An Analysis of the Count in Major League Baseball Using Markov Chains

Mychelle Hale

Advisor: Dr. Fogel

California Lutheran University

Math 475 Capstone

December 13, 2016

1 Abstract

In the game of baseball, batters are taught not to swing when they have two balls and no strikes, that is, a 2-0 count (pronounced two-oh). We use Markov chains and statistical analysis to analyze if there is mathematical support for this advice.

Using Major League Baseball (MLB) data to determine individual batting ability, we split the batters into quartiles and define the average batter in the MLB. We model how batters move up the count, i.e. 2-0 to 2-1 or 2-0 to 3-0, and determine the probability of doing so with play-by-play data. This forms our Markov chain for the average batter and is used to determine the probability of an average batter getting on base. Repeating the process for other batters enables us to compare batters of different ability and determine the variance between batters in different quartiles. We then draw our conclusion of whether a batter should swing on a 2-0 count and about the variances between batters of different quartiles.

2 Background

In baseball, there are two teams, home and away. The objective of the game is to score more runs than the opposing team. Often in youth softball and baseball, batters are told not to swing on a 2-0 count, meaning don't swing when there are two balls and no strikes. This is a tactic batters and coaches use to force the pitcher to refocus and throw a strike or to be so distracted by their poor performance that he or she will walk the batter. The idea behind this is that if the pitcher has thrown two balls and no strikes, there is a higher probability that he will throw a ball next. In this paper, we use Markov Chains to determine if there is any mathematical support for this advice.

2.1 Baseball Terms

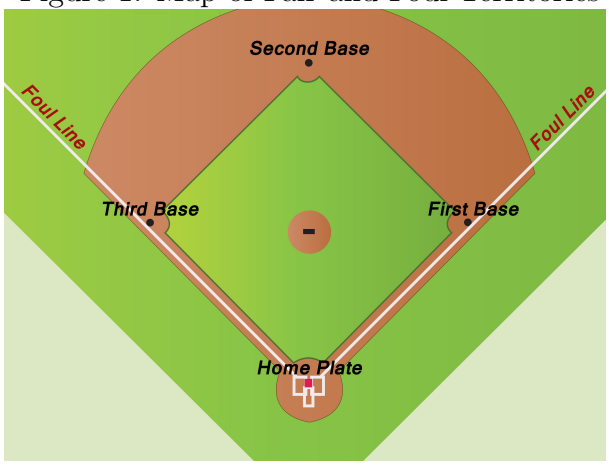
We assume most readers are familiar enough with baseball to know the most basic rules of the game and that a batter hits a ball and runs to first base, then second, third and home. When trying to hit the ball, the batter can get a hit, foul ball, foul tip, ball or strike. After

three strikeouts or four balls, the batter's plate appearance ends. Below we describe some more complicated terms relevant to baseball that are essential to understanding aspects of the project. These terms are relevant to modeling and analyzing the game.

The Count: Refers to the number of balls and strikes a pitcher has thrown during the batter's current plate appearance. The number on the left is the number of balls and the number of strikes is on the right. Note they are listed in alphabetical order. For example, 3-2 would be read as "three-two" and means that the batter has three balls and two strikes.

On the field, there is fair territory which is inside of the foul lines and poles, and foul territory outside the base lines and poles. In the diagram below (Figure 1), we visualize the baseball diamond and the playable fielding area with fair territory between the white base lines and foul outside of them. The following definitions explain what happens when the ball is hit into fair or foul territory.

Figure 1: Map of Fair and Foul Territories



Foul ball: A foul ball is called when the batter hits the ball outside of the fair territory. There are several scenarios, which the MLB defines on its website, that can lead to foul balls [9]. If a batter hits a foul ball with fewer than two strikes on him, then the foul ball is counted as a strike. As a result, we code the first two fouls "s" in our R program. If the batter hits a foul ball with two strikes on him, he can have an infinite number of foul balls.

Foul tip: A foul tip is a batted ball that goes sharply and directly to the catcher's hand or glove and is legally caught. A foul tip is considered equivalent to a ball in which the batter swings or bunts and misses. Should the batter produce a foul tip after previously accruing two strikes, the foul tip is considered strike three and the batter is out.

Hit By Pitch: The batter is touched by a pitched ball which he is not attempting to hit unless:

- the ball is in the strike zone when it touches the batter, or
- the batter makes no attempt to avoid being touched by the ball.

This is important because it contributes to the count by either allowing the batter to get on base or counting as a strike if the ball is in the strike zone when it hits the batter. It is used during the computation of the On Base Percentage Plus Slugging.

Slugging: Slugging percentage represents the total number of bases a player records per at-bat. Unlike on-base percentage, slugging percentage deals only with hits and does not include walks and hit-by-pitches in its equation. This gives an indication of how well the batter hits for power and on average.

On Base Percentage Plus Slugging (OPS): OPS adds on-base percentage and slugging percentage to get one number that unites the two. It's meant to combine how well a hitter can reach base, with how well he can hit for average and for power. As a result, OPS is widely considered one of the best evaluative tools for hitters.

2.2 Assumptions About Baseball

Here we list some important rules of baseball and assumptions specific to the model of this project. We also assume that the batter is an average batter facing an average pitcher in the MLB. We also need to assume that no external factors, such as weather, audience or defensive play, contribute to the probability of success of the batter. For the purpose of simplification, we will assume that once a batter gets a "hit," he has successfully made it on base. Thus, we define a "hit" if he makes contact with the ball into the playable

fielding area or hits a home run. Alternatively, the batter could get on base if the pitcher throws four balls (4-X count) and "walks" the batter, also known as base on balls (BB) which we consider a hit in our model.

As stated in the MLB rule book, we will consider a "foul ball" a strike as long as the batter has fewer than two strikes. Fouls do not reset when the pitcher throws a ball [9]. Once the batter has two strikes, he can bat a foul ball up infinitely many times before his plate appearance (PA) ends. Because we know that a batter in the MLB has it a ball, gotten a strike, or been walked before, we assume that the batter will hit the ball, and his plate appearance will not continue forever. Therefore, we do not put a limit on the number of foul balls a batter can produce before ending his plate appearance. However, if the batter "foul tips" a ball with two strikes, then he is out.

2.3 Markov Chain Terms for Setup of Model

A Markov Chain allows us to model a situation that can be in one of several states and containing information about the probability of going from one state to another state. We refer to this transition as a step and are interested in how many steps it takes to get from one state to another. To get precise, we first need to define some terms. Let $X_1, \ldots, X_n$ be a collection of random variables that take on values or states in the set $S = \{s_1, s_2, \ldots, s_n\}$. In this paper, when it does not lead to confusion, we refer to the state by its subscript, i.e. $s = \{1, 2, ..., n\}$.

Markov property: The probability of going from one state i to another state j is independent of how we arrived at state i. This is represented by the equation

$$P\{X_{t_n} = x_n \big| X_{t_{n-1}} = x_{n-1}, ..., X_{t_0} = x_0\} = P\{X_{t_n} = x_n \big| X_{t_{n-1}} = x_{n-1}\}.$$

with n exhaustive and mutually exclusive states, and the probabilities at a specific point in time $t = 0, 1, 2, \ldots$ and $X_i \in S$. This is sometimes referred to as the memoryless property.

One-Step Transition Probability: The probability of going from state i to state j when the Markov property holds is called a transition probability from state i, s_i, to

state j, s_j and denoted by p_{ij}.

$$p_{ij} = P(X_{t_n} = j \big| X_{t_{n-1}} = i), (i,j) \in \{1,2,...,n\}, t = \{0,1,2...,T\}$$

$$p_{ij} \geq 0, (i,j) = 1,2,...,n.$$

Markov Chain: A collection of states together with the probabilities of going from one state i to another state j where the probabilities satisfy the Markov property and the sum of the probabilities of leaving a state is equal to 1.

If a Markov sequence of random states X_n take the discrete values $_1, ..., p_N$, then

$$P(x_n = p_{i_n} | x_{n-1} = p_{i_{n-1}}, ..., x_1 = p_{i_1}) = P(x_n = p_{i_n} | p_{n-1} = p_{i_{n-1}}).$$

[12]

We organize these transition probabilities p_{ij} in a matrix, P, where the sum of the entries in each row is 1.

$$\sum_j p_{ij} = 1, i = 1,2,...,n$$

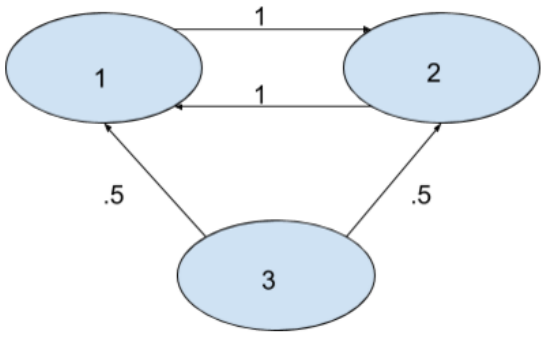
Row i contains all probabilities of starting in a state i and going to another state in column j

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} & \cdots & p_{1n} \\ p_{21} & p_{22} & p_{23} & \cdots & p_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ p_{n1} & p_{n2} & p_{n3} & \cdots & p_{nn} \end{bmatrix}$$

Example 1. *Referring to the state diagram in Figure 2 we see there are three states labeled 1, 2 and 3 in the circles. The transition probability are indicated by the numbers on the arrows. For example, the probability of going from state 3 to state 2 is* 0.5, *thus the transition probability* $p_{32} = 0.5$. *We organize the probabilities of going from state i to state j in the transition probability matrix in Figure 2.*

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & .0 \\ .5 & .5 & 0 \end{bmatrix}$$

Figure 2: State Diagram



We now consider the long-term behavior of a Markov chain when it starts in a state chosen by a probability distribution on the set of states, which we will call an initial probability vector.

Initial Probability: A probability vector with r components is a row vector whose entries are non-negative and sum to 1. If a is a probability vector which represents the initial state of a Markov chain, then we call it an initial probability vector and the i^{th} component of a represents the probability that the chain starts in state i.[5]

Next, we classify the state depending on how it behaves after a number of steps. The classifications we use include:

Absorbing State: A state j is absorbing if returns to itself with certainty in one transition $p_{jj} = 1$.

Transient State: state j can reach another state but cannot return to itself from any other state. As a result, we can expect that as time passes and n approaches infinity, the probability of being in a transient state approaches zero or $\lim_{n\to\infty} p_{ij}^n = 0, \forall i$.

Our model in the following section shows why the states for this project are only be classified as either absorbing or transient.

2.4 A Markov Chain Model for Baseball

Here we return to baseball by describing the game as a collection of states that relate to the question "Should a batter swing on a 2-0 count?" We then show why the states form

a Markov chain and how to determine the transition probabilities. To do so, we make the following additional assumptions:

1. the pitcher does not get tired, and
2. his pitching ability does not diminish over the course of several innings and no external factors contribute to the batter's ability or the pitcher's ability.

For our model, we have 14 states: one for each of the counts e.g. 2-0 or 3-2, etc. and one for each of the two ways batters can end their plate appearance (hit or out). We now classify the states used in this project, which are show in the a state diagram (Figure 3). A state diagram with circles for all of the states can be found in Appendix B.1 in Figure 9. Imagine the batter is in the middle of his plate appearance at a 2-0 count, then at each step, the batter has the option of getting a ball, strike or hit. If the pitcher throw another ball, the count becomes 3-0. Once the umpire calls four balls, the count will increase to 4-0 and the batter automatically goes to first base. Notice a transient state occurs when there is no way for the batter to return to the previous count in any amount of steps, but he can still continue his plate appearance. Then, we see that the batter enters an absorbing state when his plate appearance ends.

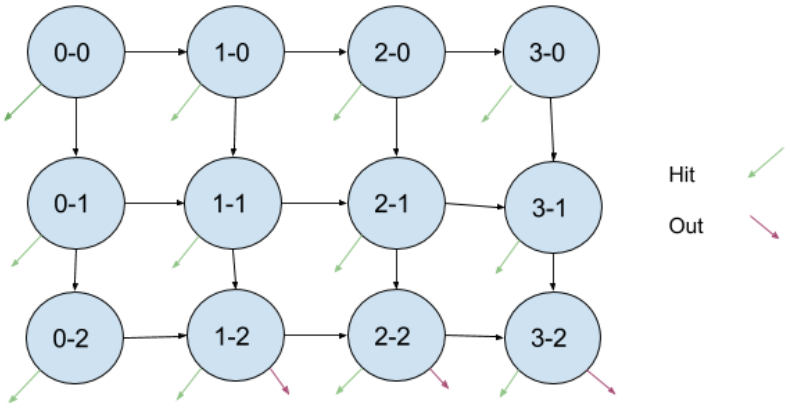


Figure 3: Modeling the Count. Recall that a batter can get a hit from any state. For ease in reading, we have included arrows leaving the states that don't seem to go anywhere. These arrows represent going from a state to a hit or out.

If the umpire calls a strike, the count will be 2-1 and this is also a transient state

since he cannot go back to no strikes. If the batter hits a foul ball with fewer than two strikes, we will count that as a strike, as described by our prior definitions and assumptions (Sections 2.1 and 2.2). Once the batter reaches three strikes, he is out. His plate appearance ends, so out is an absorbing state.

If the batter hits the ball, then we assume for our model that he was successful in reaching first base before being thrown out. This is also a success and so hit is classified as an absorbing state since his plate appearance ends and at each additional time step for the current plate appearance, he will still be considered on base. Each state is only dependent on the previous state. We are assuming that any time the pitcher attempts to throw out the runner on any base, it does not affect the pitch count. This is unique to our model and not the game. Incorporating getting thrown out is one of the weaknesses of our model, which we will see explanation for in Section 6.2.

3 Markov Chains Terms and Theorems for Analysis

Given an initial probability vector and a transition matrix, we can compute the probability of being in a given state after one step, e.g. $a^{(0)}P$, to know the probability of being in a given state after two steps. We consider $a^{(0)}P = a^{(1)}$

n-step Transition Probabilities: Given the initial probabilities $a^{(0)} = \{a_j^{(0)}\}$ of starting in state j and the transition matrix P of a Markov chain, the absolute probabilities $a^{(n)} = \{a_j^{(n)}\}$ of being in state j after n steps where $n > 0$ are given by

$$a^{(n)} = a^{(0)}P^n, n = 1, 2, ...$$

The absolute "$n^{th} - step$ probability" are given by multiplying the initial probability of being in state j by the n-step probability matrix. This is possible because the initial probability will stay the same through each time step, but the transition probabilities change because they are a function of the number of steps. This allows us to quickly calculate the probabilities of going from state-to-state at any given time.

Example 2. *Suppose we begin in state 3 in Example 1. Let the probability of beginning in state 3 equal 1 which is represented by the matrix* $a = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$. *Then, using our*

matrix P, also from Example 1, we can determine the probability of being in state 3 after n steps by computing

$$a^{(n)} = a^{(0)}P^n, n = 1, 2, ...$$

Specifically, if we want to know the probability of being in state 3 after two time steps, then $n = 2$ *and* $P = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ .5 & .5 & 0 \end{bmatrix}$.

$$\begin{aligned}
&= a^{(0)}P \\
&= a^{(1)}P \\
&= a^{(0)}PP \\
&= a^{(0)}P^2 \\
&= \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ .5 & .5 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ .5 & .5 & 0 \end{bmatrix} \\
&= \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ .25 & .25 & 0 \end{bmatrix} \\
&= \begin{bmatrix} .5 & .5 & 0 \end{bmatrix}
\end{aligned}$$

Focusing on the transition matrices we have the straightforward matrix multiplication calculations. The Chapman- Kolmogrov Equations hold for more general processes that don't satisfy the Markov process, but in the Markov chain situation they state the following:

Chapman-Kolmogorov Equations P^n is the $n-$step probability matrix where

$$P^n = P^{n-1}P$$

or more generally,

$$P^n = P^{n-m}P^m, 0 < m < n.$$

We use these equations to directly find the transition probability without needing the initial probability vector. We use this equation to quickly find the transition matrix P^n which is necessary for computing the n-step transition probabilities.

The following linear algebra concepts are needed to analyze the results of the transition matrix.

Matrix Inverse The inverse of a square matrix A, sometimes called a reciprocal matrix, is a matrix A^{-1} such that $AA^{-1} = I$ where I is the identity matrix. [11]

Canonical Form of a Transition Matrix Suppose a Markov chain with r states. The set of transient states, T, has size s and the absorbing, $\widetilde{T}$, has size $r - s$. Grouping the absorbing sets as the first $r - s$ steps in the transition matrix and following them by the s transient states puts the transition matrix in Canonical form. The Canonical form has four sub-matrices as shown in Figure 4. The $(r - s) \times (r - s)$ submatrix containing the probability of going from a transient state to a transient state, thus S is sometimes represented as the identity matrix I because we can never leave an absorbing state once it is entered. The zero matrix O is the $(r - s) \times s$ submatrix at the top right of the transition matrix. The entries in O are all zero because it is impossible to go from an absorbing state to leave an absorbing state. The submatrix R of size $s \times (r - s)$ on the bottom left of the transition matrix indicates submatrix containing from a transient state to an absorbing state. Finally, Q is the submatrix of size $s \times s$ organizing from a transient state to a transient state. To organize the matrix P in canonical form, we list all of the $r - s$ absorbing states first and then the s transient states.

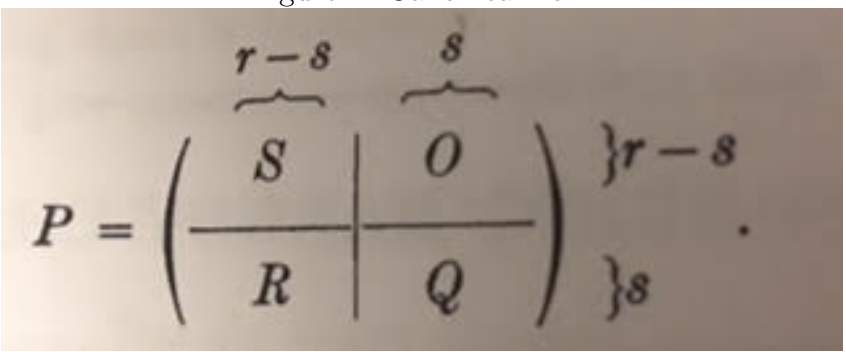
Theorem 1. *For any absorbing Markov chain $I - Q$ has an inverse and*

$$(I - Q)^{-1} = I + Q + Q^2 + Q^3 + ... = \sum_{k=0}^{\infty} Q^k.$$

Proof. Consider the identity

$$(I - Q)(I + Q + Q^1 + Q^2 + ... + Q^{n-1}) = I - Q^n,$$

Figure 4: Canonical Form



which we verify by multiplying the left side. Recall Q is the submatrix of transition probabilities transient states and will never equal I because once we leave a transient state we cannot return to it. Recall also that the identity matrix, I, has a determinant of 1. So, the determinant of $(I - Q)$ is not zero and is denoted, $det(I - Q) \neq 0$. By definition of a transient state, we know $Q^n \to 0$ because $p_{ij}^n \to 0$. Thus, $I - Q^n$ goes to I. For an n that is sufficiently large, $I - Q^n$ must have $det(I - Q^n) \neq 0$. More specifically, because $I - Q^n$ goes to I, $det(I - Q^n) = det(I) = 1$. The determinant of the product of two matrices is the product of the determinants

$$det((I-Q)(I+Q+Q^2+Q^{n-1})) = det(I-Q)det[(I-Q)(I+Q+Q^2+Q^{n-1})] = det(I-Q)^{-1}(I-Q^n).$$

and since Thus, $I - Q$ has an inverse. We can multiply both sides of the identity by the inverse: $I + Q + Q^2 + Q^{n-1} = (I - Q)^{-1}(I - Q^n)$. Then, since $(I - Q)^n$ tends to I as n tends to infinity, $(I - Q)^{-1}(I - Q^n)$ tends to $(I - Q)^{-1}$.

$$\begin{aligned} det((I - Q)(I + Q + Q^2 + Q^{n-1}) &= (I - Q)^{-1}(I - Q^n) \\ &= det(I - Q)det[(I - Q)(I + Q + Q^2 + Q^{n-1})] \\ &= (I - Q)^{-1}(I - Q^n). \end{aligned}$$

□

Fundamental matrix For an absorbing Markov chain, we define the fundamental matrix to be $N = (I - Q)^{-1}$.

Thus, every transition matrix for an absorbing Markov chain can be written in canonical form from which we can find the fundamental matrix from the submatrix Q.

In Figure 4, we see how we organize the transition probabilities for the various counts into the canonical matrix in Figure 10 in Appendix B.1. After doing so, we extract the Q submatrix of size 12×12 to find fundamental matrix N. From N, we can determine the probability of a batter starting in any of the counts in our baseball model, but for this project we assume that the batter always starts at 2-0 and the probability is 1. The next two theorems explain how we use matrix N to determine the probability of being in an absorbing state. Now that we have a matrix $N = \sum Q^k$ that contains the probability

Figure 5: Fundamental Matrix, N

(I-Q)^-1												
N Matrix	1.00	0.49	0.25	0.39	0.38	0.35	0.13	0.20	0.29	0.04	0.08	0.17
	0.00	1.00	0.51	0.00	0.40	0.50	0.00	0.14	0.33	0.00	0.04	0.16
	0.00	0.00	1.23	0.00	0.00	0.69	0.00	0.00	0.34	0.00	0.00	0.14
	0.00	0.00	0.00	1.00	0.48	0.27	0.34	0.33	0.33	0.11	0.15	0.24
	0.00	0.00	0.00	0.00	1.00	0.56	0.00	0.34	0.48	0.00	0.10	0.26
	0.00	0.00	0.00	0.00	0.00	1.27	0.00	0.00	0.62	0.00	0.00	0.26
	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.49	0.30	0.32	0.32	0.33
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.61	0.00	0.28	0.44
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.33	0.00	0.00	0.54
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.56	0.37
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.66
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.41

of moving from state i to state j in any number of steps, we develop the concepts needed to determine the probability of a batter getting on base, i.e. into a specific absorbing state after any number of steps.

Theorem 2. $QN = NQ = N - I$

Proof. Following from Theorem 3 and the definition of the Fundamental Matrix,

$$N = (I - Q)^{-1} = I + Q + Q^2 + Q^3 + ...$$

So

$$N - I = Q + Q^2 + Q^3 + ...$$

But also,

$$QN = Q(I + Q + Q^2 + Q^3 + ...) = Q + Q^2 + Q^3 + ...$$

and similarly

$$NQ = (I + Q + Q^2 + Q^3 + ...)Q = Q + Q^2 + Q^3 + ...$$

Thus, $QN = NQ = N - I$. □

From N, we can compute a new matrix B which organizes the probabilities b_{ij} of the process starting in a transient state and ending in an absorbing state. We are interested in the batters' probability of ending his plate appearance with a "hit" or an "out", which we determine with the matrix B.

B Matrix We define the new matrix $s \times (r - s)$ matrix B as
$B = NR$ where s there are transient states and $(r - s)$ there are absorbing states.

Theorem 3. *If b_{ij} is the probability that the process starting in a transient state s_i ends up in an absorbing state s_j, then*

$$\{b_{ij}\} = B = NR, \quad s_i \in T, \quad s_j \in \tilde{T}.$$

Proof. A process starting in transient state s_i may be absorbed into s_j in one or more steps. The probability of being absorbed in a single step is p_{ij}. If this does not happen, the process can move either to an absorbing state (in which case it is impossible to reach s_j), or to another transient state s_k. In the latter case there is a probability b_{kj} of being captured in the desired state. Hence, we have

$$b_{ij} = p_{ij} + \sum_{s_k \in T} p_{ik} b_{kj},$$

which can be written in matrix form as

$$B = R + QB.$$

Thus,

$$B = (I - Q)^{-1}R = NR.$$

□

Theorem 4. *If B^* is the $r \times r$ transition matrix whose entry b^*_{ij} gives the probability of being absorbed in s_j, starting in s_i, for all states s_i and s_j, then*

$$PB^* = B^*.$$

Proof. If $s_j \in T$, then $b^*_{ij} = 0$. Thus, the last s column of B^* are 0. Consider s_j absorbing. If $s_i \in T$, then $b^*_{ij} = b_{ij}$. If s_i is also absorbing, then $b^*_{ij} = d_{ij}$. So, we have

$$B^* = \begin{bmatrix} I & 0 \\ R & Q \end{bmatrix}$$

$$PB^* = \begin{bmatrix} I & 0 \\ R & Q \end{bmatrix} \begin{bmatrix} I & 0 \\ B & 0 \end{bmatrix} = \begin{bmatrix} I & 0 \\ R + QB & 0 \end{bmatrix}.$$

Recall that by Theorem 2, $QN = (N - I)$. So $R + QB = R + QNR = R + (N - I)R = NR = B$. $\Rightarrow\Leftarrow$

Therefore,

$$PB^* = B^*.$$

□

We use this theorem to determine the probability of a batter getting a hit if he starts at a 2-0 count. This is found in our B submatrix in Figure 6.

4 Generating the Data

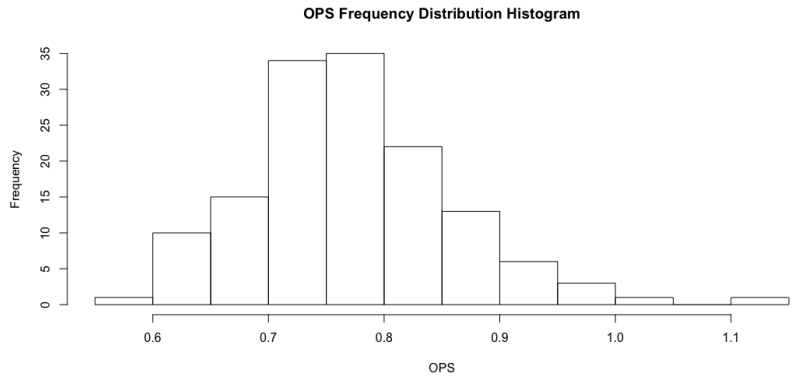
We need to understand the following statistical concepts to develop the transition probabilities and then analyze the results.

Figure 6: B Submatrix: This should be before this section

B-Submatrix		
	H	O
0-0	0.79	0.21
0-1	0.72	0.28
0-2	0.56	0.44
1-0	0.83	0.17
1-1	0.75	0.25
1-2	0.59	0.41
2-0	0.87	0.12
2-1	0.80	0.20
2-2	0.65	0.35
3-0	0.93	0.07
3-1	0.88	0.12
3-2	0.75	0.25

To get an idea of the On-Base-Percentage Plus Slugging (OPS) distribution of the players, we constructed a histogram and computed a five number summary which reports values including quartile ranges and the median. We also computed the OPS average for the 2015 batters. The histogram below shows that the OPS are normally distributed with a few outliers, which is important for conducting the hypotheses tests we use in Section 5 and Appendix C. In the statistical package R, we defined hits as successes and outs

Figure 7: OPS Frequency Histogram



as failures. The algorithm that we programmed in R tidied the play-by-play Retrosheet data necessary for the analysis of the count required this categorization because we had

various cases of successes and failures that could appear to be in multiple states. For example, suppose a batter is at a 3-0 count for his current plate appearance. If the pitcher throws a ball, the play gets recorded as a "ball," but since it is the fourth one, the batter advances to first base for a base on balls (BB). In order to avoid plays being counted twice, we tidied the data in R to accurately classify each play according to its state.

We were interested in how the batters go up the count, in terms of balls, strikes and foul balls, because these values are used to track how many times we go from a state i to a state j. These values help us calculate the transition probabilities. To compute the transition probabilities, we kept track of how the batters' plate appearances ended: hit, out, base on balls and foul tip. Then, to find our transition probabilities we computed W_{ij} the number of times the batters moved from state i to a specific state j, as well as $\sum_{k=1}^{r} W_{ik}$ the number of times the batter transitioned from state i to any state j. The transition probability p_{ij} for moving from a state i to a state j is

$$p_{ij} = \frac{W_{ij}}{\sum_{k=1}^{r} W_{ik}}.$$

We computed these transition probabilities after running our program included in Appendix A and organized them in our transition matrix using the markovchain package in R found in Appendix A. This equation for p_{ij} follows from the Markov property. Since our model is a Markov chain, we assume that the transition probability only depends on the prior state and not on previous states. The transition probability p_{ij} is a conditional probability where we are given that the prior state was i, thus the sample space for this finite probability is any play that starts in state i. Thus,

$$\begin{aligned} p_{ij} &= P(X_n = j | X_{n-1}) \\ &= \frac{\#(i \to j)}{\#(i \to \forall j)} \end{aligned}$$

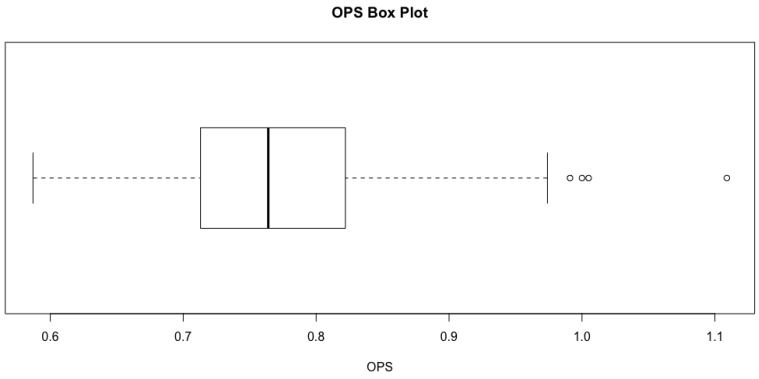
We might think to sum the transition probabilities of the batters and then divide by the number of batters, but each batter can have his own unique transition probabilities and Markov chain. To do this, let M be the set of batters and $|M|$ be the number of batters. Let m represent a batter in $|M|$. We use $W_{ij}(m)$ to be the number of times batter m moves from state i to state j and $p_{ij}(m)$ ot be the transition probabilities for batter m.

Then, the average of all the batter specific transition probabilities is represented by

$$\begin{aligned}\tilde{p_{ij}} &= \frac{\sum_{m\in|M|} p_{ij}(m)}{|M|} \\ &= \frac{\sum_{m\in|M|} \frac{W_{ij}(m)}{\sum_{k=1}^{r} W_{ik}(m)}}{|M|} \\ &= \sum_{m\in|M|} \frac{W_{ij}(m)}{|M| \sum_{k=1}^{r} W_{ik}(m)}.\end{aligned}$$

We compute the transition probabilities for the 2015 MLB batters and compare their probability of success to the population probability of success. Since we are testing the advice not to swing on a 2-0 count, we assume that advice is as good as a guess and the probability of success, $p = 0.5$. Using our five number summary in R, which is represented by the box plot below, we compute our population mean. Then, we randomly select a sample of batters to compute transition probabilities for. That is, we compute the transition probabilities for a random sample of batters, and compare their probability of success to the population mean probability of success, or the mean probability that a batter reaches base.

Figure 8: OPS Box Plot and Five Number Summary



Recall that analyzing the transition probabilities we determine the probability of a batter reaching base. The transition probability matrix can be found in Section B. We use a Markov chain because there are several routes for the batter to take before reaching an absorbing state. This is indicated in the Figure 3 as we can see how batters can start from a 0-0 count and move to 0-1 or 1-0, then from either of those states there are two

more states the batter can reach. This process continues until we reach an absorbing state. Each transition is only dependent on the state immediately prior, thus we use a Markov chain to model the outcomes.

5 Results and Analysis

Now that we have run our program to compute transition probabilities to determine our probabilities of success, we can analyze our findings using statistical tests. First, we need to understand some concepts related to statistics.

First, we choose a random sample of batters for which compute the mean probability of success.

Random Sample Random sampling is a sampling technique where we select a group of subjects (a sample) for study from a larger group (a population). Each individual is chosen entirely by chance and each member of the population has a known, but possibly non-equal, chance of being included in the sample. [13]

Theorem 5 (Bernoulli's Law of Large Numbers)**.** *As the number of identically distributed, randomly generated variables increases, their sample mean approaches their theoretical mean.*

Following Bernoulli's Law of Large Numbers in Theorem 5, we expect the sample proportion $\hat{p}$ to approach the population proportion.

When estimating a population proportion, p, we must have the following basic conditions to determine whether the distribution of the sampling proportion $\hat{p}$ is close to the Normal Distribution N. The conditions state:

1. *Random and Independent* The sample is a random sample from the population, and observations are independent of each other. The sample can be collected with or without replacement.

2. *Large Sample* The sample size, n is large enough that the sample expects at least ten successes and ten failures.

3. *Big Population* If the sample is collected without replacement, then the population size must be at least ten times larger than the sample size.

Then, we can use the Central Limit Theorem to determine if the data is normally distributed.

Theorem 6 (Central Limit Theorem for Sample Proportions (CLT))**.** *If we take a random sample n from a population, and if the sample size is a large and the population size much larger than the sample size, then the sampling distribution for $\hat{p}$ is approximately*

$$N\left(p, \sqrt{\frac{p(1-p)}{n}}\right).$$

If the value of p, the population proportion, is unknown we can substitute the value of $\hat{p}$ to calculate the estimated standard error.

We use the Central Limit Theorem in our hypothesis test.

Hypothesis Test For hypothesis testing, we refer to our null hypothesis H_0 as the hypothesis we want to test. Then, we have our alternative hypothesis H_A which provides some alternative situations. We provide the conditions necessary for hypothesis testing in Appendix C

In terms of this project, for our one sample tests our null hypothesis H_0 states a batter should swing on a 2-0 count, and our alternative hypothesis H_A states that a batter should not swing on a 2-0 count.

Using random sampling, we selected 141 batters, the number of batters in the 2015 season, from the MLB population of batters. We ran the R program to determine their transition probabilities and average probability of getting a hit from 2-0, which can be found in our B Matrix in Figure 6. From our B submatrix, we determine that the probability of being absorbed into a hit from a 2-0 count is .87, and we use this as our $\hat{p}$.Following from Bernoulli's Law of Large numbers (Theorem5), we hypothesized that the mean probability of the random sample would approach the population mean. Using 95% confidence. We hypothesized (our H_0) stated that the average batter should not swing on a 2-0 count, and determined that the average batter in the MLB should not

swing on a 2-0. There is not sufficient evidence to reject H_0. This hypothesis test can be found under Hypothesis Tests tests in Appendix C.

6 Conclusion/Future Work

6.1 Conclusion

After writing a program to analyze the 2015 Major League Baseball pitch sequences, we computed the necessary transition and fundmental matrices. Using these matrices, we determined that the probability of eventually getting on base from a 2-0 count is 87 percent. From our hypothesis test where we used this probability, we determined that there is mathematical support for the advice not to swing on a 2-0 count.

6.2 Future Work

- Someone could compare pitchers in the MLB and try to determine how the average batter would fair against the average pitcher.
- Someone could compare lefties and righties and run this program.
- Someone could reprogram the model to include recurring states when pitcher tries to throw the runner out
- Run all Retrosheet data and see how different years compare to population mean.
- Compare the variances for batters in different quartiles.

A R Code

```
#make sure the markovchain package is installed. if not execute next commen
#install.packages("markovchain")

pbp15 <- read.csv("all2015.csv")
```

```
#getting headers
headers <- read.csv("fields.csv")
names(pbp15) <-headers$Header
View(pbp15)
# We're only dealing with one variable, so let's grab it.
Z <- pbp15$`Pitch sequence`
#remove all non-pitches
Z <- gsub("[.>123+*N]","",Z)
Z <- gsub("[BIPVY]", "b",Z)
#[XFR]
#pbp15$`Pitch sequence`<-gsub("[b]$", "H", pbp15$`Pitch sequence`)
#coding strikes
Z <- gsub("[CKLMOQST]", "s",Z)
#coding last strike as out, i want to code the fouls as strikes as long as t
#pbp15$`Pitch sequence`<-gsub("[s]$", "O", pbp15$`Pitch sequence`)
#0 is out 1 is hit

dealwithfouls <- function(z) {
  zz <- strsplit(z,"")[[1]]
  scount <- 0
  pmark <- 1
  while (scount < 2 && pmark < length(zz)){
    if (zz[pmark] == "F"){
      zz[pmark] <- "s"
    }
    if (zz[pmark] == "s"){
      scount <- scount + 1
    }
    pmark <- pmark + 1
  }
```

```
  return(paste(zz, collapse = ""))
}

Z <- sapply(Z, dealwithfouls)

one.string <- function(ex){
  # replace s and b and X with O for strikeouts
  # and H for hits
  ex <- gsub("s$", "O", ex)
  ex <- gsub("b$", "H", ex)
  ex <- gsub("X$", "H", ex)
  # create a vector of individual outcomes
  ex.v <- unlist(strsplit(ex,""))
  # remove last X from vector
  last_state <- ex.v[length(ex.v)]
  ex.v <- ex.v[-length(ex.v)]
  # compute cumulative total of balls and strikes
  n.balls <- as.vector(cumsum(ex.v == "b"))
  n.strikes <- as.vector(pmin(cumsum(ex.v == "s"), 2))
  # create pitch count variable
  S <- paste(n.balls, n.strikes, sep="-")
  # add a beginning and end outcome.
  #If the count goes over 3-X or X-2, then it should print either an "H" or
  S <- c("0-0", S, last_state)
  # before and after counts
  #b.count <- unname(S[1:(length(S)-1)])
  #e.count <- unname(S[2:length(S)])
  #return(list(b.count, e.count))
  return(S)
}
```

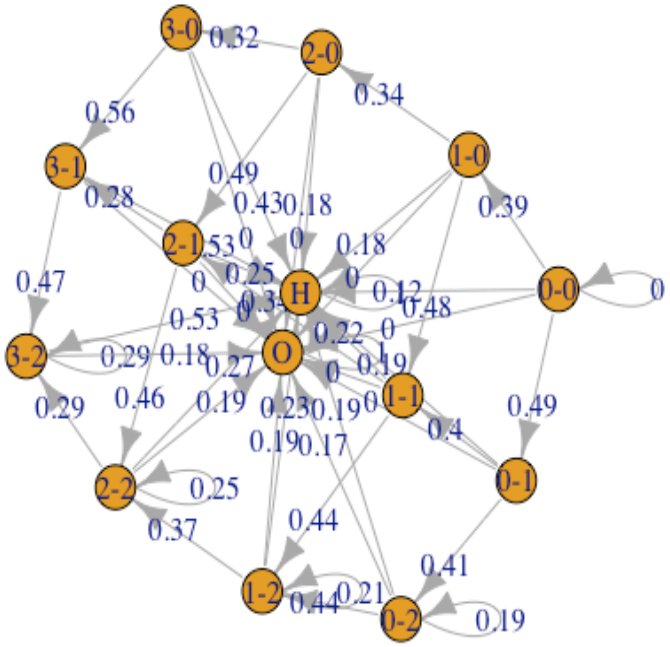
```
S <- sapply(Z, one.string)
S <- unname(unlist(S))
a <- S[1:(length(S)-1)]
b <- S[2:length(S)]

TR <- table(a,b)
P <- prop.table(TR[1:(nrow(TR)-5),c(1:12,16,17)], 1)
P <- rbind(P, c(rep(0, ncol(P)-2),1,0),c(rep(0,ncol(P)-1),1))
dimnames(P)[[1]][nrow(P)-1] <- "H"
dimnames(P)[[1]][nrow(P)] <- "O"
# convert this table P into a matrix,
# specifically the transition matrix,
# that I can use in the markov chain
# package without having to manually input the data in our table P
P <- as.matrix(P)
#check your work
View(P)
require(markovchain)
mc <- new("markovchain", transitionMatrix = P, name = "The Count")
summary(mc)
plot(mc,package='diagram')
#change to Canonical Form
mcCanonic <- canonicForm(mc)
C<-as(mcCanonic,"matrix")
#View the matrix
View (C)
#write a csv file to export for analyzing and computations in Excel
write.csv(C, file = "CanForm.csv")
T<-as(mc,"matrix")
```

```
write.csv(T, file = "CanForm.csv")
```

B Data and Transition Matrices

B.1 Canonical Form

Figure 9: State Diagram in R



C Hypothesis Test

Hypothesis Test $\overline{x}, \mu, z - score$

Let $p_0 = 0.5$ because coaches guess that batters should not swing on a 2-0 count. We computed the probability of getting a hit sometime after starting in a 2-0 count , $\hat{p} = 0.87$. To find our t-score, we must compute the sample standard deviation using the equation $s_x = \sqrt{\frac{p_0(1-p_0)}{n-1}}$, and find that $s_x = 8.757$.

Figure 10: Canonical Form of the Transition Matrix for the Count

CanForm														
	H	O	0-0	0-1	0-2	1-0	1-1	1-2	2-0	2-1	2-2	3-0	3-1	3-2
H	1	0	0	0	0	0	0	0	0	0	0	0	0	0
O	0	1	0	0	0	0	0	0	0	0	0	0	0	0
0-0	0.12	0.00	0	0.49	0	0.39	0	0	0	0	0	0	0	0
0-1	0.19	0.00	0	0	0.41	0	0.4	0	0	0	0	0	0	0
0-2	0.19	0.17	0	0	0.19	0	0	0.44	0	0	0	0	0	0
1-0	0.17	0.00	0	0	0	0	0.48	0	0.34	0	0	0	0	0
1-1	0.22	0.00	0	0	0	0	0	0.44	0	0.34	0	0	0	0
1-2	0.23	0.19	0	0	0	0	0	0.21	0	0	0.37	0	0	0
2-0	0.18	0.00	0	0	0	0	0	0	0	0.49	0	0.32	0	0
2-1	0.25	0.00	0	0	0	0	0	0	0	0	0.46	0	0.28	0
2-2	0.27	0.19	0	0	0	0	0	0	0	0	0.25	0	0	0.29
3-0	0.43	0.00	0	0	0	0	0	0	0	0	0	0	0.56	0
3-1	0.53	0.00	0	0	0	0	0	0	0	0	0	0	0	0.47
3-2	0.53	0.18	0	0	0	0	0	0	0	0	0	0	0	0.29

Using a right-tailed test and an $\alpha = 0.05$, We compute our z-score using the equation $z = \frac{\hat{p}-p_0}{s_x}$, and find $z = 0.04$, and the p-value associated with it is 0.4840. Since $p > \alpha$, we fail to reject the null. Thus, we conclude that a batter should not swing on a 2-0 count.

References

[1] Albert, Jim. Sequences of PItch Counts. Exploring Baseball with R. Retrieved October 13, 2016, from $https://baseballwithr.wordpress.com/2015/07/20/sequences-of-pitch-counts/$.

[2] Bukiet, B., Harold, E. R., & Palacios, J. L.. (1997). Markov Chain Approach to Baseball. *Operations Research*, 45(1), 1423. Retrieved from http://www.jstor.org/stable/171922

Figure 11: R Submatrix

R-Submatrix		
	H	O
0-0	0.12	0.00
0-1	0.19	0.00
0-2	0.19	0.17
1-0	0.17	0.00
1-1	0.22	0.00
1-2	0.23	0.19
2-0	0.18	0.00
2-1	0.25	0.00
2-2	0.27	0.19
3-0	0.43	0.00
3-1	0.53	0.00
3-2	0.53	0.18

[3] Macdonald, Larry. "Imagine Sports Baseball." Imagine Sports Baseball. N.p., n.d. Web. 11 Sept. 2016.

[4] Marchi, Max, and Jim Albert. Analyzing Baseball Data with R. N.p.: n.p., n.d. Print.

[5] Markov Chains. (n.d.). Retrieved October 13, 2016, from *https* : *//www.dartmouth.edu/ chance/teaching_aids/books_articles/probability_book/Chapter*11.*pdf*

[6] Random Variables. (n.d.). Retrieved May 03, 2016, from *http* : *//www.stat.yale.edu/Courses/*1997 − 98/101/*ranvar.htm*

[7] Katz, Stanley M. Study of 'The Count' Yields Fascinating Data. (n.d.). Retrieved May 03, 2016, from *http* : *//research.sabr.org/journals/study* − *of* − *the* − *count* − *yields* − *fascinating* − *data*

[8] Glossary. (n.d.). Retrieved September 19, 2016, from http://m.mlb.com/glossary

[9] Glossary/Rules. (n.d.). Retrieved September 19, 2016, from http://m.mlb.com/glossary/rules

[10] Sheskin, Theodore J. "Computing the Fundamental Matrix for the Reducible Markov Chain." Mathematics Magazine Dec. 1995: 393-98. Mathematics Association of America. Web. 11 Sept. 2016.

[11] Wolfram Alpha. (n.d.). Retrieved November 28, 2016 from $http : //mathworld.wolfram.com/MatrixInverse.html.$

[12] Wolfram Alpha. (n.d.). Retrieved November 28, 2016 from $http : //mathworld.wolfram.com/MatrixInverse.html.http : //mathworld.wolfram.com/MarkovChain.html$

[13] $http : //www.stats.gla.ac.uk/steps/glossary/sampling.html\#srs$