Markov Chains in the Game of Baseball

Will Bliss

5/10/2022

Introduction

The goal of this project is to understand the usage of Markov Chains in baseball and how they can give a team competitive advantages.

The average revenue for a baseball team associated the Major League Baseball is $122.1 million. However, one can expect better performing teams bring in more money. Thus, teams have begun looking at all ways possible to get the most amount of wins during their season. One way to gain a competitive edge is to view the game from a statistical standpoint. Markov Chains have been brought into the game of baseball as a way to predict outcomes in order to make better strategic choices during play.

While much more can be done with Markov Chains to compute event probabilities for a baseball game than what is discussed below, it is important to see the foundation for such calculations.

Methods

Baseball is an excellent application for Markov Chains because baseball is a game comprised of states. There are always 0, 1, 2, or 3 outs and there are always 0, 1, 2, or 3 runners on base at a given time.

More precisely, there are eight possible baserunner positions (three bases, each of which can be occupied or not, for a total of $2^3 = 8$ possibilities). Then, with either 0, 1, or 2 outs during these 8 baserunner possibilities, there is a total of $8 \cdot 3 = 24$ states. However, there is the situation where the 3^{rd} out is recorded, which ends the inning. Thus, we create states 25, 26, 27, and 28 for when the third out is recorded with 0, 1, 2, or 3 runners left on base. For example, state 25 is the state where the third out has been recorded and there were 0 runners left on base.

We can implement the Markov property, saying that the probability of moving to the next state in an inning of a baseball game is determined only by the current state of the inning, and not how the team got to that state. Then, to satisfy the requirements for a Markov Chain, we need a finite number of states and a transition matrix with positive entries where the row sums are equal to 1. A Markov chain can be created for an inning of a baseball game satisfying these criteria. We have our finite (28) number of states, and can create a transition matrix determining how the process will move from state to state. However, it is impossible to give the exact probabilities of transitions. What can be done, though, is to create a transition matrix calculated from real life outcomes over many games.

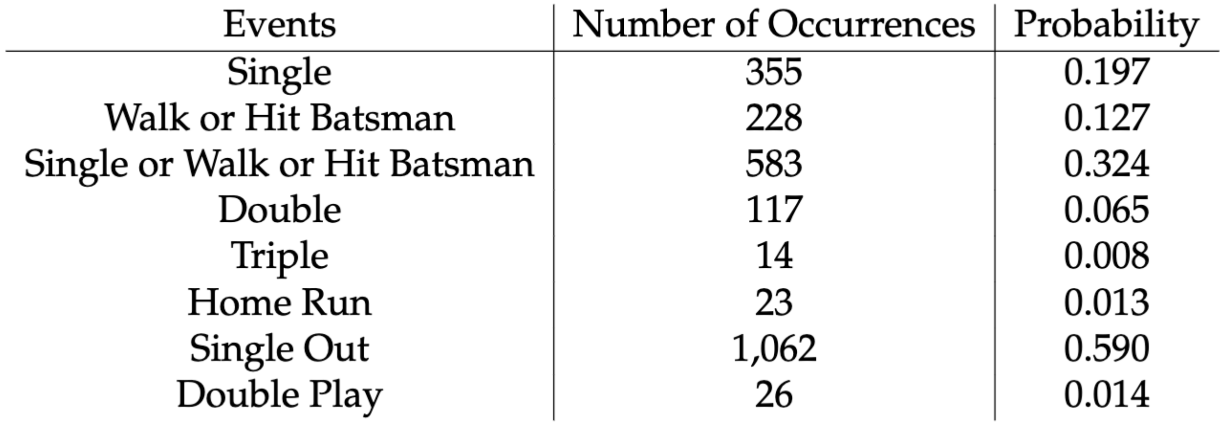
First, let's formally define each state:

State	Bases Occupied	Number of Outs
1	None	0
2	1st	0
3	2nd	0
4	3rd	0
5	1st and 2nd	0
6	1st and 3rd	0
7	2nd and 3rd	0
8	1st, 2nd, and 3rd	0
9	None	1
10	1st	1
11	2nd	1
12	3rd	1
13	1st and 2nd	1
14	1st and 3rd	1
15	2nd and 3rd	1
16	1st, 2nd, and 3rd	1
17	None	2
18	1st	2
19	2nd	2
20	3rd	2
21	1st and 2nd	2
22	1st and 3rd	2
23	2nd and 3rd	2
24	1st, 2nd, and 3rd	2
25	None	3
26	1 base	3
27	2 bases	3
28	3 bases	3

Observe that states 25 through 28 are absorbing states, since an inning ends when 3 outs are recorded. We only note how many baserunners are on since their placement doesn't matter with the inning being over.

Results

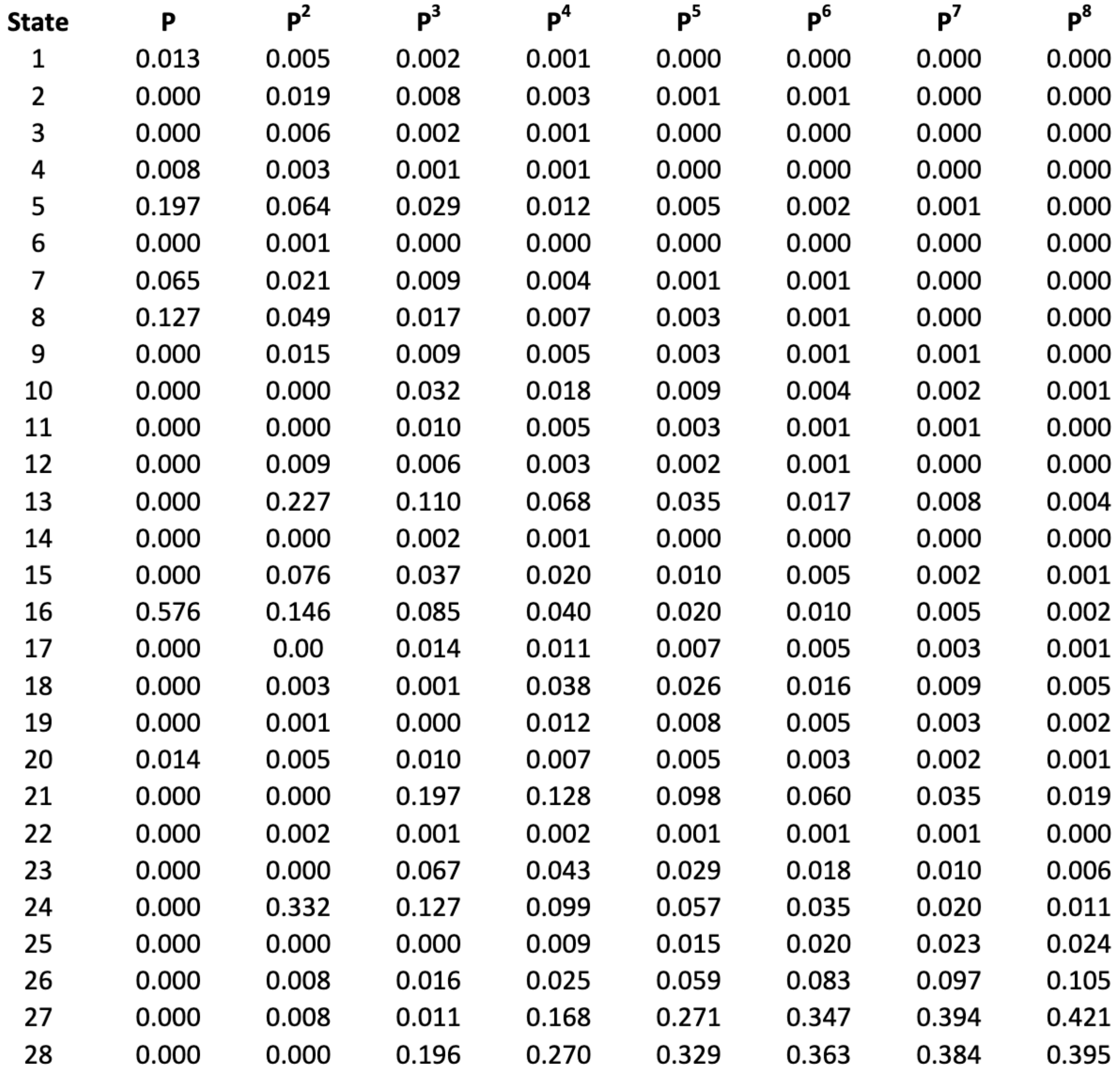
Let's look into the probabilities of certain events for a given team. Observe "Markov Chain Theory with Applications to Baseball" by Cal D. Thomay of the College of Wooster. In his thesis, he provides data for the school's baseball team, seen in the figure below:



Events	Number of Occurrences	Probability
Single	355	0.197
Walk or Hit Batsman	228	0.127
Single or Walk or Hit Batsman	583	0.324
Double	117	0.065
Triple	14	0.008
Home Run	23	0.013
Single Out	1,062	0.590
Double Play	26	0.014

Figure 1: Probabilities of events for the College of Wooster in 2013

With these event probabilities recorded, Thomay then computed the transistion matrix. As this results in a 28×28 matrix, it is difficult to provide the full matrix. Below is a display of the transition probabilities from the 8th state:



State	P	P^2	P^3	P^4	P^5	P^6	P^7	P^8
1	0.013	0.005	0.002	0.001	0.000	0.000	0.000	0.000
2	0.000	0.019	0.008	0.003	0.001	0.001	0.000	0.000
3	0.000	0.006	0.002	0.001	0.000	0.000	0.000	0.000
4	0.008	0.003	0.001	0.001	0.000	0.000	0.000	0.000
5	0.197	0.064	0.029	0.012	0.005	0.002	0.001	0.000
6	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000
7	0.065	0.021	0.009	0.004	0.001	0.001	0.000	0.000
8	0.127	0.049	0.017	0.007	0.003	0.001	0.000	0.000
9	0.000	0.015	0.009	0.005	0.003	0.001	0.001	0.000
10	0.000	0.000	0.032	0.018	0.009	0.004	0.002	0.001
11	0.000	0.000	0.010	0.005	0.003	0.001	0.001	0.000
12	0.000	0.009	0.006	0.003	0.002	0.001	0.000	0.000
13	0.000	0.227	0.110	0.068	0.035	0.017	0.008	0.004
14	0.000	0.000	0.002	0.001	0.000	0.000	0.000	0.000
15	0.000	0.076	0.037	0.020	0.010	0.005	0.002	0.001
16	0.576	0.146	0.085	0.040	0.020	0.010	0.005	0.002
17	0.000	0.00	0.014	0.011	0.007	0.005	0.003	0.001
18	0.000	0.003	0.001	0.038	0.026	0.016	0.009	0.005
19	0.000	0.001	0.000	0.012	0.008	0.005	0.003	0.002
20	0.014	0.005	0.010	0.007	0.005	0.003	0.002	0.001
21	0.000	0.000	0.197	0.128	0.098	0.060	0.035	0.019
22	0.000	0.002	0.001	0.002	0.001	0.001	0.001	0.000
23	0.000	0.000	0.067	0.043	0.029	0.018	0.010	0.006
24	0.000	0.332	0.127	0.099	0.057	0.035	0.020	0.011
25	0.000	0.000	0.000	0.009	0.015	0.020	0.023	0.024
26	0.000	0.008	0.016	0.025	0.059	0.083	0.097	0.105
27	0.000	0.008	0.011	0.168	0.271	0.347	0.394	0.421
28	0.000	0.000	0.196	0.270	0.329	0.363	0.384	0.395

Figure 2: Transition probabilities from the 8th state to any of the 28 states in the baseball transition matrix and powers of the transition matrix

For example, the inning can go from state 8 to 5 with a single. It is shown in Figure 1 that this happens with probability 0.197, which appears again in the transition matrix **P**. The other calculations were made with similar decisions. When looking at this table, the 0.000 probability in the 10th state under the column **P**, for example, means that there is no probability of transitioning from state 8 to state 10 in one step. However, in this same row there is a 0.032 probability under the column $\mathbf{P^3}$. This means that transitioning from state 8 to state 10 after three steps has a probability of 0.032.

We can take the full matrix and reorder the states in canonical decomposition, such that the transition matrix is in block matrix form:

$$\mathbf{P} = \begin{bmatrix} \mathbf{P_T} & \mathbf{R} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$$

where

- $\mathbf{P_T}$ is a 24×24 submatrix of transient → transient transition probabilities
- $\mathbf{R}$ is a 24×4 submatrix of transient → recurrent transition probabilities
- $\mathbf{0}$ is a 4×24 matrix of 0's
- $\mathbf{I}$ is a 4×4 identity matrix

Let B be the number of batters who come to the plate in an inning, let R be the number of runs scored, and let L be the number of runners left on base at the end of the inning. Then, $B = 3 + R + L$. This can be equivalently written as $R = B - L - 3$. Taking the expected value of this equation, we are then left with $E(R) = E(B) - E(L) - 3$. Therefore, if we find the expected number of batters that appear in an inning and the expected number of runners left on base during that inning, we can find the expected number of runs for an inning.

Note the following theorem: Consider an absorbing Markov chain. The matrix $\mathbf{I} - \mathbf{Q}$ has an inverse $\mathbf{N}$, with $\mathbf{N} = \mathbf{I} + \mathbf{Q} + \mathbf{Q^2} + \cdots$. The entry n_{ij} of the matrix $\mathbf{N}$ is the expected number of times the chain is in state s_j, given that it started in state s_i. If $i = j$, then the initial state is counted.

Thus, in baseball terms, the sum of the i^{th} row of $\mathbf{N}$ is the expected number of batters that will come to bat during the remainder of the inning starting from state s_i. Thomay uses this theorem to show that $E(B)$ is the sum of the first row of $\mathbf{N}$, since we are only looking for the expected number of batters to come to the plate starting from the beginning of an inning. Thomay states that he used Matlab and Microsoft Excel to compute the matrix $\mathbf{N}$, which results in $E(B) = 5.051$.

Then, to calculate R, Thomay uses a matrix $\mathbf{B} = \mathbf{NR}$. The entry b_{ij} of $\mathbf{B}$ gives the probability that an absorbing Markov Chain will be absorbed in an absorbing state s_j, given that the chain starts in the transient state s_i. Thus, for our model, the i^{th} row of $\mathbf{B}$ gives the probabilities of the process being aborbed with 0, 1, 2, or 3 runners left on base, starting from state s_i. This allows us to use the first row of $\mathbf{B}$ to calculate the expected number of runners left on base starting from the beginning of an inning. Thomay uses this to calculate $E(L) = 1.267$.

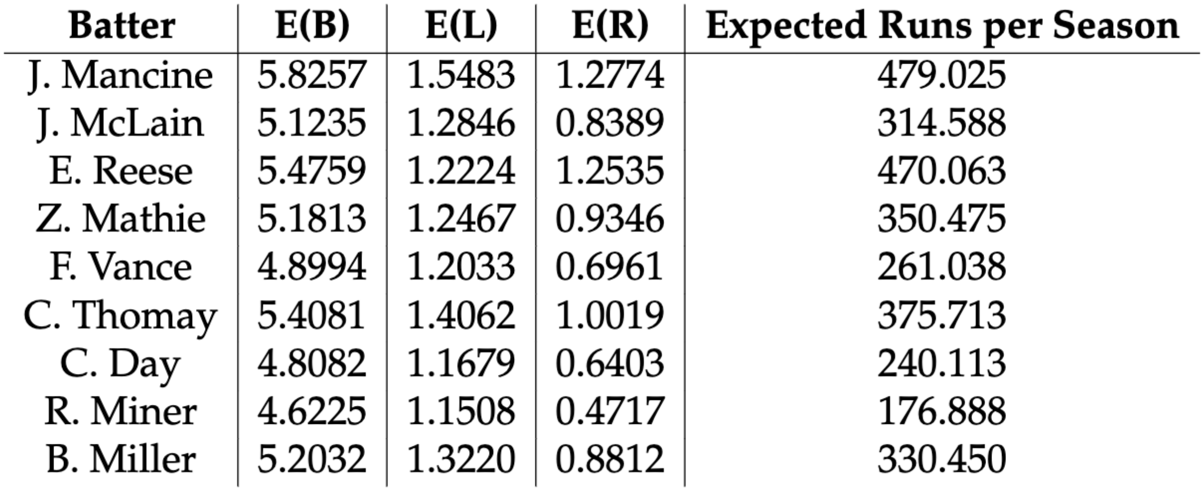
Thus, we have

$$\begin{aligned} E(R) &= E(B) - E(L) - 3 \\ &= 5.051 - 1.267 - 3 \\ &= 0.784 \end{aligned}$$

In other words, we expect the team to score 0.784 runs per inning. The College of Wooster baseball team expects to play 375 innings during the season. Thus, the model expects the team to score $0.784 \cdot 375 = 294$ runs for the season.

Discussion

Instead of team event probabilities listed in Figure 1, we can use an individual player's history to compute an individual's probability of certain outcomes. We can compute an individual player's transition probability matrix, similarly to the team computation. From the player's transition probabilities, we can compute the expected number of runs for an individual. This is important as it can provide useful insight to which players to put into your lineup. For example, Thomay calculated the expected number of runs for an individual on the College of Wooster baseball team over a 375 inning season:

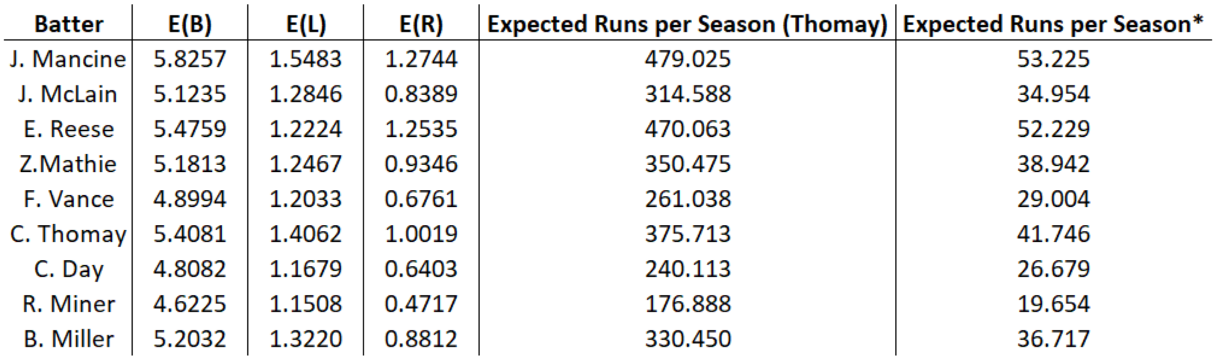


Batter	E(B)	E(L)	E(R)	Expected Runs per Season
J. Mancine	5.8257	1.5483	1.2774	479.025
J. McLain	5.1235	1.2846	0.8389	314.588
E. Reese	5.4759	1.2224	1.2535	470.063
Z. Mathie	5.1813	1.2467	0.9346	350.475
F. Vance	4.8994	1.2033	0.6961	261.038
C. Thomay	5.4081	1.4062	1.0019	375.713
C. Day	4.8082	1.1679	0.6403	240.113
R. Miner	4.6225	1.1508	0.4717	176.888
B. Miller	5.2032	1.3220	0.8812	330.450

Figure 3: Individual batter's expected number of runs scored

According to the base model, the team expects to score 294 runs per year. According to the individual player model, six of the nine batters have an expected run total higher than the team's run expected run total.

I would like to note that these numbers may be off a little. It seems as though the expected values were calculated as per game, but listed as per inning. I suggest this because it seems rare to expect 5 or 6 at bats per inning, but 5 to 6 at bats per game seem much more reasonable. Thus, I would expect all of these values to be divided by 9, since there are 9 innings in a game. If we would like to consider the values to be per game, we get the following, fixed table.



Batter	E(B)	E(L)	E(R)	Expected Runs per Season (Thomay)	Expected Runs per Season*
J. Mancine	5.8257	1.5483	1.2744	479.025	53.225
J. McLain	5.1235	1.2846	0.8389	314.588	34.954
E. Reese	5.4759	1.2224	1.2535	470.063	52.229
Z.Mathie	5.1813	1.2467	0.9346	350.475	38.942
F. Vance	4.8994	1.2033	0.6761	261.038	29.004
C. Thomay	5.4081	1.4062	1.0019	375.713	41.746
C. Day	4.8082	1.1679	0.6403	240.113	26.679
R. Miner	4.6225	1.1508	0.4717	176.888	19.654
B. Miller	5.2032	1.3220	0.8812	330.450	36.717

Figure 4: Individual batter's expected number of runs scored; *: fixed by dividing expected runs per season by 9

For example, we expect J. Mancine to score about 53 runs over a 40 game (360 inning) season.

Baseball has long been thought to be a game of chance where individual performance is what decides the outcome of the game. However, statistics have allowed managers to edit lineups and put players in better situations to win more games. While player performance still matters, it is always better to see firm numbers and get the best possible probability of winning.

Acknowledgements

Dr. Deena Schmidt

Jakob Lovato

References

Calestini, L. (2018, September 10). The Elegance of Markov Chains in Baseball. Medium. Retrieved May 9, 2022, from https://medium.com/sports-analytics/the-elegance-of-markov-chains-in-baseball-f0e8e02e7ac4

Statshacker. (2018, July 26). The Markov Chain Model of Baseball. Statshacker. Retrieved May 9, 2022, from http://statshacker.com/blog/2018/05/07/the-markov-chain-model-of-baseball/

Thomay, Cal D., "Markov Chain Theory with Applications to Baseball" (2014).Senior Independent Study Theses.Paper 5722.https://openworks.wooster.edu/independentstudy/5722

Ursin, Daniel Joseph, "A Markov Model for Baseball with Applications" (2014).Theses and Dissertations. 964.https://dc.uwm.edu/etd/964