*Journal of Quantitative Analysis in Sports*

*Volume* 6, *Issue* 4 2010 *Article* 6

# A Markov Process Model of the Number of Years Spent in Major League Baseball

**Anthony C. Krautmann,** *DePaul University*
**James E. Ciecka,** *DePaul University*
**Gary R. Skoog,** *DePaul University*

**Recommended Citation:**
Anthony C. Krautmann, James E. Ciecka, and Gary R. Skoog (2010) "A Markov Process Model of the Number of Years Spent in Major League Baseball," *Journal of Quantitative Analysis in Sports*: Vol. 6 : Iss. 4, Article 6.
Available at: http://www.bepress.com/jqas/vol6/iss4/6
**DOI:** 10.2202/1559-0410.1263

# A Markov Process Model of the Number of Years Spent in Major League Baseball

Anthony C. Krautmann, James E. Ciecka, and Gary R. Skoog

## Abstract

We treat the number of years spent in major league baseball as a random variable and estimate probability distributions for this random variable through the use of recursive formulae. Distributional characteristics, including major league baseball worklife expectancies, are estimated for players by age and current activity status in the major leagues. Data from a recent time period (1977-2007) are used to calculate current characteristics of time spent in major league baseball. However, the contemporaneous nature of our data leads to censoring because many players in our data set had not completed their major league careers by the end of 2007. We deal with censoring through a Markov process model that captures transitions between activity and inactivity in major league baseball.

**KEYWORDS:** Markov, model, career, MLB

**Author Notes:** The authors thank Alex Krautmann for data compilation.

## I. Introduction

We estimate probability distributions of the number of years players spend in major league baseball. Given age and whether a player is currently active or inactive in the major leagues, we estimate the expected number of future years in major league baseball. Other characteristics, such as the probability that a currently inactive player will never return to major league play, also are estimated. We calculate average remaining career length aggregated over all players currently in the major leagues. Finally, we estimate the number of years a new player can expect for the length of his major league career.

We view the amount of time (measured in years) that a player spends in major league baseball as a random variable. The probability distribution of this random variable depends on transition probabilities between two living states (active in major league baseball or inactive) and an absorbing (death) state. A Markov-process model (*i.e*., an auto-regressive model of order one) allows us to account for transitions between living states and between living states and the death state. We use recursive formulae which, when combined with estimated transition probabilities, enable us to estimate the probability distribution of years spent in major league baseball.

Because our data set covers the period 1997-2007, we have estimates of career length characteristics that are very current. The methodological approach used in previous work (*e.g*., Witnauer, Rogers, and Saint Onge, 2007) tabulated the lengths of careers after players stopped playing in the major leagues. That type of approach has the benefit of hind sight, but it precludes using data on currently active players and incorporating any factors that may influence careers of current players. The approach used in this paper is more contemporaneous, but censoring problems could occur because many players in our data set had not completed their major league careers by the end of 2007. We deal with censoring through a Markov process (increment-decrement) model which accounts for players exiting and reentering major league baseball any number of times based on estimated transition probabilities.

The paper is organized as follows. Section II contains notation, definitions, and recursive formulae. Empirical results for batters (*i.e*., non-pitchers) are in Section III. Section IV compares the approach and results presented in the Sections II and III with an important recent contribution to the literature on career lengths in major league baseball. Section V is the conclusion.

## II. Notation and Recursions for Markov (Increment-Decrement) Model

We use the following notation. A player's age is denoted by $x$ with $BA$ (*Beginning Age*) being the youngest age at which major league baseball activity can occur.

Everyone must have either ended his playing career in the major leagues or died by *Truncation Age TA*. Let $m \in \{a,i\}$, and $n \in \{a,i,d\}$ where *a* denotes active in major league baseball, *i* denotes inactive in major league baseball, and *d* the death state.[1] The transition probability ${}^{m}p_x^n$ denotes the probability that a player in state *m* at age *x* will be in state *n* at age *x* + 1. Any player alive, whether in state *a* or *i*, at age *TA* − 1 has transitioned to state *i* or state *d* at age *TA*. We assume that mortality probability and activity status are independent, *i.e.*, ${}^{a}p_x^d = {}^{i}p_x^d = {}^{\cdot}p_x^d$. Transition probabilities incorporate mortality,[2] and the only restrictions on them are that ${}^{m}p_x^n \geq 0$, ${}^{a}p_x^a + {}^{a}p_x^i + {}^{\cdot}p_x^d = 1$, and ${}^{i}p_x^a + {}^{i}p_x^i + {}^{\cdot}p_x^d = 1$.

Let $YA_{x,m}$ (*Years of Baseball Activity*) denote the random variable of future major league active time for a player who is in state *m* at age *x*, and $p_{YA}(x,m,y)$ measures the probability that a player who is in state *m* at age *x* will accumulate *y* additional years of major league activity. That is, $p_{YA}(x,m,y)$ measures the probability that $YA_{x,m} = y$; and the mapping of each *y* into its probability of occurring, $p_{YA}(x,m,y)$, constitutes the probability mass function (pmf) for $YA_{x,m}$. We refer to $YA_{x,m}$ as the additional-years-in-baseball random variable. Its pmf ultimately depends on the transition probabilities ${}^{m}p_x^n$ which are primitive to the Markov model. Transitions between states are assumed to occur at the end of the periods between ages. We let ${}^{m}p_x^n$ denote the probability that a player in state *m* at age *x* stays in that state until age *x*+1, at which point in time he enters state *n*. In effect, this assumption means that a player is credited with an entire year of activity if he is in the major leagues for any part of a year.[3]

---

[1] A player is defined to be active at age *x* if he played in at least one major league game in the season he was age *x*. Age is defined to be a player's age on the first day of July. We assume *BA* = 19 and *TA* = 46.

[2] The probability of death of an active major league baseball player is small. However, active major league players have been known to die. For example, Darryl Kile died in 2002, Cory Lidle died in 2006, and Nick Adenhart died in 2009. Although the results we report would not change much if we were to assume that the probability of death were zero, we include mortality because our methodology seamlessly accommodates it and thereby obviates an assumption of zero mortality.

[3] Other transition points are possible. For example, we might think of transitions between states occurring at mid year. The pmf based on mid-year transitions is one-half year to the left of its end-of-year transitions counterpart. Therefore, the mean and all percentile points decrease one-half year as we move from end to mid-year transitions; but the variance, standard deviation, skewness, and kurtosis remain unchanged. When starting inactive, pmfs for activity do not depend on the time of transitions; therefore all measures of activity are the same under mid-year and end-of-year timing assumptions.


DOI: 10.2202/1559-0410.1263

Recursions defining $p_{YA}(x,m,y)$ consist of global conditions, boundary conditions, and main recursions. Global conditions (1a)–(1d) refer to extreme values of *y* and *x* as well as conditions that hold at all ages. Boundary conditions (2a)–(2d) deal with probabilities of future activity being either 0 or 1 year. Main recursions (3a)–(3b) capture probabilities of future activity for years *y* exceeding values defined by the boundary conditions (Skoog and Ciecka 2002).

*Years of Activity Probability Mass Functions for* $YA_{x,m} = y$ *for* $m \in \{a,i\}$ *with End-of-Period Transitions*

Global Conditions

(1a) $p_{YA}(x,a,y) = p_{YA}(x,i,y) = 0$ if $y < 0$ or $y > TA - x.$

(1b) $p_{YA}(TA,a,0) = p_{YA}(TA,i,0) = 1$

(1c) ${}^{a}p_x^d = {}^{i}p_x^d = {}^{\cdot}p_x^d$ for $x = BA,\ldots,TA-1$

(1d) ${}^{a}p_{TA-1}^i + {}^{a}p_{TA-1}^d = {}^{i}p_{TA-1}^i + {}^{i}p_{TA-1}^d = 1$

Boundary Conditions

(2a) $p_{YA}(x,a,0) = 0$ for $x = BA,\ldots,TA-1$

(2b) $p_{YA}(x,a,1) = {}^{a}p_x^d + {}^{a}p_x^i p_{YA}(x+1,i,0)$ for $x = BA,\ldots,TA-1$

(2c) $p_{YA}(TA-1,i,0) = 1$

(2d) $p_{YA}(x,i,0) = {}^{i}p_x^d + {}^{i}p_x^i p_{YA}(x+1,i,0)$ for $x = BA,\ldots,TA-1$

Main Recursions for $x = BA,\ldots,TA-1$

(3a) $p_{YA}(x,a,y) = {}^{a}p_x^a p_{YA}(x+1,a,y-1) + {}^{a}p_x^i p_{YA}(x+1,i,y-1),\ y = 2,3,\ldots,TA-x$

(3b) $p_{YA}(x,i,y) = {}^{i}p_x^a p_{YA}(x+1,a,y) + {}^{i}p_x^i p_{YA}(x+1,i,y),\ y = 1,2,\ldots,TA-x$

Global condition (1a) says that future years of activity cannot be negative nor can they exceed *TA –x* years. The latter condition holds because we assume that everyone alive at age *x* dies or leaves baseball before or at age *TA*. Again, assuming that everyone has died or left baseball by age *TA*, the probability of zero active time at age *TA* is 1 as expressed in (1b). Global condition (1c) expresses the assumption that transition to the death state is independent of whether a player is active or inactive in the major leagues. The last global condition, (1d) says that every player alive at age *TA* – 1 must transition to inactivity or the death state.

Boundary condition (2a) expresses the impossibility of no future time in major league baseball if a player were active at age *x* since one year of activity is credited to a currently active player. Condition (2b) gives the probability that a

currently active player accrues exactly one year of activity by having died by age $x$+1 or by transitioning to inactive and remaining inactive thereafter. In boundary condition (2c), anyone inactive at age $TA - 1$ accumulates zero years of activity with certainty; and (2d) gives the probability an inactive player can accumulate no additional labor force time by transitioning to $d$ or remaining in the $i$ state thereafter.

The remaining probability mass values are defined by the main recursions. The right-hand side of (3a) is the sum of two terms that contribute to the probability that an active player age $x$ will accumulate $y$ years of future activity. (1) The first term ${}^{a}p_x^a p_{YA}(x+1,a,y-1)$ is the product of two factors. The second factor, $p_{YA}(x+1,a,y-1)$, is the probability that a player active at age $x$ + 1 will have $y$ – 1 years of future activity and, when multiplied by ${}^{a}p_x^a$ adds another year of activity at age $x$. (2) The second term ${}^{a}p_x^i p_{YA}(x+1,i,y-1)$ also is the product of two factors. The latter factor $p_{YA}(x+1,i,y-1)$ is the probability that a player inactive at age $x$ + 1 will have $y$ – 1 years of future activity and, when multiplied by ${}^{a}p_x^i$ yields an additional one year of activity at age $x$. The second factors in both terms aggregate sample paths resulting from remaining active for $y$ – 1 years from age $x$+1, and their respective multipliers ${}^{a}p_x^a$ and ${}^{a}p_x^i$ induce an additional one whole year of activity.[4] The right-hand side of (3b) is the sum of two terms that contribute to the probability that an inactive player age $x$ will accumulate $y$ years of activity: (1) The first term ${}^{i}p_x^a p_{YA}(x+1,a,y)$ is the product of two factors. $p_{YA}(x+1,a,y)$ is the probability that a player active at age $x$ + 1 will have $y$ years of future activity and, when multiplied by ${}^{i}p_x^a$ yields no additional years of activity at age $x$ since the transition to activity comes at the end of the year. (2) The second term ${}^{i}p_x^i p_{YA}(x+1,i,y)$ also is the product of two factors. The latter factor $p_{YA}(x+1,i,y)$ is the probability that a person inactive at age $x$ + 1 will have $y$ years of future activity and, when multiplied by ${}^{i}p_x^i$, adds no additional activity at age $x$.

Our goal is to estimate the probability distribution of the years-of-activity random variable $YA_{x,m}$ within the context of the Markov model using the

[4] It clear that ${}^{a}p_x^a$ leads to another year of activity since there is a transition from active to active. However, ${}^{a}p_x^i$ also ensures one more year of activity because the transition to inactive comes at the end of the year with our assumption of end-of-year transitions.


DOI: 10.2202/1559-0410.1263

foregoing recursions and then compute $YA_{x,m}$'s characteristics with the following formulae. The expected value of $YA_{x,m}$, major league baseball worklife expectancy or average career length for a player in state *m* (active or inactive) at age *x*, is defined by

(4a) $$E(YA_{x,m}) = \sum_{y} y_x p_{YA}\left(x,m,y\right) = {}^{m}e_x^a .$$

The median value $y_{x,med}$ of $YA_{x,m}$ possess the property that

(4b) $$\Pr(YA_{x,m} \le y_{x,med}) \ge .50 \text{ and } \Pr(YA_{x,m} \ge y_{x,med}) \ge .50 ,$$

and the mode $y_{x,mode}$ of $YA_{x,m}$ is the value of $YA_{x,m}$ that fulfills the inequality

(4c) $$p_{YA}\left(x,m,y_{x,mode}\right) \ge p_{YA}\left(x,m,y\right) \text{ for all values of } y.$$

The variance, standard deviation, skewness, and kurtosis are defined by

(4d) $$V(YA_{x,m}) = \sum_{y} (y_x - {}^{m}e_x^a)^2 p_{YA}(x,m,y) ,$$

(4e) $$SD(YA_{x,m}) = \sqrt{V(YA_{x,m})} ,$$

(4f) $$SK(YA_{x,m}) = (1/SD(YA_{x,m}))^3 \sum_{y} (y_x - {}^{m}e_x^a)^3 p_{YA}(x,m,y) , \text{ and}$$

(4g) $$KU(YA_{x,m}) = (1/SD(YA_{x,m}))^4 \sum_{y} (y_x - {}^{m}e_x^a)^4 p_{YA}(x,m,y) .$$

Cumulative probabilities occur at values of $YA_{x,m}$ where

(4h) $$\Pr(YA_{x,m} \le y_{x,\alpha}) \ge \alpha \text{ and } \Pr(YA_{x,m} \ge y_{x,\alpha}) \ge \alpha \text{ for } \alpha = .10,.25,.75,.90 .$$

## III. Empirical Results

Comprehensive year-by-year records exist for all players for the entire history of major league baseball since 1871. We consider the eleven-year period 1997 –

2007, during which time 1,536 batters played major league baseball.[5] Our data set contains 16,896 potential observations (= 1,536 batters x 11 seasons). However, some players entered major league baseball after 1997 and therefore had no record between 1997 and the date they entered the major leagues. We deleted players after they were inactive for four consecutive seasons in order to limit their impact on the estimate of ${}^{i}p_x^{i}$ after their careers had effectively ended.[6] This left 8,555 observations for batters. A player is defined to be an active major league player in a particular year if he appeared in at least one major league game during that year. [7] A player is inactive in a particular year if he did not play in any major league games during the year but only after commencing his major league career. The end-of-period transitions assumption seems most appropriate given our definition that an active player receives credit for a year of activity if he appears in one major league game during a year (*i.e.*, once active, the player cannot become inactive until the following season with our counting convention).

*PMFs and Their Characteristics for Major League Baseball Batters*

Table 1 shows the age distribution of major league batters, who range in ages from 19 to 48, with a mean age of 28.62 and a standard deviation of 4.31 years.[8] Table 1 and Figure 1 are the pmf for batters by age; the mass function is skewed to the right with the mean exceeding the median, 28.00, which exceeds the mode, 26.00. Table 2 contains estimated transition probabilities ${}^{m}p_x^{n}$, $m \in \{a,i\}$, and $n \in \{a,i,d\}$ for $x = 20,21,...,45$.[9] As expected, ${}^{a}p_x^{a} > {}^{i}p_x^{a}$ at all ages prior to age 45, the age at which we assume all active players transition to inactivity. Table 3 contains examples of pmfs (graphed in Figures 2-5) for ages 25, 30, 35, and 40 computed by applying recursions (1a)-(1d), (2a)-(2d), and (3a)-(3b) to the

[5] We use the term "batters" to refer to "non-pitchers" who play field positions when opposing teams bat or who play more specialized roles like designated hitters (in the American League), pinch hitters and runners, or defensive players. Our data sources are T*he ESPN Baseball Encyclopedia* (Gillette and Palmer 2008) and the *Major League Handbook* (James 1998-2008). Mortality probabilities are from Center for Disease Control and Prevention (CDCP 2007).

[6] When interpreting our estimates for inactive players, the reader should keep in mind that they apply to players who have been inactive for four or fewer seasons.

[7] Of course, other definitions of "active" could be used such as requiring a player to appear in (say) 20 games in a year in order to be considered active in that year. Another definition might relate to some minimum time on the active roster of a major league team during a year.

[8] Among the youngest players were B. J. Upton, and Justin Upton (age 19) and Andres Blanco (20). Julio Franco (age 48), Rickey Henderson (44), Andres Galarraga (43) were the oldest. Julio Franco has remained active beyond our assumed truncation age *TA* = 46. This is exceptional; cases like this occur with practically zero probability.

[9] See the Appendix for counts of active and inactives and transitions between states.


DOI: 10.2202/1559-0410.1263

transition probabilities in Table 2. At each age, the first column by age refers to currently active batters and the second column to currently inactive batters.

**Table 1. Age PMF for Major League Baseball Batters**

| Age | Probability | Age | Probability |
|---|---|---|---|
| 19 | 0.0004 | 34 | 0.0408 |
| 20 | 0.0037 | 35 | 0.0346 |
| 21 | 0.0158 | 36 | 0.0246 |
| 22 | 0.0303 | 37 | 0.0169 |
| 23 | 0.0549 | 38 | 0.0115 |
| 24 | 0.0744 | 39 | 0.0074 |
| 25 | 0.0872 | 40 | 0.0047 |
| 26 | 0.0926 | 41 | 0.0025 |
| 27 | 0.0895 | 42 | 0.0012 |
| 28 | 0.0849 | 43 | 0.0004 |
| 29 | 0.0796 | 44 | 0.0003 |
| 30 | 0.0701 | 45 | 0.0001 |
| 31 | 0.0631 | 46 | 0.0001 |
| 32 | 0.0574 | 47 | 0.0001 |
| 33 | 0.0507 | 48 | 0.0001 |

Sum of Probabilities: 1.0000
Average Age of Batters: 28.62
Median Age of Batters: 28.00
Modal Age of Batters: 26.00
Standard Deviation: 4.31

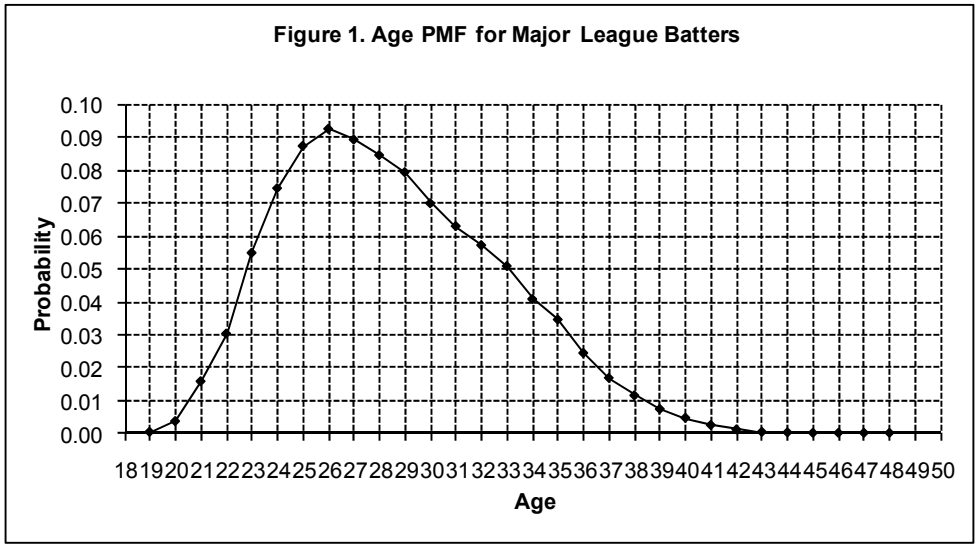


**Table 2. Transition Probabilities for Major League Baseball Batters**

| Age | ${}^{a}p_x^{a}$ | ${}^{a}p_x^{i}$ | ${}^{i}p_x^{a}$ | ${}^{i}p_x^{i}$ |
|---|---|---|---|---|
| 20 | 0.9119 | 0.0868 | 0.4994 | 0.4994 |
| 21 | 0.8137 | 0.1849 | 0.6657 | 0.3329 |
| 22 | 0.8371 | 0.1614 | 0.5991 | 0.3994 |
| 23 | 0.8506 | 0.1479 | 0.4510 | 0.5476 |
| 24 | 0.8277 | 0.1708 | 0.4029 | 0.5956 |
| 25 | 0.8366 | 0.1620 | 0.3073 | 0.6914 |
| 26 | 0.7958 | 0.2029 | 0.2023 | 0.7963 |
| 27 | 0.8096 | 0.1891 | 0.1600 | 0.8387 |
| 28 | 0.8100 | 0.1887 | 0.1039 | 0.8948 |
| 29 | 0.7903 | 0.2083 | 0.1110 | 0.8877 |
| 30 | 0.8111 | 0.1876 | 0.1243 | 0.8743 |
| 31 | 0.8189 | 0.1797 | 0.1051 | 0.8935 |
| 32 | 0.7926 | 0.2060 | 0.0930 | 0.9055 |
| 33 | 0.7691 | 0.2293 | 0.0617 | 0.9368 |
| 34 | 0.7788 | 0.2196 | 0.0421 | 0.9563 |
| 35 | 0.7566 | 0.2417 | 0.0177 | 0.9806 |
| 36 | 0.6793 | 0.3189 | 0.0295 | 0.9686 |
| 37 | 0.6618 | 0.3362 | 0.0129 | 0.9851 |
| 38 | 0.6335 | 0.3643 | 0.0152 | 0.9826 |
| 39 | 0.6122 | 0.3855 | 0.0097 | 0.9880 |
| 40 | 0.5371 | 0.4604 | 0.0130 | 0.9845 |
| 41 | 0.5817 | 0.4155 | 0.0188 | 0.9784 |
| 42 | 0.4985 | 0.4985 | 0.0000 | 0.9970 |
| 43 | 0.6645 | 0.3322 | 0.0000 | 0.9967 |
| 44 | 0.4982 | 0.4982 | 0.0000 | 0.9964 |
| 45 | 0.0000 | 0.9961 | 0.0000 | 0.9961 |


DOI: 10.2202/1559-0410.1263

**Table 3. Probability Mass Functions for Years of Future Activity for Batters at Various Ages, Given Activity Status**

| | Age 25 | | Age 30 | | Age 35 | | Age 40 | |
|---|---|---|---|---|---|---|---|---|
| Years | Active | Inactive | Active | Inactive | Active | Inactive | Active | Inactive |
| 0 | 0 | 0.213 | 0 | 0.5695 | 0 | 0.8893 | 0 | 0.9685 |
| 1 | 0.051 | 0.0884 | 0.1233 | 0.0871 | 0.2205 | 0.0416 | 0.4543 | 0.0147 |
| 2 | 0.0811 | 0.0901 | 0.1221 | 0.0769 | 0.2353 | 0.0263 | 0.229 | 0.0069 |
| 3 | 0.0889 | 0.0896 | 0.1332 | 0.0659 | 0.1766 | 0.0174 | 0.1581 | 0.0043 |
| 4 | 0.0935 | 0.0873 | 0.1299 | 0.0552 | 0.1313 | 0.0126 | 0.0537 | 0.0043 |
| 5 | 0.0993 | 0.0796 | 0.1068 | 0.0442 | 0.0906 | 0.0063 | 0.0534 | 0.0012 |
| 6 | 0.0929 | 0.0724 | 0.0941 | 0.0349 | 0.0666 | 0.0032 | 0.0516 | 0 |
| 7 | 0.0868 | 0.0648 | 0.0911 | 0.0245 | 0.0339 | 0.0017 | 0 | 0 |
| 8 | 0.0838 | 0.0548 | 0.0671 | 0.0167 | 0.0228 | 0.0008 | 0 | 0 |
| 9 | 0.0754 | 0.0441 | 0.0487 | 0.0108 | 0.0082 | 0.0006 | 0 | 0 |
| 10 | 0.0612 | 0.0355 | 0.0329 | 0.0066 | 0.0074 | 0.0002 | 0 | 0 |
| 11 | 0.0512 | 0.0279 | 0.0233 | 0.0036 | 0.0068 | 0 | 0 | 0 |
| 12 | 0.0445 | 0.0194 | 0.012 | 0.0021 | 0 | 0 | 0 | 0 |
| 13 | 0.0319 | 0.0132 | 0.0078 | 0.0011 | 0 | 0 | 0 | 0 |
| 14 | 0.0224 | 0.0085 | 0.003 | 0.0007 | 0 | 0 | 0 | 0 |
| 15 | 0.0148 | 0.0053 | 0.0026 | 0.0003 | 0 | 0 | 0 | 0 |
| 16 | 0.0099 | 0.0028 | 0.0021 | 0 | 0 | 0 | 0 | 0 |
| 17 | 0.0052 | 0.0016 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18 | 0.0032 | 0.0008 | 0 | 0 | 0 | 0 | 0 | 0 |
| 19 | 0.0014 | 0.0005 | 0 | 0 | 0 | 0 | 0 | 0 |
| 20 | 0.001 | 0.0003 | 0 | 0 | 0 | 0 | 0 | 0 |
| 21 | 0.0007 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 22 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 23 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 24 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | | | | | | | |
| Sum | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

**Figure 2. PMFs for Batters Age 25**

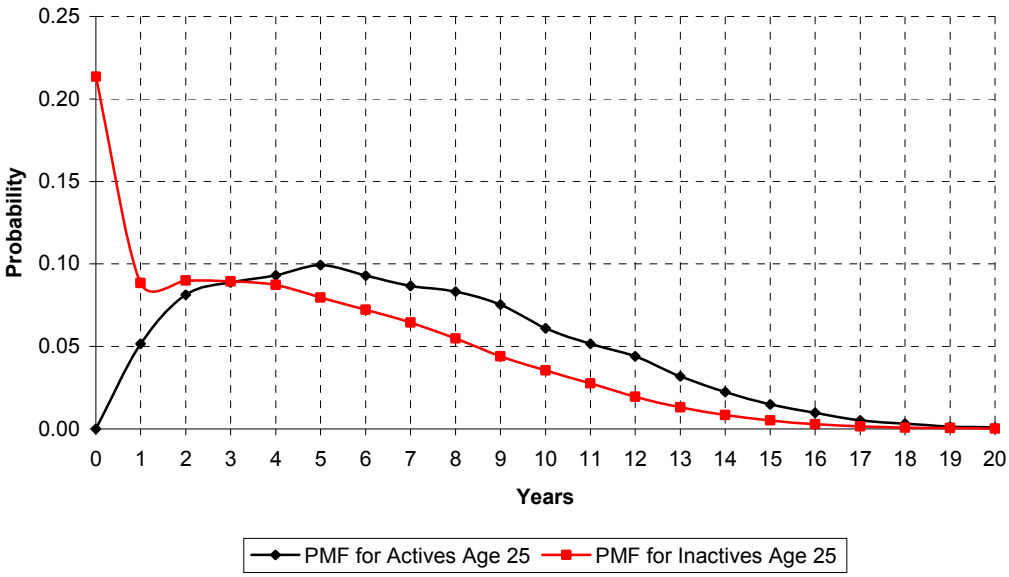
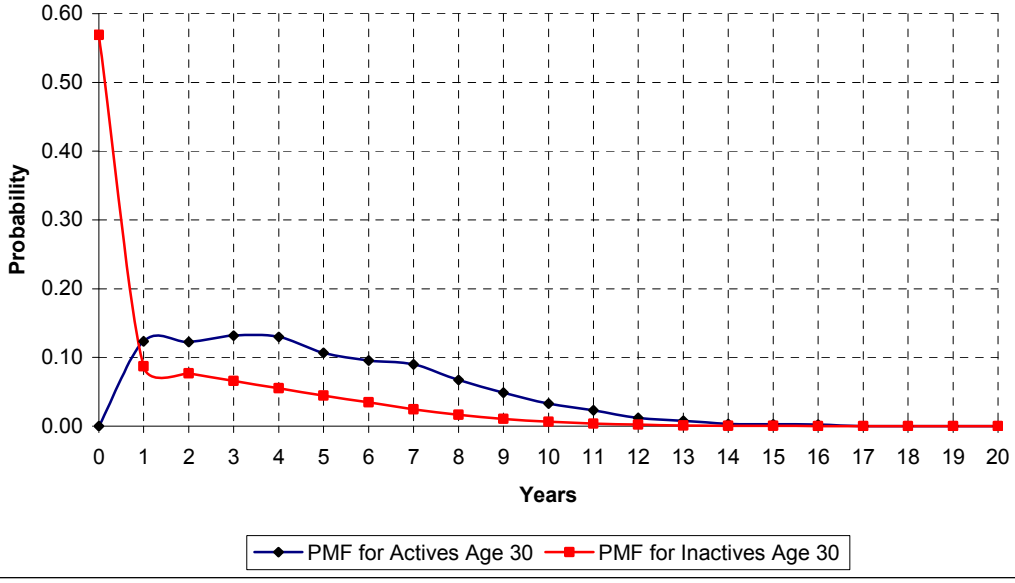



**Figure 3. PMFs for Batters Age 30**





DOI: 10.2202/1559-0410.1263

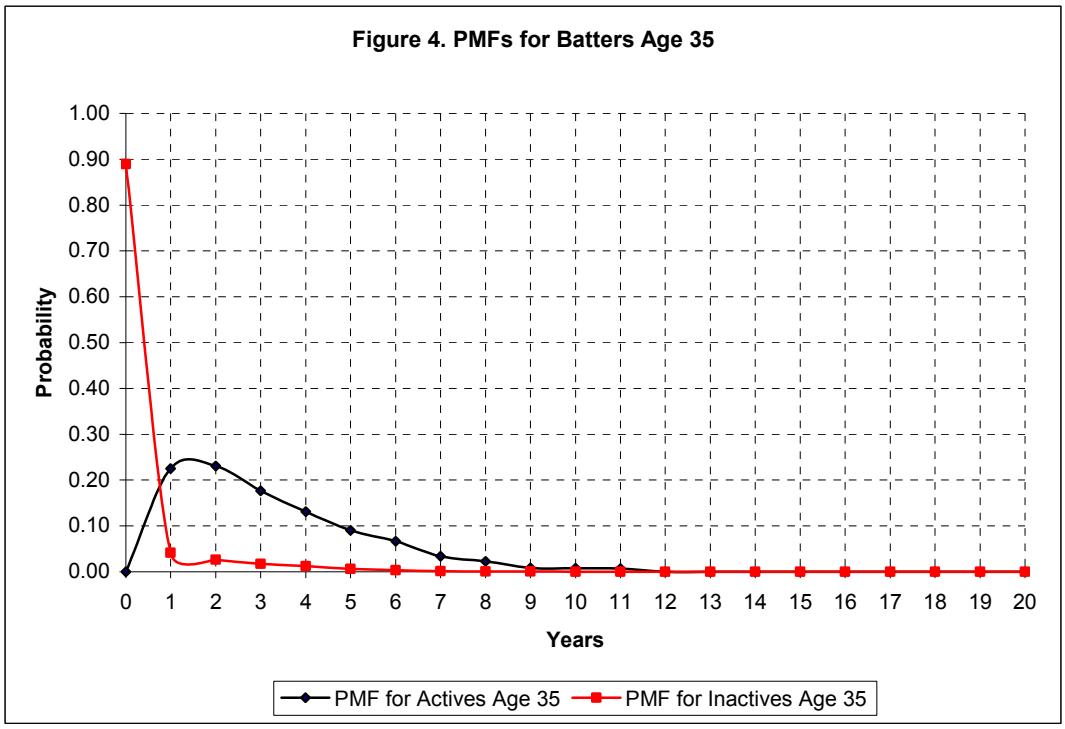


Figure 4. PMFs for Batters Age 35
Probability
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Years
PMF for Actives Age 35
PMF for Inactives Age 35

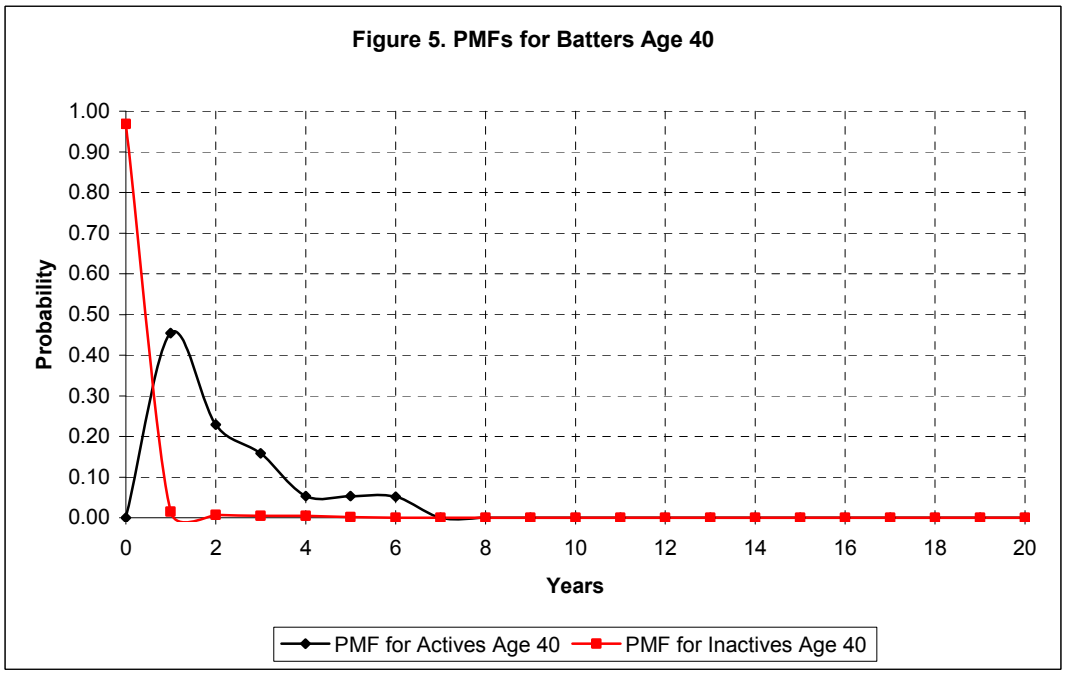


Figure 5. PMFs for Batters Age 40
Probability
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
0 2 4 6 8 10 12 14 16 18 20
Years
PMF for Actives Age 40
PMF for Inactives Age 40

Table 4 shows average years of major league activity (*i.e.*, baseball worklife expectancies) for batters between ages 20-45 using (4a). We expect currently active players to have longer worklives than currently inactive players, and indeed ${}^{a}e_{x}^{a} > {}^{i}e_{x}^{a}$. More strikingly, the differences $({}^{a}e_{x}^{a} - {}^{i}e_{x}^{a})$ grow between ages 21-31; and the relative differences expressed as either $({}^{a}e_{x}^{a} - {}^{i}e_{x}^{a}) / {}^{a}e_{x}^{a}$ or $({}^{a}e_{x}^{a} - {}^{i}e_{x}^{a}) / {}^{i}e_{x}^{a}$ increase for all ages beyond age 20. The final column in Table 4, $p_{YA}(Age, i, 0)$, shows probabilities that inactive players will not return to the major leagues. These probabilities increase with age. There is, for example, only approximately a one-in-four chance of reentry for an inactive 32-year-old batter (since the probability of an inactive player never returning to major league play is .7257).

As an example of the use of Table 4, consider an active 25-year-old batter who has a baseball worklife expectancy of ${}^{a}e_{25}^{a} = 6.89$ years. Suppose this player already had accumulated 2 years of activity before age 25. Then, expected career length would be 2 years prior to age 25 plus 6.89 years in expected future activity, or 8.89 total years. A batter similar in all respects (age, activity prior to age 25) but inactive at age 25 has an expected career of 2 years prior to age 25 plus 4.38 years of expected future activity, or 6.38 years.

The age and status-specific expectancies and other characteristics in Tables 5 and 6 give additional insight into career profiles. Tables 5 and 6 expand Table 4; they contain distributional characteristics for active and inactive batters age 20-45 computed with (4a)-(4h). Typically, distributions are skewed to the right with means exceeding medians which exceed modes. Standard deviations are large relative to means, especially for currently inactive batters. Wide probability intervals like the inter-quartile range also portray the high variability in future years in the major leagues.

What can an active 25-year-old batter anticipate in regard to his future career? We already noted a 6.89 expected future years in the major leagues. In addition, Table 5 tells us that the distribution of future years in the major leagues is skewed to the right (*SK* = .55). Thus, 50% of active batters will accumulate less than 6.89 future years. The median is 6 years, and the modal number of years is 5. The inter-quartile range is 4 to 9 years. The inactive 25-year-old counterpart with 4.38 expected years in the major leagues has median additional years of 4 and modal years of zero (which occurs with probability .2130 shown in Table 4). The inter-quartile range is 1 to 7 years.


DOI: 10.2202/1559-0410.1263

**Table 4. Worklife Expectancies for Major League Baseball Batters and the Probability an Inactive Player's Career is Over**

| Age | ${}^{a}e_x^a$ | ${}^{i}e_x^a$ | $p_{YA}(Age,i,0)$ |
|---|---|---|---|
| 20 | 10.43 | 8.93 | 0.0070 |
| 21 | 9.55 | 8.33 | 0.0114 |
| 22 | 8.82 | 7.39 | 0.0300 |
| 23 | 8.13 | 6.31 | 0.0717 |
| 24 | 7.45 | 5.38 | 0.1283 |
| 25 | 6.89 | 4.38 | 0.2130 |
| 26 | 6.36 | 3.50 | 0.3061 |
| 27 | 6.01 | 2.87 | 0.3828 |
| 28 | 5.63 | 2.35 | 0.4548 |
| 29 | 5.25 | 2.01 | 0.5069 |
| 30 | 4.94 | 1.65 | 0.5695 |
| 31 | 4.58 | 1.24 | 0.6498 |
| 32 | 4.17 | 0.89 | 0.7257 |
| 33 | 3.85 | 0.59 | 0.7997 |
| 34 | 3.58 | 0.40 | 0.8521 |
| 35 | 3.24 | 0.27 | 0.8893 |
| 36 | 2.89 | 0.23 | 0.9052 |
| 37 | 2.71 | 0.15 | 0.9326 |
| 38 | 2.53 | 0.12 | 0.9447 |
| 39 | 2.36 | 0.09 | 0.9592 |
| 40 | 2.18 | 0.06 | 0.9685 |
| 41 | 2.16 | 0.04 | 0.9812 |
| 42 | 1.99 | 0.00 | 1.0000 |
| 43 | 2.00 | 0.00 | 1.0000 |
| 44 | 1.50 | 0.00 | 1.0000 |
| 45 | 1.00 | 0.00 | 1.0000 |

**Table 5. Distributional Characteristics of Active Major League Baseball Batters***

| Age | Mean | Median | Mode | SD | SK | KU | 10% | 25% | 75% | 90% |
|---|---|---|---|---|---|---|---|---|---|---|
| 20 | 10.43 | 10.00 | 9.00 | 4.23 | 0.39 | 2.77 | 5.00 | 7.00 | 13.00 | 16.00 |
| 21 | 9.55 | 9.00 | 8.00 | 4.20 | 0.41 | 2.76 | 4.00 | 6.00 | 12.00 | 15.00 |
| 22 | 8.82 | 8.00 | 8.00 | 4.14 | 0.44 | 2.75 | 4.00 | 6.00 | 12.00 | 15.00 |
| 23 | 8.13 | 8.00 | 7.00 | 4.07 | 0.47 | 2.75 | 3.00 | 5.00 | 11.00 | 14.00 |
| 24 | 7.45 | 7.00 | 6.00 | 3.98 | 0.51 | 2.75 | 3.00 | 4.00 | 10.00 | 13.00 |
| 25 | 6.89 | 6.00 | 5.00 | 3.86 | 0.55 | 2.77 | 2.00 | 4.00 | 9.00 | 12.00 |
| 26 | 6.36 | 6.00 | 4.00 | 3.73 | 0.58 | 2.80 | 2.00 | 3.00 | 9.00 | 12.00 |
| 27 | 6.01 | 6.00 | 3.00 | 3.56 | 0.60 | 2.84 | 2.00 | 3.00 | 8.00 | 11.00 |
| 28 | 5.63 | 5.00 | 2.00 | 3.39 | 0.64 | 2.90 | 2.00 | 3.00 | 8.00 | 10.00 |
| 29 | 5.25 | 5.00 | 1.00 | 3.22 | 0.67 | 2.98 | 1.00 | 3.00 | 7.00 | 10.00 |
| 30 | 4.94 | 4.00 | 3.00 | 3.02 | 0.72 | 3.10 | 1.00 | 3.00 | 7.00 | 9.00 |
| 31 | 4.58 | 4.00 | 3.00 | 2.82 | 0.79 | 3.24 | 1.00 | 2.00 | 6.00 | 9.00 |
| 32 | 4.17 | 4.00 | 2.00 | 2.65 | 0.85 | 3.39 | 1.00 | 2.00 | 6.00 | 8.00 |
| 33 | 3.85 | 3.00 | 1.00 | 2.46 | 0.92 | 3.58 | 1.00 | 2.00 | 5.00 | 7.00 |
| 34 | 3.58 | 3.00 | 1.00 | 2.26 | 1.01 | 3.85 | 1.00 | 2.00 | 5.00 | 7.00 |
| 35 | 3.24 | 3.00 | 2.00 | 2.09 | 1.13 | 4.12 | 1.00 | 2.00 | 4.00 | 6.00 |
| 36 | 2.89 | 2.00 | 1.00 | 1.95 | 1.21 | 4.29 | 1.00 | 1.00 | 4.00 | 6.00 |
| 37 | 2.71 | 2.00 | 1.00 | 1.81 | 1.26 | 4.37 | 1.00 | 1.00 | 4.00 | 5.00 |
| 38 | 2.53 | 2.00 | 1.00 | 1.68 | 1.30 | 4.35 | 1.00 | 1.00 | 3.00 | 5.00 |
| 39 | 2.36 | 2.00 | 1.00 | 1.55 | 1.31 | 4.15 | 1.00 | 1.00 | 3.00 | 4.00 |
| 40 | 2.18 | 2.00 | 1.00 | 1.45 | 1.23 | 3.61 | 1.00 | 1.00 | 3.00 | 5.00 |
| 41 | 2.16 | 2.00 | 1.00 | 1.32 | 0.96 | 2.69 | 1.00 | 1.00 | 3.00 | 4.00 |
| 42 | 1.99 | 1.00 | 1.00 | 1.15 | 0.66 | 1.89 | 1.00 | 1.00 | 3.00 | 4.00 |
| 43 | 2.00 | 2.00 | 1.00 | 0.82 | 0.01 | 1.50 | 1.00 | 1.00 | 3.00 | 3.00 |
| 44 | 1.50 | 1.00 | 1.00 | 0.50 | 0.01 | 1.00 | 1.00 | 1.00 | 2.00 | 2.00 |
| 45 | 1.00 | 1.00 | 1.00 | 0.00 | --- | --- | 1.00 | 1.00 | 1.00 | 1.00 |

*See formulae (4a)-(4h) for definitions of all distributional characteristics.


DOI: 10.2202/1559-0410.1263

**Table 6. Distributional Characteristics of Inactive Major League Baseball Batters***

| Age | Mean | Median | Mode | SD | SK | KU | 10% | 25% | 75% | 90% |
|---|---|---|---|---|---|---|---|---|---|---|
| 20 | 8.93 | 9.00 | 7.00 | 9.90 | 1.27 | 1.78 | 4.00 | 6.00 | 12.00 | 15.00 |
| 21 | 8.33 | 8.00 | 7.00 | 4.22 | 0.40 | 2.76 | 3.00 | 5.00 | 11.00 | 14.00 |
| 22 | 7.39 | 7.00 | 6.00 | 4.20 | 0.43 | 2.74 | 2.00 | 4.00 | 10.00 | 13.00 |
| 23 | 6.31 | 6.00 | 5.00 | 4.16 | 0.49 | 2.73 | 1.00 | 3.00 | 9.00 | 12.00 |
| 24 | 5.38 | 5.00 | 0.00 | 4.06 | 0.59 | 2.77 | 0.00 | 2.00 | 8.00 | 11.00 |
| 25 | 4.38 | 4.00 | 0.00 | 3.86 | 0.76 | 2.94 | 0.00 | 1.00 | 7.00 | 10.00 |
| 26 | 3.50 | 3.00 | 0.00 | 3.58 | 0.96 | 3.28 | 0.00 | 0.00 | 6.00 | 9.00 |
| 27 | 2.87 | 2.00 | 0.00 | 3.30 | 1.14 | 3.70 | 0.00 | 0.00 | 5.00 | 8.00 |
| 28 | 2.35 | 1.00 | 0.00 | 3.01 | 1.33 | 4.23 | 0.00 | 0.00 | 4.00 | 7.00 |
| 29 | 2.01 | 0.00 | 0.00 | 2.80 | 1.48 | 4.73 | 0.00 | 0.00 | 3.00 | 6.00 |
| 30 | 1.65 | 0.00 | 0.00 | 2.54 | 1.71 | 5.56 | 0.00 | 0.00 | 3.00 | 6.00 |
| 31 | 1.24 | 0.00 | 0.00 | 2.20 | 2.06 | 7.18 | 0.00 | 0.00 | 2.00 | 5.00 |
| 32 | 0.89 | 0.00 | 0.00 | 1.86 | 2.51 | 9.64 | 0.00 | 0.00 | 1.00 | 4.00 |
| 33 | 0.59 | 0.00 | 0.00 | 1.49 | 3.13 | 13.96 | 0.00 | 0.00 | 0.00 | 2.00 |
| 34 | 0.40 | 0.00 | 0.00 | 1.18 | 3.80 | 19.67 | 0.00 | 0.00 | 0.00 | 1.00 |
| 35 | 0.27 | 0.00 | 0.00 | 0.95 | 4.47 | 26.29 | 0.00 | 0.00 | 0.00 | 1.00 |
| 36 | 0.23 | 0.00 | 0.00 | 0.85 | 4.83 | 30.18 | 0.00 | 0.00 | 0.00 | 0.00 |
| 37 | 0.15 | 0.00 | 0.00 | 0.67 | 5.67 | 40.15 | 0.00 | 0.00 | 0.00 | 0.00 |
| 38 | 0.12 | 0.00 | 0.00 | 0.59 | 6.18 | 46.55 | 0.00 | 0.00 | 0.00 | 0.00 |
| 39 | 0.09 | 0.00 | 0.00 | 0.49 | 7.03 | 58.05 | 0.00 | 0.00 | 0.00 | 0.00 |
| 40 | 0.06 | 0.00 | 0.00 | 0.42 | 7.81 | 69.82 | 0.00 | 0.00 | 0.00 | 0.00 |
| 41 | 0.04 | 0.00 | 0.00 | 0.31 | 9.93 | 109.50 | 0.00 | 0.00 | 0.00 | 0.00 |
| 42 | 0.00 | 0.00 | 0.00 | 0.00 | --- | --- | 0.00 | 0.00 | 0.00 | 0.00 |
| 43 | 0.00 | 0.00 | 0.00 | 0.00 | --- | --- | 0.00 | 0.00 | 0.00 | 0.00 |
| 44 | 0.00 | 0.00 | 0.00 | 0.00 | --- | --- | 0.00 | 0.00 | 0.00 | 0.00 |
| 45 | 0.00 | 0.00 | 0.00 | 0.00 | --- | --- | 0.00 | 0.00 | 0.00 | 0.00 |

*See formulae (4a)-(4h) for definitions of all distributional characteristics.

*Aggregate Measures of Years in Major League Baseball*

Formula (5) gives us the average remaining active time for batters, be they presently active or inactive, in the major leagues:[10]

$$(5) \qquad \sum_x \left[ \frac{{}^aN_x + {}^iN_x}{\sum_x ({}^aN_x + {}^iN_x)} \right] \left[ \frac{{}^aN_x \, {}^ae_x^a + {}^iN_x \, {}^ie_x^a}{{}^aN_x + {}^iN_x} \right] \qquad x = 19, 20, \ldots 45$$

$$= \sum_x \left[ \frac{{}^aN_x \, {}^ae_x^a + {}^iN_x \, {}^ie_x^a}{\sum_x ({}^aN_x + {}^iN_x)} \right]$$

where ${}^aN_x$ and ${}^iN_x$ denote the number of active and inactive players age *x*. This is a weighted average of baseball worklife expectancies for currently active and inactive batters where the weights are the fractions of active and inactive batters at those ages. Formula (5) yields 4.14 years for batters.[11] Currently active batters have remaining worklife of 5.58 years using formula (6).

$$(6) \qquad \sum_x \left[ \frac{{}^aN_x \, {}^ae_x^a}{\sum_x {}^aN_x} \right] \qquad x = 19, 20, \ldots 45$$

Both (5) and (6) depend on the age distribution of batters [active and inactive in (5) and only active in (6)]. Formula 5 has the interpretation that, if we consider

[10] This formula resembles a calculation that might be performed with a survivor table to calculate the average life expectancy in a stationary population. Let the number of survivors age *x* and beyond be $l_x, l_{x+1}, \ldots, l_\omega$, where $\omega$ denotes the youngest age after which no one survives; and let $\dot{e}_x, \dot{e}_{x+1}, \ldots, \dot{e}_\omega$ denote life expectancies at ages $x, x+1, \ldots, \omega$. Then, average life expectancy in the stationary population would be $(l_x \dot{e}_x + l_{x+1}\dot{e}_{x+1} + \cdots + l_\omega \dot{e}_\omega) / (l_x + l_{x+1} + \cdots + l_\omega)$. The latter formula is a weighted average of life expectancies with weights being the fraction of the population alive at each age.

[11] Baseball worklife expectancies in Table 4 start at age 20. At age 19, the estimated value of worklife is ${}^ae_{19}^a = 9.92$ years, but it utilizes transition probabilities from only two observations. We also assume that ${}^ie_{19}^a = 0$. These expectancies have little impact because there are so few 19-year-old batters in the major leagues.


DOI: 10.2202/1559-0410.1263

the pool of currently active and inactives batters, they have remaining careers of 4.14 years on average. Restricting the pool to active batters, the average remaining years of play is 5.58 years.

We consider a final aggregation question. Assume a youngster will become a major league batter but with unknown age of entry into the major leagues. What is expected career length? Table 7 and Figure 6 show the pmfs for entry age. Denote the probability of entry (debut) age by $p_{DA}(x)$, then expected career length is computed with formula (7).[12]

$$(7) \qquad \sum_x p_{DA}(x)({}^a e_x^a) \qquad\qquad x = 19, 20, \ldots 45$$

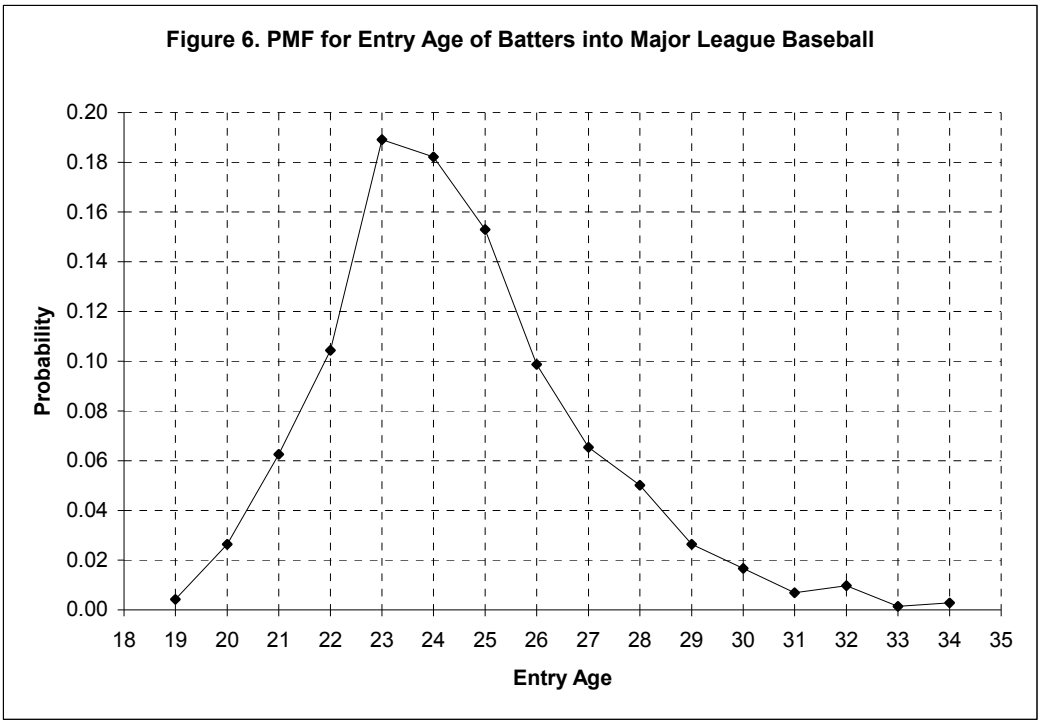
Expected career length using formula (7) is 7.40 years for batters.

**Table 7. Age of Entry of Batters into Major League Baseball**

| Age | Probability | Age | Probability |
|---|---|---|---|
| 19 | 0.0042 | 27 | 0.0654 |
| 20 | 0.0264 | 28 | 0.0501 |
| 21 | 0.0626 | 29 | 0.0264 |
| 22 | 0.1043 | 30 | 0.0167 |
| 23 | 0.1892 | 31 | 0.0070 |
| 24 | 0.1822 | 32 | 0.0097 |
| 25 | 0.1530 | 33 | 0.0014 |
| 26 | 0.0987 | 34 | 0.0028 |

Sum of Probabilities: 1.0000
Average Age of Entry: 24.43
Median Age of Entry: 24.00
Modal Age of Entry: 23.00
Standard Deviation: 2.46

[12] If a player was inactive for at least 1997, 1998, and 1999, we assumed his entry (debut) year occurred in the first subsequent year in which he became active. It is possible, but unlikely, that a player could have debuted earlier than 1997, subsequently been inactive for at least 1997, 1998, and 1999, and than turned active. However, Felix Jose is an example of a player who debuted in 1988, played through 1995, was out of major league baseball in 1996-1999, and reentered in 2000 at age 35. Chris Donnels is another example; he started his major league career in 1991, played through 1995, was inactive in 1996-99, and returned to the major leagues in 2000 as a 34-year-old player. To the extent other instances like these have occurred, our entry age would be older than the actual debut age. We also evaluated formula (7) using the Lahman baseball database for 1997-2008 for debut age and got very similar results for expected career length.




## IV. Recent Work on Baseball Worklife Expectancies

In an important recent paper, Witnauer, Rogers, and Saint Onge (WRO 2007), calculate career lengths of major league position players (batters) in the 20th century. Their data set consists of 5,989 players who started their major league careers between 1902 and 1993. Their data end in 2003 when a very high fraction, but not all, of the players they consider had completed their major league careers. They have three objectives: (1) to determine average career lengths (baseball worklife expectancies) through the use of life table methods, (2) to determine the relationship between age of entry and average career length, and (3) to estimate average career length in three baseball eras.

WRO adopt life table notation. They let $l_x$ denote the number of players who have completed $x$ years in the major leagues, $d_x$ is the number who exit play between $x$ and $x$+1, $L_x = .5(l_x + l_{x+1})$ is the average player years in the interval from $x$ to $x$+1, and total player years from $x$ is $T_x = \sum_{y=x}^{\omega} L_y$ where $\omega$ is the last year of play in the table. Average career length (baseball worklife expectancy) after completing $x$ years is calculated with $e_x = T_x / l_x$.


DOI: 10.2202/1559-0410.1263

WRO's Table 1 accomplishes their first objective. In that table, they relate years of play completed to average remaining career length. For example, a new player has $x = 0$ and $e_0 = 5.6$ remaining years in the major leagues, whereas a player who has had five years in the major leagues has an average remaining career length of $e_5 = 5.4$ years. Here, we would add one comment: In a life table, the number of years a person has lived is identical to a person's age, and therefore life expectancy for a person age *x* also refers to a person who has lived *x* years. WRO's Table 1 refers to players who have played *x* years in the major leagues, but players' ages are not shown in the table. For example, players of all ages contribute to $e_5 = 5.4$ as long as they have accumulated exactly five years in the major leagues at their particular ages. WRO consider some relationships between age and baseball careers in their Table 2.

Panel B of WRO's Table 2 has average career lengths for players who started major league play at ages 20,21,…,28. These expectancies apply to players of specific ages because entry age is, by definition, a player's age. Given entry age, WRO also show average remaining years after the accumulation of 1,2,…,11 years. These expectancies are specific to entry age but not specific to current age. For example, a player who enters the major leagues at age 25 has an expected career of 3.96 years; but it is 2.97 years after the accumulation of six years in the major leagues. Such a player must be at least age 31 but would be older if a player were out of the majors during any of the six years. WRO's second objective is accomplished by their Table 2.[13] Their final objective is achieved in Panel B of their Table 3, which contains average major league careers by three eras (Early Era 1902-45, Golden Age Era 1946-68, and Modern Era 1969-2003). They found that entry players had expected careers of 4.3, 6.6, and 6.9 years in the Early Era, Golden Age Era, and Modern Era, respectively.

How do the estimates in this paper compare to WRO's work? First, WRO let players play-out their careers and, with the benefit of hindsight, calculated expected time in major league baseball. This is a clever application of life table concepts. It is the strength of their work, but it simultaneously creates a problem. As WRO indicate, the data are right censored. They deal with right censoring by only selecting players who began their careers in 1993 or earlier. On the other hand, we include players only for the period 1997-2007 – what might be termed the Very Modern Era. We use our data to estimate age-specific transition probabilities, and right censoring ceases to be a problem because we do not have to follow players to the end of their careers. Our work is similar to a typical life

[13] Panel A of WRO's Table 2 has the proportion of players whose careers end during the years of play interval *x* to *x*+1 for starting ages 20,21,…,28. The entries in this part of their table are $((l_x - l_{x+1})/l_x$.

table based on age-specific mortality probabilities usually generated cross sectionally over a short period of time and not over a very long time period in which everyone born has died. Our transition probabilities also are estimated over a short time period. However, the similarity with a life table ends there because transitions only flow in one direction in a life table (*i.e.*, from living to dead), but our recursive formulae allow two way transitions between active and inactive states in major league baseball. Second, our baseball worklives are age specific; but, because of the Markov nature of our recursions, we do not know the history of a player before he reaches active or inactive status. An active player age 25 may have been in the major leagues for several years or in his first year. On the other hand, WRO worklives are based on years in the major leagues; but ages of players are not specified except for entry age players in the first row of their Table 2, Panel B. Third, of all the empirical results in both papers, there are two estimates that are directly comparable. WRO show average career years remaining for an entry player in their Modern Era (1969-2003) to be 6.85 years. Players entering major league baseball at various ages contribute to this figure. Our formula (7) for batters is directly comparable since batters also enter the major leagues at various ages. Formula (7) for batters evaluates to 7.40 years for the Very Modern Era (1997-2007). The estimates are close and the difference of .55 years is easily explainable. The time periods under consideration are somewhat different; and, as WRO show in their Table 3, career length for entry players has increased in more current times. In addition, a counting convention may be important. We have already indicated that end-of-year transitions lead to worklives .5 years longer than mid-period transitions. We use end of period transitions and WRO use mid-period transitions as manifested by the averaging embedded in their calculation of $L_x = .5(l_x + l_{x+1})$.[14] Thus, we can account for a half year of the difference in our estimate of 7.40 years and WRO's estimate of 6.85 years. Fourth, WRO's and our methods are complementary. Their approach works well for players who have finished their careers; and our approach enables us to focus on very current players, many of which are currently in the major leagues.

[14] Had WRO used end-of-period transitions, total player years $T_x'$ from $x$ would have been $T_x' = \sum_{y=x}^{\omega} l_y = .5l_x + \sum_{y=x}^{\omega} L_y$ and an average career would have been $e_x' = T_x'/l_x = .5 + e_x$. That is, average careers would have been one-half year longer.


DOI: 10.2202/1559-0410.1263

## V. Conclusion

We use current data and recursive formulae to estimate probability distributions for time spent in major league baseball. Distributions for both initially active and inactive players are shown in Table 3 at ages 25, 30, 35, and 40.

Tables 4 summarizes major league worklife expectancies for batters. At age 25, a typical entry age, active batters have expected careers of 6.89 years. As little as one year of inactivity is associated with much shorter baseball worklives: inactive 25-year-old batters have expected careers of 4.38. The probability of never playing again looms large for inactive players. As shown in Table 4, this probability is .21 for batters at age 25; it is .57 at age 30; and .89 at age 35.

Baseball worklife expectancy is the average of the additional-years-in-baseball random variable. Not only is there variation about the average but it is not symmetrical. Tables 5 and 6 give details. Some players have long careers which cause worklives to exceed medians which exceed modes. For active 25-year-old batters with average remaining careers of 6.89 years, the median and modal years are 6.00 and 5.00, respectively. Similar relationships hold for inactive batters. If we can think of age of entry into the major leagues as a random variable (conditional on entry actually occurring), the expected career for batters who randomly enter the major leagues is 7.40 years.

The Markov process nature of our model should be kept in mind when interpreting career expectancies. Active 25-year-old batters, for example, consist of players in their initial year in the major leagues as well as batters who have been in the majors for one or more years. When we report a worklife of 6.89 years, it incorporates all active batters age 25 regardless of their prior major league years. Players with a few years of major league experience presumably have demonstrated more baseball human capital than a rookie of the same age, and the former may have longer careers with greater probability. On the other hand, we include all players in our sample, even players called to the major leagues in September when rosters can be expanded. These players may have shorter careers and are more likely than others to not play in the major leagues in the following year. These are important factors that cause variance in careers of same-aged players.

The main shortcoming of the Markov process model is that it is only of order one – it has only a one period memory. An auto-regressive model of order two, for example, would have transition probabilities like ${}^{aa}p_x^a$, ${}^{ia}p_x^a$, ${}^{ai}p_x^a$, ${}^{ii}p_x^a$ where the left superscripts indicate activity status at ages $x-1$ and age $x$ and the right superscript indicates activity status at age $x+1$. It might be the case that ${}^{aa}p_x^a > {}^{ia}p_x^a$ and ${}^{ai}p_x^a > {}^{ii}p_x^a$, but the Markov model does not allow us to capture these inequalities and their implications for time spent in major league

baseball. Players who were active and players inactive at age $x-1$ are put in the same group if they are active at age $x$. Similarly, players who were active and players inactive at age $x-1$ are put in the same group if they are inactive at age $x$. Additional development of this topic would entail specifying models of order two or perhaps order three as well as the corresponding recursive formulae of probability mass functions. Such models would give us baseball worklife expectancies and other distributional characteristics based on more information than simply the current activity/inactivity state used in the Markov model.

## Appendix

### Table 1. Counts of Batters by Age[a]

| Age | ${}^{a}N_x$ | ${}^{a}N_x^a$ | ${}^{a}N_x^i$ | ${}^{i}N_x$ | ${}^{i}N_x^a$ | ${}^{i}N_x^i$ |
|---|---|---|---|---|---|---|
| 19 | 2 | 0 | 2 | 0 | 0 | 0 |
| 20 | 23 | 21 | 2 | 2 | 1 | 1 |
| 21 | 81 | 66 | 15 | 3 | 2 | 1 |
| 22 | 167 | 140 | 27 | 15 | 9 | 6 |
| 23 | 297 | 253 | 44 | 31 | 14 | 17 |
| 24 | 415 | 344 | 71 | 57 | 23 | 34 |
| 25 | 487 | 408 | 79 | 104 | 32 | 72 |
| 26 | 512 | 408 | 104 | 153 | 31 | 122 |
| 27 | 486 | 394 | 92 | 206 | 33 | 173 |
| 28 | 487 | 395 | 92 | 221 | 23 | 198 |
| 29 | 441 | 349 | 92 | 234 | 26 | 208 |
| 30 | 394 | 320 | 74 | 249 | 31 | 218 |
| 31 | 339 | 278 | 61 | 228 | 24 | 204 |
| 32 | 320 | 254 | 66 | 204 | 19 | 185 |
| 33 | 283 | 218 | 65 | 178 | 11 | 167 |
| 34 | 241 | 188 | 53 | 166 | 7 | 159 |
| 35 | 190 | 144 | 46 | 169 | 3 | 166 |
| 36 | 144 | 98 | 46 | 169 | 5 | 164 |
| 37 | 95 | 63 | 32 | 155 | 2 | 153 |
| 38 | 63 | 40 | 23 | 131 | 2 | 129 |
| 39 | 44 | 27 | 17 | 103 | 1 | 102 |
| 40 | 26 | 14 | 12 | 77 | 1 | 76 |
| 41 | 12 | 7 | 5 | 53 | 1 | 52 |
| 42 | 6 | 3 | 3 | 34 | 0 | 34 |
| 43 | 3 | 2 | 1 | 23 | 0 | 23 |
| 44 | 2 | 1 | 1 | 14 | 0 | 14 |
| 45 | 1 | 1 | 0 | 9 | 0 | 9 |
| 46 | 1 | 1 | 0 | 3 | 0 | 3 |
| 47 | 1 | 1 | 0 | 1 | 0 | 1 |
| 48 | 0 | 0 | 0 | 0 | 0 | 0 |
| 49 | 0 | 0 | 0 | 0 | 0 | 0 |
| 50 | 0 | 0 | 0 | 0 | 0 | 0 |

${}^{a}p_x^a = ({}^{a}N_x^a / {}^{a}N_x)(1 - \dot{p}_x^d)$ and ${}^{a}p_x^i = 1 - {}^{a}p_x^a - \dot{p}_x^d$

${}^{i}p_x^i = ({}^{i}N_x^i / {}^{i}N_x)(1 - \dot{p}_x^d)$ and ${}^{i}p_x^a = 1 - {}^{i}p_x^i - \dot{p}_x^d$

[a] We use The *ESPN Baseball Encyclopedia* (2008) and the *Major League Handbook* (1998-2008) to determine ${}^{a}N_x$, ${}^{a}N_x^a$, ${}^{a}N_x^i$, ${}^{i}N_x$, ${}^{i}N_x^a$, and ${}^{i}N_x^i$ Mortality probabilities are from Center for Disease Control and Prevention (2007).


DOI: 10.2202/1559-0410.1263

## References

Center for Disease Control and Prevention (CDCP), U.S. Department of Health and Human Services. (2007). *National Vital Statistics Reports*, *United States Life Tables, 2003*. 54(14). Hyattsville, MD.

Gillette, Gary and Pete Palmer. (2008). *The Baseball Encyclopedia*, Fifth Edition. New York, NY: Sterling Publishing Company.

James, Bill. *Major League Handbook,* 1998 – 2008 Editions. Skokie, IL: STATS Publishing.

Skoog, Gary R. and James E. Ciecka. (2002). Probability mass functions for labor market activity induced by the Markov (increment-decrement) model of labor force activity. *Economics Letters*, 77(3), 425-431.

Witnauer, William D., Richard G. Rogers, and Jarron M. Saint Onge. (2007). Major league baseball career length in the 20th century. *Population Research and Policy Review*, 26(4), 371-3.