4-2008

Quasigeometric Distributions and Extra Inning Baseball Games

Darren B. Glass
Gettysburg College

Philip J. Lowry
City University of New York

Glass, Darren B., and Philip J. Lowry. "Quasigeometric Distributions and Extra Inning Baseball Games." Mathematics Magazine 81.2 (April 2008), 127- 137.

Quasigeometric Distributions and Extra Inning Baseball Games

Abstract
Each July, the eyes of baseball fans across the country turn to Major League Baseball's All-Star Game, gathering the best and most popular players from baseball's two leagues to play against each other in a single game. In most sports, the All-Star Game is an exhibition played purely for entertainment. Since 2003, the baseball All-Star Game has actually 'counted', because the winning league gets home field advantage in the World Series. Just one year before this rule went into effect, there was no winner in the All-Star Game, as both teams ran out of pitchers in the 11th inning and the game had to be stopped at that point. Under the new rules, the All-Star Game must be played until there is a winner, no matter how long it takes, so the managers need to consider the possibility of a long extra inning game. This should lead the managers to ask themselves what the probability is that the game will last 12 innings. What about 20 innings? Longer?

In this paper, we address these questions and several other questions related to the game of baseball. Our methods use a variation on the well-studied geometric distribution called the quasigeometric distribution. We begin by reviewing some of the literature on applications of mathematics to baseball. In the second section we will define the quasigeometric distribution and examine several of its properties. The final two sections examine the applications of this distribution to models of scoring patterns in baseball games and, more specifically, the length of extra inning games.

Keywords
Major League Baseball, MLB, All-Star Game, Extra Innings, Baseball Rules

Disciplines
Mathematics | Probability | Sports Sciences | Statistics and Probability

Quasigeometric Distributions and Extra Inning Baseball Games

DARREN GLASS
Gettysburg College
200 N. Washington St.
Gettysburg, PA 17325
dglass@gettysburg.edu

PHILIP J. LOWRY
City University of New York
New York, NY 10016
plowry1176@aol.com

Each July, the eyes of baseball fans across the country turn to Major League Baseball's All-Star Game, gathering the best and most popular players from baseball's two leagues to play against each other in a single game. In most sports, the All-Star Game is an exhibition played purely for entertainment. Since 2003, the baseball All-Star Game has actually 'counted', because the winning league gets home field advantage in the World Series. Just one year before this rule went into effect, there was no winner in the All-Star Game, as both teams ran out of pitchers in the 11th inning and the game had to be stopped at that point. Under the new rules, the All-Star Game must be played until there is a winner, no matter how long it takes, so the managers need to consider the possibility of a long extra inning game. This should lead the managers to ask themselves what the probability is that the game will last 12 innings. What about 20 innings? Longer?

In this paper, we address these questions and several other questions related to the game of baseball. Our methods use a variation on the well-studied geometric distribution called the quasigeometric distribution. We begin by reviewing some of the literature on applications of mathematics to baseball. In the second section we will define the quasigeometric distribution and examine several of its properties. The final two sections examine the applications of this distribution to models of scoring patterns in baseball games and, more specifically, the length of extra inning games.

1. Sabermetrics

While professional baseball has been played for more than a century, it has only been in the last few decades that people have applied mathematical tools to analyze the game. Bill James coined the term 'Sabermetrics' to describe the analysis of baseball through objective evidence, and in particular the use of baseball statistics. The word Sabermetrics comes from the acronym SABR, which stands for the Society for American Baseball Research [**12**].

Before SABR was ever organized, and before sabermetrics was a word, the influence of statistics over the strategy used by a manager in professional baseball was minimal. No manager would have ever thought of having charts on what each batter had done against each pitcher in the league. Now things are different. Since the publication of Michael Lewis's book *Moneyball* in 2003 [**10**], even most casual baseball fans have become familiar with Sabermetric statistics such as OPS ("on-base plus slugging", which many people feel is a better measure of offensive skill than the traditional statistics such as batting average or RBIs) and Win Shares (a statistic developed by Bill

James in [**8**] which attempts to measure the all-around contributions of any player), and there has been a proliferation of books and websites for the more dedicated fans to pursue these interests.

Sabermetrics has had a profound influence not just in the living room, but also in the clubhouse as it has begun to affect the strategy of the game. In the last decade, Sabermetrics devotees such as Billy Beane, Theo Epstein, Paul DePodesta, and Bill James himself have all worked in the front offices of Major League baseball teams, and these approaches are often given some of the credit for the Red Sox winning the 2004 World Series [**6**].

Sabermetricians attempt to use statistical analysis to answer all sorts of questions about the game of baseball: whether teams should intentionally walk Barry Bonds, whether Derek Jeter deserves his Gold Glove, which players are overpaid (or underpaid), when closing pitchers should be brought into the game, and whether or not batting order matters are just some of the questions that have had many words written about them. For readers interested in these questions, websites such as *Baseball Prospectus* [**2**] and journals such as *By The Numbers* [**5**] are a great place to start reading. Alan Schwarz's book *The Numbers Game* [**13**] provides an excellent historical perspective, and Albert and Bennett's book *Curve Ball: Baseball, Statistics, and the Role of Chance in the Game* [**1**] is a good introduction to some of the mathematical techniques involved.

One recurring theme in the sabermetric literature is the question of how likely certain records are to be broken and how unlikely these records were to begin with. For example, now that Barry Bonds has set the career homerun record, many people are curious whether we should expect to see any player pass Bonds in our lifetime. Several recent articles ([**3**], [**4**]) in *The Baseball Research Journal* have asked the question "How unlikely was Joe DiMaggio's 56 game hitting streak?" and have come to different answers depending on the methods they use to look at the question. This question is of the same flavor as the question we address in Section Four, as we use the mathematical models developed to examine how likely a 20 inning game is to occur, and how unlikely the longest recorded game of 45 innings really was.

2. Distributions

Geometric distributions. To begin, let us recall what we mean by a distribution in the first place.

DEFINITION 2.1. A probability distribution on the natural numbers is a function $f : \mathbb{N}_0 \to [0, 1]$ (where $\mathbb{N}_0$ denotes the nonnegative integers) such that $\sum_{n=0}^{\infty} f(n) = 1$. The mean (or expected value) of a discrete distribution f is given by $\mu = \sum nf(n)$ and the variance is given by $\sigma^2 = \sum (n - \mu)^2 f(n)$.

DEFINITION 2.2. A geometric distribution is a distribution such that for all $n \geq 1$, $f(n) = f(0)\ell^n$ for some fixed $0 < \ell < 1$.

We note that geometric distributions are the discrete version of the exponential decay functions which are found, for example, in half-life problems. In particular, if f is a geometric distribution, then we see that

$$1 = \sum_{n=0}^{\infty} f(n)$$

$$= \sum_{n=0}^{\infty} f(0)\ell^n$$

$$= f(0) \sum_{n=0}^{\infty} \ell^n$$

$$= \frac{f(0)}{1-\ell},$$

therefore $f(0) = 1 - \ell$. Thus, the entire distribution is determined by the value of ℓ. It is a straightforward computation to see that the mean of this distribution is $\frac{\ell}{1-\ell}$ and the variance is $\frac{\ell}{(1-\ell)^2}$.

Quasigeometric distributions. In this paper, we wish to discuss a variation of geometric distributions which can reasonably be referred to as quasigeometric distributions, as they behave very similarly to geometric distributions. These distributions are defined so that they are geometric other than at a starting point. In particular, we want there to be a common ratio between $f(n)$ and $f(n+1)$ for all $n \geq 1$ but not (necessarily) the same ratio between $f(0)$ and $f(1)$. To be explicit, we make the following definition:

DEFINITION 2.3. A quasigeometric distribution is a distribution so that for all $n \geq 2$, $f(n) = f(1)d^{n-1}$ for some $0 < d < 1$. We call d the *depreciation constant* associated to the distribution.

Just as geometric distributions are completely determined by the value of k, a quasigeometric distribution is entirely determined by the values of d and $f(0)$ (which we will often denote by a). In particular, a computation analogous to the one above shows that for $n \geq 1$, $f(n) = (1-a)(1-d)d^{n-1}$. Given this, it is possible to compute the mean and variance of the distribution as follows:

$$\begin{aligned}
\mu &= \sum_{n=0}^{\infty} nf(n) \\
&= \sum_{n=1}^{\infty} n(1-a)(1-d)d^{n-1} \\
&= (1-a)(1-d) \sum_{n=1}^{\infty} nd^{n-1} \\
&= (1-a)(1-d)(1-d)^{-2} \\
&= \frac{1-a}{1-d}, \qquad (1) \\
\sigma^2 &= \sum_{n=0}^{\infty} n^2 f(n) - \mu^2 \\
&= \sum_{n=1}^{\infty} n^2(1-a)(1-d)d^{n-1} - \mu^2 \\
&= (1-a)(1-d) \sum_{n=1}^{\infty} n^2 d^{n-1} - \mu^2
\end{aligned}$$

$$= \frac{(1-a)(1+d)}{(1-d)^2} - \frac{(1-a)^2}{(1-d)^2}$$
$$= \frac{(1-a)(a+d)}{(1-d)^2}. \tag{2}$$

Conversely, we note that a quasigeometric distribution is uniquely determined given μ and σ^2 (although not all pairs (μ, σ^2) determine a quasigeometric distribution). In particular, if $s > |m - m^2|$ and we set

$$a = \frac{m + s - m^2}{m + s + m^2} \quad \text{and} \quad d = \frac{m^2 + s - m}{m^2 + s + m},$$

then the quasigeometric distribution given by a and d will have mean m and variance s. In statistics, this method of describing a distribution is called the method of moments.

3. Baseball scoring patterns

Runs scored per inning. It has been observed by several people (see [**9**], [**15**], [**16**]) that the number of runs scored per inning by a given baseball team fits a quasigeometric distribution (although they do not use this language). In TABLE 1, we have provided a table of the probabilities that a given number of runs is scored in an inning based on several different datasets and we see that the same general pattern persists. Woolner's data [**16**] separates teams by their strength, trying to see if teams that score an average of 3.5 runs per game have different scoring patterns than those that score 5.5 runs per game. The data compiled by Jarvis [**9**] separates teams by league to see how scoring patterns are affected by the different rules (designated hitter, etc.) as well as the different cultures in the American League and the National League.

TABLE 1: Probability of scoring a given number of runs in an inning

Dataset	0 runs	1	2	3	4	5
Woolner (all)	0.730	0.148	0.068	0.031	0.014	0.006
Woolner (3.5 RPG)	0.760	0.140	0.059	0.024	0.011	0.004
Woolner (5.5 RPG)	0.679	0.161	0.079	0.042	0.022	0.009
Jarvis (AL)	0.722	0.151	0.070	0.032	0.014	0.006
Jarvis (NL)	0.731	0.150	0.067	0.030	0.013	0.006

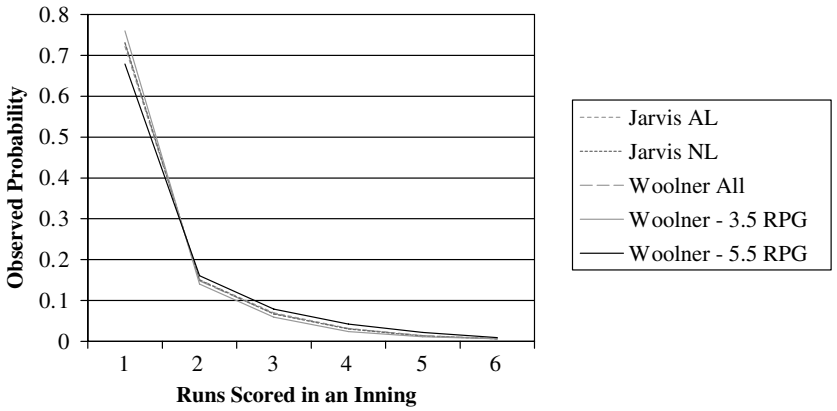

Figure 1 Runs scored per inning

One notes from FIGURE 1 that, after an initial dropoff, the probability of scoring a given number of runs appears to fit a typical exponential curve. This is not a surprising result to baseball fans, because it is what one would intuitiviely expect if the only way that a runner reached base was by hitting a single: the first run would be more difficult to score as it requires multiple hits, but the probability that each additional run will score coincides with the probability that the batter gets a hit.

Over each of these different data sets, one can compute the mean and standard deviation, and in turn compute the associated values of a and d that would define the appropriate quasigeometric distribution from the equations at the end of Section 2.

TABLE 2: Computations of a and d for datasets

Dataset	m	s	a	d
Woolner (all)	0.484	0.999	0.727	0.436
Woolner (3.5 RPG)	0.408	0.896	0.772	0.444
Woolner (5.5 RPG)	0.627	1.173	0.642	0.429
Jarvis (AL)	0.503	1.024	0.715	0.435
Jarvis (NL)	0.478	0.986	0.730	0.434

We see from TABLE 2 that the value of the depreciation constant d does not change very much, even when we look across leagues or across varying strengths of teams. In fact, it does not change significantly even if we compare different eras. This observation will be the key assumption of our model. For the duration of this paper, we will assume that scoring patterns in a given inning fit a quasigeometric distribution with a value of $d = 0.436$ for the depreciation constant as suggested by the full database in [**16**]. The value of a, on the other hand, does change significantly with the strength of a team. One way of interpreting this result is that the difference between the quality of teams is mainly in the probability that they score a single run in a given inning. Then, after the first batter crosses the plate, all teams are more or less equally successful at continuing the scoring drive.

The work of Smith, featured in [**14**], shows that an average major league team scores 0.487 runs per inning. TABLE 3 computes the probability that a team which scores at this rate will score a given number of runs in an inning according to this quasigeometric model, and compares this with the probability observed in Woolner's dataset.

TABLE 3: Runs per inning: Quasigeometric Model vs. Woolner's Data

Number of Runs	Predicted Prob.	Observed Prob.
0	0.725	0.730
1	0.155	0.148
2	0.068	0.068
3	0.029	0.031
4	0.013	0.014
5	0.006	0.006

Runs scored per game. Of course, as any baseball fan who has watched his team squander a lead can tell you, games are not won or lost by the number of runs scored in a given inning but instead by the number of runs scored in the full nine (or more) innings. So one would like a formula to determine the number of runs scored in a

nine-inning game. In order to do so, we first make the assumption that all innings are independent of one another. While this assumption is almost certain to be overly strong—teams are likely to face similar pitchers, weather, and park effects in consecutive innings—it greatly simplifies the problem. Furthermore, we will see that it leads to mathematical results that match with actual game data.

We denote the probability that a team scores n runs in nine innings as $F(n)$, and note that

$$F(n) = \sum f(n_1) f(n_2) \dots f(n_8) f(n_9),$$

where the sum ranges over all 9-tuples of nonnegative integers $(n_1, \dots, n_9)$ which sum to n and $f(n_j)$ is the probability that the team scores n_j runs in inning j.

If a team scores n runs in a game, then we know that the team must score in between one and $\min(n, 9)$ different innings. Breaking up by these cases, we can compute

$$F(n) = \sum_{i=1}^{\min(n,9)} \left(\begin{array}{c} 9 \\ i \end{array} \right) f(0)^{9-i} \left(\sum f(n_1) \dots f(n_i)\right),$$

where the interior sum is over all ordered i-tuples of positive integers summing to n. If we now invoke our assumption that the probability of scoring a given number of runs in an inning is quasigeometric (and independent of the inning), and therefore that $f(0) = a$ and $f(n_i) = (1-a)(1-d)d^{n_i - 1}$ for all $n_i \geq 1$, we can calculate that

$$F(n) = \sum_{i=1}^{\min(n,9)} \left(\begin{array}{c} 9 \\ i \end{array} \right) \left(\begin{array}{c} n-1 \\ i-1 \end{array} \right) a^{9-i} d^{n-i} (1-a)^i (1-d)^i.$$

In this formula, i represents the number of innings in which the team scores, a represents the probability that a team goes scoreless in a given inning, and d represents the depreciation constant, which we are assuming is equal to 0.436 for all teams. One way to view the $\binom{n-1}{i-1}$ term is that it counts the number of ways to divide n runs among i innings. It will be more useful to us to translate this result in terms of the strength of a given team. To do this, we note that Equation (1) showed that to model a team that scores an average of m runs per inning we should choose $a = 1 - m(1-d)$. Doing so, we compute:

$$F(n) = \sum_{i=1}^{\min(n,9)} \left(\begin{array}{c} 9 \\ i \end{array} \right) \left(\begin{array}{c} n-1 \\ i-1 \end{array} \right) m^i (1-(1-d)m)^{9-i} (1-d)^{2i} d^{n-i},$$

where again d is the depreciation constant 0.436 and m represents the average number of runs per inning that a team scores. TABLE 4 computes $F(n)$ for a team that scores the historical average of 0.487 runs per inning and compares these values with the empirical distribution of runs per game scored by National League teams between 1969 and 2002.

One sees that this quasigeometric model appears to give a good approximation of reality, and therefore we might want to see how this type of model can be used to answer many different types of questions. In the following section, we will look at the question of how often we should expect games to last 20 innings or more, but before moving on to that, we think it would be interesting to note that one could use this model to compute the odds that a team of a given strength would beat another team of a given strength. In particular, we note that the 2003 Atlanta Braves scored an average of 0.618 runs per inning, whereas the 2003 New York Mets scored an average of 0.443 runs per inning. While this is clearly a lopsided matchup, one of the beautiful things

TABLE 4: Number of runs per game predicted by model vs. actual game data

Number of Runs	$F(n)$ = Prob in game	% of NL Games
0	0.055	0.062
1	0.107	0.108
2	0.138	0.139
3	0.145	0.148
4	0.135	0.134
5	0.114	0.113
6	0.091	0.088
7	0.068	0.068
8	0.046	0.049
9	0.034	0.033
10	0.023	0.023

about the game of baseball is that underdogs often win, and one wonders what the probability of the Mets winning a given game against the Braves would have been.

One can use the quasigeometric model in order to approach this question. In particular, we can use the strengths of each team to calculate $F_B(n)$ (resp. $F_M(n)$), the probability that the Braves (resp. the Mets) will score n runs in nine innings. Given these functions and the assumption that their scoring is independent of each other, we can compute that there is roughly a 31% chance that the Mets will be ahead after nine innings, a 60% chance that the Braves will win, and a 9% chance that the game will go into extra innings. If one looks at what actually happened in the games played between the two teams in 2003, we see that the Braves won 11 of the 19 (or 58%) of the games, with none going into extra innings. These results correspond quite closely with the predictions of our model, given the small sample size involved.

4. Extra inning games

One of the things about baseball that its fans love the most, and its detractors like the least, is the fact that it is free of the artificial boundaries of time within which the clock confines other sports. This freedom from time constraints helps to shape the unique charm that is an evening at the ballpark, for fans never know when they may be the first to be enchanted until past sunrise by the first-ever wild ten-hour 46-inning slugfest.

This idea brings us back to the question posed in the introduction: what is the probability that a given baseball game lasts twenty innings or more? Alternatively, there has only been one Major League Baseball game to last twenty-six innings in history, and one could ask if the mathematical models predict more or fewer than have actually occurred.

To answer these questions, one must first consider what the probability is that a game goes into extra innings at all. In particular, this asks whether or not the two teams have scored the same number of runs after nine innings of play. To compute this, we make the assumption that the scoring of the two teams is independent of one another, and thus that T, the probability that the game is tied after 9 innings, can be computed as

$$T = \sum_{i=0}^{\infty} F_A(i) F_B(i)$$

where $F_A(i)$ and $F_B(i)$ are the probabilities that Team A and Team B score i runs in nine innings, the formula for which was given above.

We note that the formula above tells us that if we assume both teams score the major league average of 0.487 runs per inning, then $T = 0.103$, so that we would expect just over 10% of games to go into extra innings. In reality, 9.22%—18,440 of the 199,906 major league games played between 1871 and 2005—have gone into extra innings. The discrepancy between this number and what our model predicts likely arises from two facts. First, our model assumes that the teams are scoring independently of one another. In reality, this assumption is likely to be not quite true, as external factors (humidity, altitude, pitching, etc.) may cause games to be either high or low scoring, and there may be a psychological factor that promotes teams to score more if the other team is a few runs ahead, or to stop trying once they are blowing out the other team.

The other factor that we can think of is trickier to get a handle on. The above calculation assumes that both teams are average, but in most games one team will be better than the other. For an extreme example, we look at the AL East in 2003, where the Detroit Tigers scored an average of 0.405 runs per inning and the Boston Red Sox scored an average of 0.659 runs per inning. This is the largest discrepancy between two teams in the same league in over 25 years. In this case, the formula predicts that only 8.4% of games will go into extra innings. While this specific example is an extreme, it suggests that when two teams have differing abilities to score runs, we should expect fewer extra inning games even if the overall average number of runs scored is held constant. This expectation is confirmed by the data in TABLE 5, where the rows and columns represent the strengths of the two teams playing, and T is the probability that they will be tied after nine innings, according to our model.

Given that a large number of games are played between teams with widely differing abilities to score runs, this would suggest that our model will predict a larger number of extra inning games than actually occur.

After the ninth inning, the game will conclude at the end of the first inning after which the score is not tied. Therefore, if we let k be the probability that the two teams score the same number of runs in a given inning, then the probability that a game is still tied after n innings is Tk^{n-9} and for $n > 9$ the probability that it ends after n innings is $Tk^{n-10}(1-k)$.

We note that we are making several assumptions here. First, we are assuming that there is no effective difference between the tenth inning and any later inning as far as offensive production is concerned. We also assume that, at least as far as extra innings go, if k is the probability that the two teams score the same number of runs in a given inning then the probability that they score the same number of runs in each of

TABLE 5: Probability of a tie game between two teams of various strengths

	0.405	0.437	0.487	0.537	0.617	0.659
0.405	0.1148	0.1119	0.1065	0.1006	0.0903	0.0848
0.437	0.1119	0.1097	0.1056	0.1007	0.0918	0.0869
0.487	0.1065	0.1056	0.1033	0.1000	0.0932	0.0892
0.537	0.1006	0.1007	0.1000	0.0982	0.0936	0.0905
0.617	0.0903	0.0918	0.0932	0.0936	0.0921	0.0904
0.659	0.0848	0.0869	0.0892	0.0905	0.0904	0.0896

n consecutive innings is k^n. We note that our intuition suggests that due to different strategies in the late parts of the game, as well as fatigue amongst the players, that the scoring distribution might be different as a game progresses, but the data seems to suggest that this difference is negligible. For details, see [**14**].

In order to proceed, it will now suffice to figure out what value k should have. Our first attempt to do so was to use an empirical number coming from the data itself, as detailed in [**11**]. In this paper, we will use the quasigeometric model of scoring which we have developed in order to construct a theoretical value of k. In particular, if we let $a = f_A(0)$ and $b = f_B(0)$ be the respective probabilities of each team going scoreless in an inning, we can compute:

$$\begin{aligned} k &= \sum_{i=0}^{\infty} f_A(i) f_B(i) \\ &= ab + \sum_{i=1}^{\infty} f_A(i) f_B(i) \\ &= ab + \sum_{i=1}^{\infty} (1-a)(1-d_A) d_A^{i-1} (1-b)(1-d_B) d_B^{i-1} \\ &= ab + \frac{(1-a)(1-b)(1-d_A)(1-d_B)}{d_A d_B (1-d_A d_B)}. \end{aligned}$$

If we continue with our assumption that $d_A = d_B = 0.436$, and we let m_A (resp. m_B) be the average number of runs per inning scored by team A (resp. team B), then this simplifies to give us

$$k = 1 - 0.564 m_A - 0.564 m_B + 0.4423 m_A m_B.$$

We are now ready to see the fruits of our labor. Let us first look at the case where both of our teams score the major league average number of runs, which means $m_A = m_B = 0.487$. Then it follows that $T = 0.103$ and that $k = 0.55588$. In particular, the probability of a game lasting n innings is $(0.103)(0.4442)(0.5558)^{n-10}$ for all $n \geq 10$. The chart below calculates this probability for games of varying lengths. We have also included the actual number of major league ballgames from 1871 through 2005 that have lasted that long, as well as the number of games that our model predicts.

Comparing the model to the past . . . and to the future. So how "rare" are extremely long marathon baseball games? The second author has built a database, discussed in detail in [**11**], of baseball games lasting 20 innings or more. Among these are included the Brooklyn at Boston 26-inning major league record game in 1920, the Rochester at Pawtucket 33-inning minor league game in 1981, and the longest known ballgame: a 45-inning amateur game in Mito, Japan in 1983. Our theoretical model predicts the 26-inning major league record game is not as rare as empirical data would indicate, but the 33-inning minor league record game and 45-inning amateur record game are significantly more rare than empirical data would indicate.

In the previous section we saw that there is approximately a 0.00029 probability that any given game lasts 20 or more innings. Assuming that the probability of any two games lasting this long is independent of one another we can compute that the probability that out of any collection of x games at least one of them lasts 20 or more innings is $1 - (1 - 0.00029)^x$. There are 2340 major league games played each year and therefore we should expect a 50% chance to experience a major league game of 20 or more innings in any given season. Similarly, our model predicts that there will

TABLE 6: Number of games of a given length predicted vs. actual number

# Innings	Prob in given game	Actual MLB	Expected MLB
≤ 9	0.8973	181,466	179,349.6
10	0.04574	8106	9142.3
11	0.02542	4561	5080.9
12	0.01413	2549	2824.3
13	0.007857	1413	1570.4
14	0.004367	831	872.8
15	0.002427	426	485.1
16	0.001349	259	269.6
17	0.0007502	140	149.9
18	0.0004170	69	83.3
19	0.0002318	40	45.9
20	0.0001288	20	25.1
21	7.163E-05	10	13.9
22	3.982E-05	8	7.8
23	2.213E-05	2	4.3
24	1.230E-05	3	2.4
25	6.839E-06	2	1.3
26	3.802E-06	1	0.74
27	2.113E-06	0	0.41
28	1.174E-06	0	0.23
29	6.530E-07	0	0.13
30	3.630E-07	0	0.071
Total	1.0	199,906	199,906

be 0.939 major league games that would have lasted 27 or more innings by now. In fact, we have not yet had such a game in 135 years of major league play. These results indicate that the 26-inning game in Boston is not an outlier from what one would expect from our model.

If we assume that the scoring patterns in minor league games are similar to those in major league games (an assumption for which there is some evidence), and in particular that scoring is quasigeometric with the same values of a and d, then we should expect 6.68 minor league games to have gone 27 or more innings. In fact, we have had 6 such games. If we look further we see that the model predicts that we will have had only 0.087 minor league games which lasted 33 innings. In fact, we have had one such game.

Furthermore, there is a 99.3% chance we will have a minor league marathon of 20 or more innings in any given season, a 0.13% chance we will have a minor league game of 34 or more innings in any given season, a 1.32% chance of seeing a minor league game of 34 innings or more in any given decade, and a 9.4% chance of seeing a minor league game of 34 innings or more in a lifetime of 75 years.

Our model allows us to estimate the probability of games lasting a certain number of innings or longer. This is an alternative method, and perhaps a more easily understood way to express how unlikely are marathons of a certain length. We will now use this approach to compare relative probabilities of breaking the current records for major league and minor league games.

Assuming that major league baseball continues to have 30 teams play a 162-game season, there is a 50% chance we will see a major league game go 27 innings or more in the next 60 years. There is a 95% chance we will see a major league game go 27

innings or more in the next 260 seasons. So the 85-year old 26-inning major league record, while rare, is not so rare that we should assume it will stand for another ninety seasons.

As far as minor league games go, if we assume that there continue to be 13,714 minor league games played per year, then there is a 50% chance we will see a minor league game go 34 innings or longer in the next 565 years. There is a 95% chance we will see a minor league game go 34 innings or more in the next 2,445 years. So the 24-year old 33-inning minor league record may be very rare, and although it could be broken at any time, we should not expect to see it broken anytime soon.

It is interesting to note that, despite several assumptions that seem like they are not entirely accurate, this model does a good job of predicting the number of marathon games. This gives us hope that the quasigeometric model of baseball scoring can be used to answer a variety of questions about the game of baseball, and that it will be a useful tool in the growing research in Sabermetrics.

REFERENCES

1. James Albert and Jay Bennett, *Curve Ball: Baseball, Statistics, and the Role of Chance in the Game,* Copernicus, New York, 2001.
2. Baseball Prospectus, `http://www.baseballprospectus.com`
3. Charles Blahous, The DiMaggio Streak: How Big a Deal Was It? *Baseball Research Journal* (1994) 41–43.
4. Bob Brown and Peter Goodrich, Calculating the Odds: DiMaggio's 56-Game Hitting Streak, *Baseball Research Journal* (2003) 35–40.
5. *By The Numbers*, Statistical Analysis Committee of SABR, issues available for download at `http://www.philbirnbaum.com/`
6. *Mind Game: How the Boston Red Sox Got Smart, Won a World Series, and Created a New Blueprint for Winning.* Steven Goldman, ed., Workman Publishing, New York, 2005.
7. Bill James, *1977 Baseball Abstract,* self-published, 1977.
8. Bill James, *Win Shares,* STATS Publishing, New York, 2002.
9. John F. Jarvis, A Collection of Team Season Statistics, `http://www.knology.net/~johnfjarvis/stats.html`
10. Michael Lewis, *Moneyball: The Art of Winning an Unfair Game*, W. W. Norton, New York, 2003.
11. Philip J. Lowry, I Don't Care If I Ever Come Back: Marathons Lasting 20 or More Innings, *Baseball Research Journal* (2004) 8–28.
12. Society for American Baseball Research, `http://www.sabr.org`
13. Alan Schwarz, *The Numbers Game: Baseball's Lifelong Fascination with Statistics*, Thomas Dunne Books, New York, 2004.
14. David W.Smith, Coming from Behind: Patterns of Scoring and Relation to Winning, presentation at SABR Denver Convention, 2003.
15. TangoTiger. Tango Distribution. Tango On Baseball website, `http://www.tangotiger.net`
16. Keith Woolner, An analytic model for per-inning scoring distributions, Baseball Prospectus, March 4, 2000. `http://www.baseballprospectus.com/article.php?articleid=472`