than 40 years. We present a structured, logical lesson that we have successfully used to both cement Markov chain concepts in students' minds, and also influence students to think more deeply and intuitively about the ideas. (The lesson is not appropriate for teaching the basic concepts; it is designed to be used after the introductory lectures on the topic.) As part of the lesson, we provide links to some of the many good sources of data that can be found and exploited on the internet.

We have used the baseball example as a Markov chain lesson in an ungraded Independent Activities Period course at MIT, a special topics course in advanced discrete mathematics at Georgia Tech, and as part of an independent study course for Georgia Tech undergraduates. In all three cases, the students were upper-level undergraduates with a mathematical background that included at least 4 semesters of calculus. The Georgia Tech students all have been Industrial and Systems Engineering majors, while the MIT students had a variety of majors (engineering, science, and mathematics). As part of this paper, we will describe the typical reaction of these students to the lesson (including the ability of the non-baseball-fans to understand the ideas).

The remainder of this paper is organized as follows. In Section 2, we describe the Markov chain model of baseball. Sections 3 through 5 contain the lesson itself, divided into three sections: creating and validating the model, using the model for basic information-gathering and analysis, and using the model as a tool when searching for suggested solutions to system issues.

2. A Basic Markov Chain Model of Baseball

In this section, we present a simple Markov chain model of run-scoring in baseball. The model has been used by several researchers, including Howard (1960); Cook (1964); Thorn and Palmer (1984); Pankin (1991); Stern (1997); Bukiet, Harold, and Palacios (1997); and Sokol (2003).

The Markov chain is used to model the progression of a half-inning of baseball, in which one team bats until three outs have been made. The states of the Markov chain correspond to the positions of the runners on base and the number of outs. There are eight possible runner locations (three bases, each of which can be occupied or not, for a total of $2^3 = 8$) and three possible numbers of outs (0, 1, or 2), for a total of 8x3=24 states. To answer some of the questions in Sections 4 and 5, we will need an absorbing 25th state, "3 outs"; for other questions, transitions past the end of the inning will simply take us back to the beginning of the next inning, at the "no runners, 0 outs" state. We will refer to these states as (*B*,*O*), where *B* is the set of baserunners (written without brackets for clarity) and *O* is the number of outs. For example, state (12,2) corresponds to "runners on first base and second base, 2 outs" and state (Ø ,1) corresponds to "no runners on base, 1 out". Figure 1 shows the state space of the Markov chain without an absorbing (*,3) state.

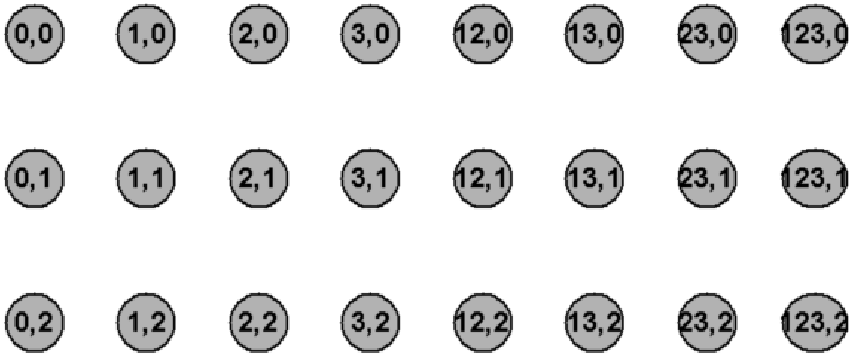

Figure 1: State space of the Markov chain; states are labeled in (B,O) format.

For this Markov chain model, state transition probabilities p_{ij} model the chance that the current batter's plate appearance will change the state of the system from *i* to *j*. For example, suppose state *i* = (1,0) ("runner on first base, 0 out") and state *j* = (Ø ,0) ("no runners on base, 0 out"). The only ways to get from state *i* to state *j* in a single transition are for the batter to hit a home run, or (very rarely) for the defense to make one or more errors that allow both the runner and the batter to score. Thus, p_{ij} is equal to the probability of the batter hitting a home run, plus the probability of the defense making the necessary errors. Figure 2 gives an example of the potential transitions from the state (3,1) ("runner on 3rd base, 1 out") using a simplified model that includes only basic baseball events; other arcs (representing events like a sacrifice fly, passed ball, or outfield-assisted double play) could easily be added.

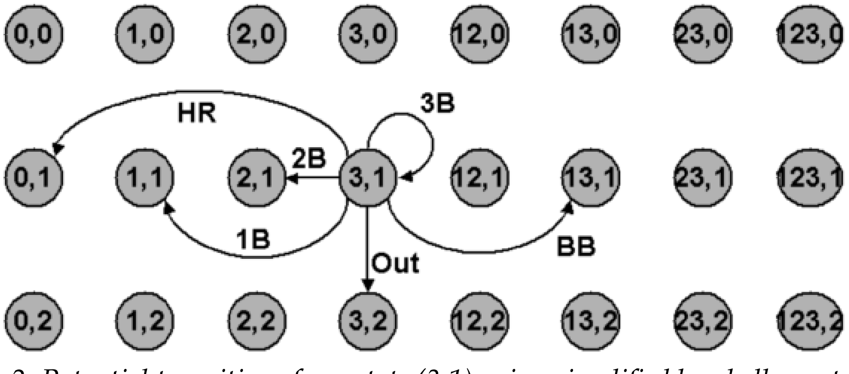

Figure 2: Potential transitions from state (3,1) using simplified baseball event model.

Students often ask at this point whether the model incorporates "clutch hitting", the idea that the probabilities differ by state because some batters perform better under pressure or with runners on base. It is simple to vary the probabilities by state, but students can be directed to easily-readable work by Cramer (1977) and Grabiner (1993a, 1993b) suggesting that clutch hitting does not exist, but is simply a product of random chance and small sample sizes. (Data sources such as CBS Sportsline[1] (2003), ESPN[2] (2003), Fox Sports[3] (2003), and Sports Illustrated[4] (2003) provide plenty of data for students to replicate the experiments of Cramer and Grabiner, if they so choose.)

3. Lesson I: Constructing and Validating the Model

After presenting the model, the most common reaction we receive from students is the question, "Is the Markov chain model really valid?" This is a perfect opportunity to turn the question around: instead of being told one way or another, the students can investigate this question on their own.

The most basic question of validity to ask is, "**Does this process satisfy the Markov property?"** In other words, is each batter's outcome dependent only on the current state?

This is an excellent point for class discussion, although the outcome will probably be inconclusive; it is clear that previous states might lend some information about the next transition (for example, if the previous few states all have O^* outs, the pitcher might be tiring and the next state will be more likely to also have O^* outs). On the other hand, such a situation might be uncommon because of the likelihood of a pitching change.

Other questions about the Markov property might arise from the notion of "lineup protection" − that a hitter will perform better if the following hitter is also better. (The idea is that the first hitter will get easier pitches to hit if the pitcher is worried about the second hitter.) Students interested in this question can read a study by Grabiner (1997), in which he concludes that lineup protection is statistically insignificant.

Thinking about whether the Markov chain exactly models the underlying process is important, but equally important from an educational standpoint is reinforcing the idea that a model can be appropriate for use even though it is not exact. The real question of importance is "**Does this Markov chain model provide a reasonable approximation of reality from which we can develop our analyses?**" With the abundance of data available on the internet, this is a question that students can readily answer by constructing and testing the model.

In order to test the model, students need to construct a transition matrix P. There is no direct data available on line that measures transition probabilities. However, there is much data on the probabilities of various events (home runs, doubles, strikeouts, etc.) that can be found at the web sites of CBS Sportsline (2003), ESPN (2003), Fox Sports (2003), Sports Illustrated (2003), and others. The data is available for individual players, for teams as a whole, and (by adding over teams) entire leagues. In fact, data breakdowns are available by base-out situation, home/away, left/right handed pitcher, day/night, turf/grass, lineup slot, month, defensive position, opposing team, and ballpark as well. Historical data − without all the detailed breakdowns − is available for every player, team, and league from the late 1800's through today at the Baseball Reference[5] (2004) site. Students familiar with databases can download a database of all players from 1871 to the present from the Baseball Archive[6] (2004); from the database, they can easily create team and league sums.

For any data set, a key question is how to translate event probabilities to transition probabilities. Some events e have a simple, deterministic mapping from initial state i to the following state $f_e(i)$; for example, $f_{strikeout}((B,O)) = (B,O+1)$, $f_{triple}((B,O)) = (3,O)$, and

[1] http://www.sportsline.com/mlb/stats
[2] http://sports.espn.go.com/mlb/statistics
[3] http://www.foxsports.com/named/FS/MLB/stats
[4] http://sportsillustrated.cnn.com/baseball/mlb/stats
[5] http://www.baseballreference.com/
[6] http://www.baseball1.com/statistics

$f_{homerun}((B,O)) = (\varnothing, O)$. However, many other events have stochastic results if they happen in certain states. For example, $f_{single}((1,O))$ could be $(12,O)$ if the baserunner stops at second base, or $(13,O)$ if the baserunner advances to third base. (In fact, the outcome could also be $(23,O)$, $(2,O+1)$, or $(1,O+1)$ if the defensive team throws to third base in an attempt to get the baserunner out; these outcomes are much less common, however.)

The variety of possible event results means that students may create models of varying complexity; in fact, we find this to be helpful in class − each student or team of students can present their model, and the class can observe and discuss the relative merits of each one. Some might choose a simple model where baserunner advancement is completely conservative − one base on a single, two on a double, zero on an out. This model was adopted by D'Esopo and Lefkowitz (1960) and also used by Bukiet, Harold, and Palacios (1997). A more realistic model can be created using baserunner advancement probabilities given by Pankin (1993); Pankin also provides data on the frequency of defensive errors.

Once a model has been created, it is easy to test. A reasonable test is to compare the number of runs scored by each team in the league to the number of runs predicted by the model (when using the transition matrix created from that team's data), or to do the same for the entire league. Using the overly-conservative model of D'Esopo and Lefkowitz, the Markov chain usually predicts approximately 7% fewer runs scored than reality; the more realistic model using Pankin's data is usually within 2% of the true number of runs scored. So, while baseball might not be a perfect Markov process, the Markov chain model still provides a reasonable approximation of reality.

Of course, the question (for students) still remains: how do we calculate the expected number of runs predicted by the Markov chain model? In the next two sections, we describe a sequence of questions, progressively more complex, that students can think through and answer using this model; the answer to the first question will also provide a way to calculate the expected number of runs scored.

4. Lesson II: Using the Model to Answer Informational Questions

In this section, we suggest some potential lines of investigation for students using the model. Rather than asking theory-based questions in terms of the Markov chain (e.g., "What are the steady-state probabilities?"), we have had more success selecting questions that pique students' interest in terms of the baseball application and that require the students to incidentally perform the same Markov chain computations they would need to answer theoretically-based questions. These questions also require students to think carefully about the meaning of Markov chain concepts.

We have found that a good first example for the students is to pose a question like **"Suppose a team has the bases loaded and nobody out. How many runs can the team expect to score this inning?"** Once the students' interest has been piqued, the question can be extended to include every base-out situation (in fact, when doing the calculation, all states' expected run values will be found simultaneously anyway). The related question **"How many runs can a team be expected to score in an inning?"**, which can be used to validate the model (see Section 3), can also be answered here.

To answer this question, students will need to think through two issues. First, they must realize which version of the model is appropriate; the variant with an absorbing state "3 outs" is necessary. Second, they need to determine the number of runs scored in each state transition. Although this is not a Markov chain calculation *per se*, we have found it to be a valuable, and straightforward, modeling experience. Because each baserunner and the batter in the current state must be accounted for (either as a baserunner, an out, or scoring a run) in the next state, the number of runs t_{ij} scored in the transition between states i and j is simply

$$t_{ij} = 1 + (|B_i| + O_i) - (|B_j| + O_j), \qquad (1)$$

where $|B_i|$ denotes the number of baserunners in state i.

Once the students have understood this formula, the expected run value v_i of each state i is easy to do using standard, straightforward Markov chain calculations:

$$v_i = \sum_j p_{ij}(t_{ij} + v_j), \ \forall \ i \neq \text{"3 outs"}, \qquad (2)$$

$$v_{3outs} = 0. \qquad (3)$$

For the purpose of validating the model, we need to consider $v_{(\varnothing,0)}$. This value is the expected number of runs scored per inning, which we can compare to the runs/inning observed from the real data.

A natural followup to finding the value of each state is to ask a question like "How much is a home run worth?" To start off, "**Suppose a batter comes to bat with the bases loaded and no outs, and hits a grand slam home run. How many runs is it worth to the team?**"

For whatever reason, this seemingly innocuous question is often the key to the entire lesson. We have observed that when presented with this question initially, students almost always answer the obvious "4" (or else they realize there must be a more complex answer − why else would we ask? − but they have no idea what it is). The telltale answer "4" indicates that the students are not yet thinking about the model together with the application, but only about the application itself. Our experience has been that the simple answer to this question is often the key to the entire lesson − the "turning point" where students begin to think in terms of the model. Thinking through the answer

$$4 + v_{(\varnothing,0)} - v_{(123,0)} \qquad (4)$$

seems to start students thinking about how the model shows something about reality that isn't immediately intuitive, and how it gives them insight that many of their familiar television sportscasters don't have.

A logical extension of calculating the expected value of a home run in one state is to "calculate the expected value of a home run in general." A home run in any state (B,O) yields a next state of $(0,O)$ with $|B| + 1$ runs scored, so the expected value u_{HR} of a home run is

$$u_{HR} = \sum_B \sum_O \pi_{(B,O)} (|B| + 1 + v_{(\varnothing,O)} - v_{(B,O)}). \qquad (5)$$

In order to finish this calculation, students first need to find the steady-state probability vector π . Thus, even though π might not be of independent interest in this model, the students still get reinforcement of the basic Markov chain steady state equations

$$\pi = \pi P, \qquad (6)$$

$$\sum_B \sum_O \pi_{(B,O)} = 1. \qquad (7)$$

In addition to calculating the expected value of a home run, students can also "**calculate the expected value of other major events included in the model (singles, doubles, triples, walks, etc.).**" Depending on the amount of detail in each student's model, this might include a double probability sum, since the outcome of an event might itself be stochastic. Denoting q_{ije} as the probability of moving to state j after event e occurs in state i, the expected value u_e of any event e can be calculated as

$$u_e = \sum_i \pi_i \left(\sum_j q_{ije} (t_{ij} + v_j - v_i) \right). \qquad (8)$$

Once the students have calculated the expected value of each event, they can begin comparing players. "**How much more (or less) valuable than average was the batting contribution of each player on the local Major League team?**"

If we denote z_{ke} as the fraction of plate appearances of player k in which event e happened, then the relative value y_k of player k is

$$y_k = \sum_e z_{ke} u_e \qquad (9)$$

In fact, calculations (8) and (9) are the basis of Thorn and Palmer's (1984) linear weights method of player evaluation.

Given that the data z_{ke} are readily available (for most important events e) in tabular form that can easily be imported to a spreadsheet, it takes almost no extra effort to extend this question to the entire league. Data for the American and National Leagues from 2001-2003 is available on our supplementary web site[7].

[7] http://www.isye.gatech.edu/~jsokol/markovball

5. Lesson III: Using the Model to Answer Prescriptive Questions

At this point, after spending a good deal of time working with expected values and ignoring the effect of each player's teammates, we find it helpful to reground the students in the real-world scenario. The event values u_e assume a league-average probability of each base-out situation; however, a player's teammates − specifically, those who bat immediately before − have a large effect on the base-out distribution seen by the player. The base-out distribution, in turn, affects the event values for each player.

To let the students gain some intuition for this concept, we give the students a challenge. "**Find the best lineup you can for the 9 most frequent starters (one per position) on the local Major League team. The student with the best lineup (the one that scores the highest average number of runs per game) will win a prize.**" This task requires the students to construct transition matrices for each of the 9 players, and extend the model to cover 9 innings (9x24 states, plus one absorbing "game over" state). By iteratively applying transition matrices in the same order as their lineup, they can simulate a game and calculate the expected number of runs scored. When teaching students who have programming experience, we ask them to perform these tests on their own; for classes where the students are not proficient programmers, we provide some Matlab code that runs the simulation for them. Either way, the task of finding a good ordering of the batters is left to the students. The Matlab code and a sample data set[8] (for the 2001 Atlanta Braves) are available online for readers.

The most important outcome we have observed from this final exercise is that students develop a good intuitive feel for the process. Many students, especially those who follow baseball from the perspective of a fan rather than an OR/MS practitioner, initially construct "traditional" baseball lineups − a fast basestealer first followed by a contact hitter second (regardless of how often they get on base), for example, with the pitcher batting last. After some experimentation and thought, they realize that the expected value of all home runs their third batter hits is lower than what they calculated for u_{HR}, because the three batters immediately before that third batter (#9, #1, and #2) are on base less frequently than average. This insight starts the students down a more productive track of thinking about the underlying run-scoring process.

Perhaps because they approach the lesson without preconceived ideas about what makes a good batting order, students who are not baseball fans (or who are unfamiliar with baseball altogether) often provide the best answers and the best intuitive explanations. Regardless of their level of interest in or knowledge of baseball, students' reactions to this lesson are almost always positive. They like the idea of applying something they have learned to a "common" situation outside the working world, and even those students who are not baseball fans seem to appreciate the idea that understanding Markov chains can give them a better understanding of baseball than the average fan. This sentiment is usually reflected in the last part of the module; our final in-class discussion often consists of students teaching the lesson, describing how some batters are good at helping the team reach high-value states, and other batters are good at helping the team obtain good value from each state. More importantly, in the process of analyzing baseball using their Markov chain model, the students have learned to think about this and future Markov chain applications in a creative, open-minded way.

References

The Baseball Archive (2004), http://www.baseball1.com/statistics/

Baseball Reference.com (2004), http://www.baseballreference.com/

Bukiet, B., E. R. Harold, and J. Palacios (1997), "A Markov Chain Approach to Baseball," *Operations Research,* Vol. 45, pp. 14-23.

CBS Sportsline (2003), MLB Stats, http://www.sportsline.com/mlb/stats

Cook, E. (1964), *Percentage Baseball,* MIT Press, Cambridge, MA.

Cramer, R. D. (1977), "Do Clutch Hitters Exist?," *Baseball Research Journal,* Vol. 6, pp. 74-79.

Denardo, E. V. (2002), *The Science of Decision Making,* John Wiley and Sons.

[8] http://www.isye.gatech.edu/~jsokol/markovball

D'Esopo, D. A., and B. Lefkowitz, (1960), "The Distribution of Runs in the Game of Baseball," *SRI Internal Report.*

Durrett, R. A. (1995), *Probability: Theory and Examples,* Duxbury Press.

ESPN (2003), MLB Statistics Index, http://sports.espn.go.com/mlb/statistics

Fox Sports (2003), MLB Stats, http://www.foxsports.com/named/FS/MLB/stats

Grabiner, D. (1993a), http://remarque.org/~grabiner/fullclutch.txt

Grabiner, D. (1993b), http://remarque.org/~grabiner/risp91.txt

Grabiner, D. (1997), http://remarque.org/~grabiner/protstudy.txt

Hillier, F. S., and Lieberman, G. J. (2002), *Introduction to Operations Research,* 7th ed. McGraw-Hill.

Howard, R. A. (1960), *Dynamic Programming and Markov Processes,* MIT Press, Cambridge, MA.

Pankin, M. D. (1991), *Finding Better Batting Orders,* SABR XXI, New York. http://www.pankin.com/markov/btn1191.htm

Pankin, M. D. (1993), *Subtle Aspects of the Game,* SABR XXIII, San Diego. http://www.pankin.com/markov/sabr23.htm

Sokol, J. S. (2003), "A Robust Heuristic for Batting Order Optimization," *Journal of Heuristics,* Vol. 9, pp. 353-370.

Sports Illustrated (2003), MLB Statistics, http://sportsillustrated.cnn.com/baseball/mlb/stats/

Stern, H. S. (1997), "A Statistician Reads the Sports Page: Baseball by the Numbers," *Chance,* Vol. 10, p. 38.

Thorn, J., and P. Palmer (1984), *The Hidden Game of Baseball: A Revolutionary Approach to Baseball and Its Statistics,* Doubleday, Garden City, New York.

Winston, W. L. (1997), *Operations Research: Applications and Algorithms,* 3rd ed. Duxbury Press.

Appendix

Sample Data and Code to Accompany "An Intuitive Markov Chain Lesson From Baseball"

1. Summary

The files below contain Matlab code and a sample data set (the 2001 Atlanta Braves) which students and instructors can use as a supplement to the Markov chain lesson described in the paper "An Intuitive Markov Chain Lesson From Baseball" (Sokol J.S. (2004), "An Intuitive Markov Chain Lesson From Baseball," *INFORMS Transactions on Education,* [9]). The code and data set use a simple event model to make it easy for users to create and analyze their own batting order data sets.

2. Files

League data[10]: This file contains league total data that can be used to calculate transition probabilities, situational values, event values, etc. Several leagues' totals are included. Note that only basic data is included here, so that students do not get bogged down in the details of baseball. Results using very detailed data (errors, baserunner advancement probabilities, steals, etc.) are not very different from those obtained with the basic data. The columns in this data file follow the same format as described below in "Creating New Data Sets".

Matlab code[11]: The Matlab files in this zip archive contain the Markov chain calculations necessary to evaluate batting orders.

Data file[12]: This file contains the input data for the 2001 Atlanta Braves. The instructions below describe how users can easily create their own data sets.

3. Usage

Installing and Running the Program

1. Extract all of the files from the zip archive[13] into the directory you plan to use for this application.
2. Copy the data file (or create your own data file) into the same directory.
3. Open Matlab from this directory.
4. Start the program by giving Matlab the command **battingorder datafilename** where "datafilename" is the name of the file containing the data set you wish to use.
5. Each time you are prompted for a batting order, enter it in the form **123456789**, where each number is a player in your data set. (So, 123456789 has the first player in your data set batting first, the second player in your data set batting second, etc. 421356789 has the fourth player in your data set batting first, etc.) Note that this allows you to duplicate players; in fact, by entering an order like 555555555 you can answer the question "how many runs would my team score if everyone was like the fifth player in my data set".

[9] http://ite.pubs.informs.org/Vol5No1/Sokol/
[10] http://ite.pubs.informs.org/Vol5No1/Sokol/leaguetotals.txt
[11] http://ite.pubs.informs.org/Vol5No1/Sokol/matlabfiles.zip
[12] http://ite.pubs.informs.org/Vol5No1/Sokol/braves.data
[13] http://ite.pubs.informs.org/Vol5No1/Sokol/matlabfiles.zip

6. In addition to appearing on your screen, all output will be saved to the file **datafilename.out**, where "datafilename" is the same as above.
7. To end the program, hit enter (without any numbers) at the batting order prompt.

Creating New Data Sets

Data sets should have 9 rows (one for each player) of 7 columns each:

1. Column 1: home runs
2. Column 2: triples
3. Column 3: doubles
4. Column 4: singles
5. Column 5: walks
6. Column 6: at-bats
7. Column 7: player name (or other comments)

As mentioned above, the model used in the Matlab code is simplified to include only these basic events, so that users can easily create their own data sets from current or historical data.

Comments

Please send comments, bugs, etc. to Joel Sokol (jsokol@isye.gatech.edu)