Mathematically Modeling Baseball

Bruce G. Bukiet

Baseball lends itself to mathematical modeling. A play in baseball mainly involves just two people, a pitcher facing a batter. The other players on the field have a much smaller influence. As a first approximation one can ignore the influence of fielding ability. The situation—runners on base, number of outs, current score—is clear before and after the batter's turn. This simplicity sets baseball apart from a sport like basketball with ten players on the court simultaneously passing, picking, dribbling, and guarding before a shot is taken or points are scored.

As a sample for *Notices* readers, we now describe a model we developed, which over the past twenty years has fared well in informal comparisons with other sportswriters.[1] For 2017 the method correctly projected all of the division winners but no wild card teams (see Figure 1).

Bruce Bukiet is associate professor of mathematical sciences and associate dean for undergraduate studies in the College of Science and Liberal Arts at New Jersey Institute of Technology. His email address is bukiet@njit.edu.

[1] *See* web.njit.edu/~bukiet/baseball/season_review_2013.htm *and the annual contests at* Baseballphd.net.

DOI: http://dx.doi.org/10.1090/noti1667

The difference between using the best possible lineup and worst possible lineup is as much as four wins in a season. Before our work, it was thought to be much less.

Trades can be modeled in a straightforward way by swapping players' transition matrices to their new teams. We showed that a fairly good home run hitter is much more valuable (in terms of team wins expected) than an excellent singles hitter.

We verified the claim in Michael Lewis's book *Moneyball* that replacing several strong and weak hitters with an equal number of average hitters should lead to similar performance.

Our model has been used to determine whether a game is worth wagering on, depending on the payoff (for entertainment purposes only, of course). It has been used to compute the relative value of highly paid players, average paid players, and the lowest paid players (in work performed with undergraduate student Iman Kazerani). We have also used the method to evaluate who should win baseball's Most Valuable Player and Cy Young Awards—which player would have added the most wins to a team of average players that season (in work with Kevin Fritz, a high school student at the time [2]).

First attempts. In the late 1980s I attempted with brute force to compute expected baseball scores from hitting data. At that time it wasn't easy to obtain much baseball data beyond "at bats," hits, doubles, triples, home runs, outs, and walks for batters, and wins, losses, strike outs, innings pitched, home runs, hits and walks allowed, and earned run averages for pitchers. A key consideration in modeling baseball is that the order of events matters: a single followed by a home run yields two runs, while the reverse yields just one run immediately. Reducing the batting data to the probability of just six events—walks, singles, double, triples, home runs, and outs—and using a simple model for runner advancement (to be described later), I programmed a computer to enumerate all possible sequences of events that the lineup could experience to get to 27 outs. I quickly learned that analyzing a single lineup would take many years. Since about 40 plate appearances occur

Simplicity sets baseball apart.

Team	Projected	Actual
BOS	91	93
NYY	80	91
TB	76	80
TOR	90	76
BAL	77	75

Team	Projected	Actual
CLE	99	102
MIN	69	85
KC	74	80
CHW	64	67
DET	86	64

Team	Projected	Actual
HOU	94	101
ANA	79	80
SEA	80	78
TEX	80	78
OAK	78	75

Team	Projected	Actual
WSH	97	97
MIA	72	77
ATL	74	72
NYM	92	70
PHI	69	66

Team	Projected	Actual
CHC	104	92
MIL	71	86
STL	88	83
PIT	80	75
CIN	62	68

Team	Projected	Actual
LAD	104	104
ARI	69	93
COL	83	87
SD	53	71
SF	95	64

Figure 1. For 2017 the method correctly projected all of the division winners but no wild card teams.

for a team in a typical 9-inning game, over 640 sequences would have to be analyzed for a given lineup. This would have taken more than a quadrillion years on a late 1980s computer. I learned that brute force methods may be easy to code but impossible to run in reasonable time. I needed to streamline the computation.

easy to code but impossible to run in reasonable time

A Markov process model. Several researchers (e.g. Bellman [1] and Trueman [4]) had considered baseball as a Markov process in order to study managerial decision-making, such as when bunting or stealing is worthwhile. In a Markov process, the probability of the next state depends only on the current state, not on the history. In baseball there are 3 x 8 + 1 = 25 states for the batting team: 0, 1, or 2 outs times 8 base-runner situations (no one on base, man on first, … , bases loaded) plus the final "absorbing" 3-out state. For every batter, pitcher, and perhaps other factors, we can develop a 25x25 matrix representing the transition probability of moving from one of these states to another.

For demonstration purposes, we make certain simplifying assumptions:

(1) on a walk, runners advance if forced;
(2) on a single, a runner on first advances to second base while other runners score;
(3) on a double, a runner on first base advances to third base and other runners score;
(4) on a triple, all base runners score;
(5) on a home run, all base runners and the batter score; and
(6) on an out, runners do not advance.

The transition matrix, P, can be written:

$$P = \begin{pmatrix} A & B & C & D \\ 0 & A & B & E \\ 0 & 0 & A & F \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

where

$$A = \begin{pmatrix} P_H^{(1)} & P_S + P_W & P_D & P_T & 0 & 0 & 0 & 0 \\ P_H^{(2)} & 0 & 0 & P_T^{(1)} & (P_S + P_W)^{(1)} & 0 & P_D & 0 \\ P_H^{(2)} & P_S^{(1)} & P_D^{(1)} & P_T^{(1)} & P_W & 0 & 0 & 0 \\ P_H^{(2)} & P_S^{(1)} & P_D^{(1)} & P_T^{(1)} & 0 & P_W & 0 & 0 \\ P_H^{(3)} & 0 & 0 & P_T^{(2)} & P_S^{(1)} & 0 & P_D^{(1)} & P_W \\ P_H^{(3)} & 0 & 0 & P_T^{(2)} & P_S^{(1)} & 0 & P_D^{(1)} & P_W \\ P_H^{(3)} & P_S^{(2)} & P_D^{(2)} & P_T^{(2)} & 0 & 0 & 0 & P_W \\ P_H^{(4)} & 0 & 0 & P_T^{(3)} & P_S^{(2)} & 0 & P_D^{(2)} & P_W^{(1)} \end{pmatrix}$$

$B = P_{out}\, I_8$; $C = 0_{8x8}$; $D = 0_8$; $E = 0_8$; $F = P_{out}$.

Here P_W, P_S, P_D, P_T, P_H, P_{out} are the probabilities of the batter walking, and getting a single, double, triple, home run, or out, respectively, and must sum to 1. I_8 is the 8x8 identity matrix. Superscripts indicate the number of runs, if any, scored on the particular transition. *D*, *E*, and *F* are column vectors with 8 entries representing transitions arising from going from zero outs to three outs (triple plays), from one out to three outs (double plays), and from two outs to three outs, respectively. Since we ignore double and triple plays, the *C*, *D*, and *E* submatrices are all zero. The

Batting orders and win probabilities interest far more people than my research in detonation theory.

A submatrix represents transitions where no outs occur. The first row of *A* represents transitions from no one on to no one on, man on first, man on second, man on third, men on first and second, men on first and third, men on second and third, and bases loaded, respectively. The other rows of *A* represent transitions *to* these states *from* man on first, man on second, etc. The "1" on the bottom corner represents the absorbing 3-out state.

Games start with no one on and no one out, which we represent by a row vector, **U**, with 25 entries—the first is 1 and the rest are zero. By multiplying $\mathbf{U}P_1$ we find the probability of being in any situation after the first batter's plate appearance. Multiplying the resulting row vector by P_2 gives the probabilities after the second batter. (Here P_1, etc. represent the transition matrix for the batter in that position in the lineup). Going through the lineup in this manner and keeping track of runs scored and returning to the no on, no out state when 3 outs are reached and keeping track of the inning, gives the probability of the lineup scoring 0, 1, 2, ... runs during the 9-inning game.

The structure given above allows for tremendous flexibility. One may revise the runner advancement model, for example by considering that, on a single, a runner on first may stop at second, or at third, or even score or be thrown out; so the probability of getting a single can be partitioned based on actual or model data. Researchers (e.g. Hirotsu and Wright [3]) have included dependence on balls, strikes, inning, or score leading to transition matrices with over one million rows and columns. Finding sufficient data to set the entries of such a transition matrix appropriately may be problematic.

Using the method above improved the very long brute force computation method from quadrillions of years down to under 1.5 seconds in the early 1990s. Other researchers have included dependence on balls, strikes, inning, or score. These analyses lead to transition matrices with over one million rows and columns. Yet, the structure described above, with the 25x25 core transition matrix for each batter, yields many interesting results, including, for example, about optimal line-ups. One manager quipped that he would use all possible lineups in spring training and then decide which one to use during the season; for nine batters, there are 9!=362,880 possible lineups. Clearly, this is not possible. Computing the expected number of runs for each lineup shows that the best possible lineup should have the "slugger" bat second or third and the pitcher (who is part of the lineup in the National League) bat seventh or eighth.

Recent years have seen a great increase in appreciation for the utility of math and statistics to improve team performance, and MLB teams are reported to have developed statistics and analytics groups. New technologies have led to new metrics. The model described above can indirectly incorporate the influence of new metrics like launch angles, exit velocity, and opportunity time.

Through this work I have been able to bring an appreciation of the value of math to a wide group of people: batting orders and win probabilities interest far more people than my research in detonation theory. I've gained experience speaking to the media, learning just to promote math's value and power and to avoid getting bogged down in the statistical nuances. I have used math modeling of baseball as a hook to recruit students to pursue math majors and minors and have provided high school and college students with research opportunities, leading to several papers and presentations with students. Recently an undergraduate student, Kelvin Rivera, performed an independent study project demonstrating that our model could be used for football, something for decades I didn't think could work. You never know the next amazing insight you'll get from pursuing math modeling.

References

[1]R. Bellman, Dynamic programming and Markovian decision process, with application to baseball, *Optimal Strategies in Sports*, eds. S. P. Ladany and R. E. Machol, Elsevier-North Holland, New York, 1977.

[2]Bruce Bukiet and Kevin Fritz, Objectively Determining Major League Baseball's Most Valuable Players, *International Journal for Performance Analysis in Sport*, vol. 10, no. 2, 152-169.

[3]Nobuyoshi Hirotsu and Mike Wright, Modelling a baseball game to optimise pitcher substitution strategies incorporating handedness of players, *IMA J. Manag. Math.* **16**, 179–194, 2005, MR2133438.

[4]Trueman, R. E., Analysis of baseball as a Markov process, *Optimal Strategies in Sports*, eds. S. P. Ladany and R. E. Machol, Elsevier—North Holland, New York, 1977.

Photo Credits

Aerial photo of baseball field by Tom Gouw; `https://www.pexels.com`.

Photo of Bruce Bukiet courtesy of New Jersey Institute of Technology.

ABOUT THE AUTHOR

Bruce Bukiet has employed math to make contributions toward better understanding of baseball, biology, bombs, and bugs. He currently works with colleagues on projects to increase the number of women pursuing study and careers in STEM.

Bruce Bukiet