Parabola Volume 58, Issue 1 (2022)

Batting survival and recurrence relations in cricket

Bernard Kachoyan[1]

1 Introduction

In cricket, a batter[2] accumulates runs during an innings, then is either dismissed (gets "out") or remains "not out" at the end of the innings[3]. A batter's career then consists of the number of runs accumulated in each of a number of innings. In a previous Parabola article [1], we discussed the ability to describe the career of a cricket batter in terms of a survival function. This survival function, denoted by $S(j)$ or S_j is the probability that a batter scores more than j runs; that is,

$$S(j) = S_j = Pr(r > j) \tag{1}$$

where r is the random variable representing the number of runs scored. This function can be used to make detailed comparisons between batters or make predictions on future performance.

Note that the inequality in (1) is strict, so that $S(0) = S_0 \neq 1$. In fact, $S_0 = 1 - Pr(r = 0)$ where $Pr(r = 0)$ is the probability that you will score no runs (which in cricket is called scoring a "duck"). Many previous analyses have determined that S_0 is indeed a special point in cricket and that getting off the mark is especially difficult for batters. Hence, from here on in it will be defined as a specific independent parameter.

The previous article [1] concentrated on how a survival function can be constructed from a batter's batting history using a statistical technique called Product Limit Estimation (PLE). The benefit of the PLE was that it was able to easily incorporate the not-out innings. However, one problem with that technique is that by following the data closely, the resultant survival curve is a step function, constant in between drops where a batter has been dismissed. It also is unable to be extrapolated beyond the data itself[4]. So some estimate of the underlying "true" survival function which can be used to predict future performance, is desirable.

On a more aesthetic level, it is also of interest to know how to construct a batting survival function based on more fundamental aspects of batting rather than a simple

[1]Dr Bernard Kachoyan is an Adjunct Associate Professor University of New South Wales. Email: Bernard.kachoyan@unsw.edu.au

[2]"batter" is now the accepted word to refer to both male and female cricketers, replacing "batsman".

[3]For those not familiar with cricket, a batter can remain not out at the end of an innings for various reasons, including that only ten out of the eleven batters need to be dismissed for the innings to be concluded.

[4]For example, if a batter has never scored more than 50 runs, we are unable to use the PLE to give an estimate of the probability of him scoring more than 50 in subsequent innings.

line of best type statistical approach. This article will outline how a batting survival function can be constructed from simple building blocks of the probability of being dismissed at each run total, and the probability of scoring a particular number of runs. We will show how under some simplifying assumptions.

The following is a summary of the analysis conducted in [2] and readers who are interested in more detail or wish further supporting references can refer to that paper. For the purposes of this article, we are concentrating more on outlining the concepts of difference equations and their solution rather than application to specific cricketers.

2 Simplified derivation

To highlight the concepts, we initially simplify the problem by considering the case when a batter can only score one run at a time. Then at every score j, there are two possibilities: either the batter is dismissed (with probability μ_j) before they score an extra run, or they score a run (with probability $1 - \mu_j$) and move on to $j + 1$ runs. So, the probability of scoring more than $j + 1$ runs is the probability of scoring more than j runs times the probability that one was not dismissed at $j + 1$ runs; that is,

$$S_{j+1} = (1 - \mu_{j+1})S_j \,. \tag{2}$$

An equation for the sequence S which defines the value at a specific point in terms of the values at previous points is termed a *recurrence relation* or *difference equation*. This type of equation should be familiar to most students in the context of geometric progressions and compound interest. The relation (2) is *of first order* because it relates the current step only to the immediately previous step.

In the most general case so far discussed, the probability of dismissal μ_j is a function of the runs already scored and equation (2) shows how a survival function can be developed in a step by step manner from this data alone. Interesting things happen, however, where the probability of dismissal is constant, μ, where $0 < \mu < 1$ [5], and this will be assumed from here on. In this case,

$$S_{j+1} = (1 - \mu)S_j \,. \tag{3}$$

This equation is simply a geometric progression with common ratio $1 - \mu$. This leads directly to a general solution:

$$S_j = Pr(r > j) = S_0(1 - \mu)^j \quad \text{for} \quad j = 0, 1, 2, \ldots \,. \tag{4}$$

The term S_0 can be seen as the necessary initial condition with which to start the recurrence (2).

This so-called geometric survival function has the interesting property of being memoryless. This property is usually defined in terms of waiting time and states that the distribution of a waiting time until a certain event occurs does not depend on how

[5] The extreme cases $\mu = 0$ and $\mu = 1$ correspond to nonrealistic and trivial cases.

much time has elapsed already[6]. For batting, the property refers to the expected number of runs scored, in particular that

$$Pr(r > j + n \mid r > j) = Pr(r > n)\,. \tag{5}$$

Here, $Pr(r > j + n \mid r > j)$ denotes the conditional probability that the value of r is greater than $j + n$ given that it is greater than j.

For example, $Pr(r > 60 \mid r > 50) = Pr(r > 10)$. In other words, the probability of scoring more than 60 given you have already scored 50 - that is, more than ten runs from 50 - is the same as the probability of scoring more than 10 runs in the first place. The system has no "memory" of what has happened previously.

This may seem intuitively obvious since we have assumed the probability of dismissal is a constant but has some profound implications in interpretation of batting statistics. What may be less clear is how this survival function relates to the traditional notion of a batting average. This relationship is explained in Box 1.

Box 1 Relationship of Survival Function to Batting Average

It is well known from survival theory that the average survival, or in this case the expected number of runs for each innings, $\overline{r}$, is given by

$$\overline{r} = \sum_{j=0}^{\infty} S_j\,.$$

For the simple case (4), this is the sum of a geometric progression which gives

$$\overline{r} = S_0/\mu\,.$$

Thus, the average score is inversely proportional to the probability of dismissal at each score. This at first might seem unrelated to the traditional batting average in cricket but can in fact be derived directly from its definition: the batting average is $\frac{N}{D}$, where N is the number of runs scored and D is the number of times the batter was dismissed. If we denote D_0 to be the number of innings where no runs were scored (ducks), then

$$\frac{N}{D} = \frac{(D - D_0)}{D} \times \frac{N}{(D - D_0)} = \frac{S_0}{\mu}\,.$$

Where now S_0 is the previously defined fraction of non-ducks, and $\mu = (D - D_0)/N$ is the ratio of the number of dismissals where at least one run was scored to the number of runs, which is indeed the probability of getting out per run.

[6]For example, how long I have to wait for a bus does not depend on the time since the last bus.

3 The full problem

Having set the scene, the question now becomes: how does the analysis change when we now consider that more than one run can be scored each time? How will that affect the theoretical survival function and can we retain the nice properties of the simple case? For simplicity, we will assume that one can only score 1, 2, 3 or 4 runs with each scoring shot and that the probability of scoring each does not depend on j.

Following the same reasoning as above, we can construct a survival function by recursively relating S_{j+4} to S_j, by considering

$$\begin{aligned} S_j = \;& Pr(\text{survived more than } j \text{ runs}) \\ = \;& \big[Pr(\text{survived more than } j-1 \text{ runs}) \times Pr(\text{scored a } 1) \\ + \;& Pr(\text{survived more than } j-2 \text{ runs}) \times Pr(\text{scored a } 2) \\ + \;& Pr(\text{survived more than } j-3 \text{ runs}) \times Pr(\text{scored a } 3) \\ + \;& Pr(\text{survived more than } j-4 \text{ runs}) \times Pr(\text{scored a } 4)\big] \times Pr(\text{survived run } j) \end{aligned}$$

or, expressed differently,

$$S_j = (1-\mu)(p_1 S_{j-1} + p_2 S_{j-2} + p_3 S_{j-3} + p_4 S_{j-4})\,. \tag{6}$$

Here, p_n represents the probability of scoring n runs, where $p_1 + p_2 + p_3 + p_4 = 1$. This fourth order[7] recurrence relation might look a bit daunting at first but, since μ, p_1, p_2, p_3, p_4 are each constant, we can make the assumption that the solution looks like the first order case, namely

$$S_j = k\beta^j \tag{7}$$

for some β and $j = 0, 1, 2, \ldots$.

If we substitute (7) into (6), then we soon find that this works, provided that β is a solution to the following fourth order polynomial equation:

$$f(\beta) = \beta^4 - (1-\mu)(p_1\beta^3 + p_2\beta^2 + p_3\beta + p_4) = 0\,. \tag{8}$$

In the most general case, there may be four solutions to this polynomial, let's say β_1, β_2, β_3 and β_4. Each of these generate their own geometric solution of the form (7) with their own constant of proportionality k_m, where $m = 1, \ldots 4$. The general solution to the recurrence relation is a linear combination of these[8]:

$$S_j = \sum_{m=1}^{4} k_m \beta_m^j\,. \tag{9}$$

This is a sum of four geometrical progressions, so it is not in itself a geometrical progression. However, as j gets larger, the solution with the largest magnitude of β, let's

[7] It is of fourth order since each step is related to the four previous steps.

[8] For those not familiar with this notation, Σ just represents the sum of terms from $m = 1$ to $m = 4$.

call it β_{max}, becomes dominant over the others and hence the survival curve will approach a geometric one $S_j \propto \beta_{max}^j$ for large j.

One can use the special knowledge of the coefficients of the polynomial, namely that $0 < \mu < 1$, $0 \leq p_1, p_2, p_3, p_4 \leq 1$ and $p_1 + p_2 + p_3 + p_4 = 1$, to prove certain things about this polynomial which confirm its practical utility (see Box 2 for an outline of the proofs). Firstly, there is only one positive solution. Secondly, the positive root is less than 1. This latter point guarantees that $S_j \to 0$ as j grows large. Finally, all the negative solutions have magnitude less that the positive one. Thus, the general solution (9) will indeed converge to a single geometric distribution, the common factor of which will be the positive root of the fourth-degree polynomial[9] (8). The value of the common factor is obviously in turn dependent on the probability of dismissal and the individual probabilities of each scoring shot.

The question now is: how quickly does this convergence happen? If it happens too slowly, then we are not justified in our assumption that a single geometric distribution may be made to fit the data of a particular batter. In order to answer this question, we need to run the recursion forward in time from the first run. A little thought should tell you that a fourth order recurrence such as (6) needs four initial conditions since the function S_j does not exist for $j < 0$. Another way of looking at it is that there are four unknown k's in the general solution (9) which requires four conditions to determine. Unfortunately, starting this recursion is a little complicated due to the possibilities of scoring more than one run at a time. The probability of surviving beyond a score of 1, S_1, is given by the probability that one survived beyond 0 then did not get out at 1:

$$S_1 = S_0(1 - \mu p_1) = S_0\big((1-\mu)p_1 + p_2 + p_3 + p_4\big) = S_0\big((1-\mu)p_1 + (1 - p_1)\big)\,.$$

The term in the final brackets is equivalent to the probability that either the batter scored 1 then did not get out at 1, or that the batter scored 2, 3 or 4. Similarly,

$$S_2 = S_1 p_1(1-\mu) + S_0 p_2(1-\mu) + S_0(1 - p_2 - p_1)$$

or, in words:

$$\begin{aligned} S_2 = Pr(&\text{survived past 1, scored a 1 and did not get out}) + \\ &Pr(\text{scored a 2 from 0 and did not get out}) + \\ &Pr(\text{scored a 3 or 4 from 0}) \end{aligned} \quad (10)$$

Continuing in this manner gives:

$$\begin{aligned} S_3 &= (1-\mu)p_1 S_2 + (1-\mu)p_2 S_1 + (1-\mu)p_3 S_0 + p_4 S_0 \\ S_4 &= (1-\mu)p_1 S_3 + (1-\mu)p_2 S_2 + (1-\mu)p_3 S_1 + (1-\mu)p_4 S_0 \end{aligned}$$

and subsequent values S_5, S_6 etc. can then be determined from the recurrence relation (6).

[9] If we add the possibility of sixes being scored as well, then the polynomial becomes one of degree 6.

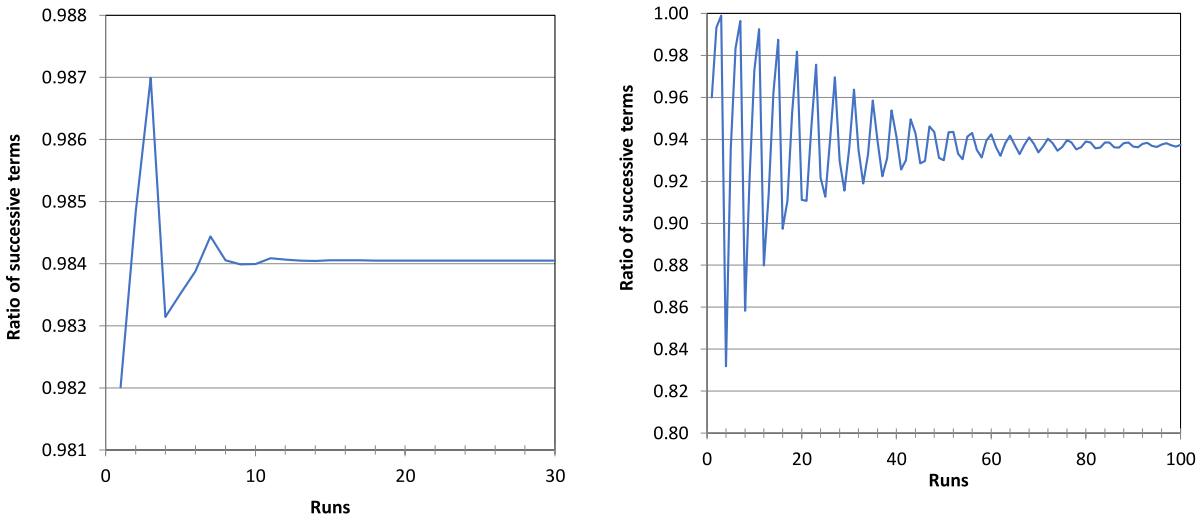

Figure 1: The ratio of successive terms in the recurrence relation (6), taken from [2]. Left (typical values): $p_1 = 0.60$, $p_2 = 0.14$, $p_3 = 0.04$, $p_4 = 0.22$ and $\mu = 0.029$. Right (unrealistic values): $p_1 = 0.2$, $p_2 = 0$, $p_3 = 0$, $p_4 = 0.8$ and $\mu = 0.2$.

Box 2: Proof of properties of the polynomial $f(\beta)$

To prove there is only on positive solution, one can use, for example, Descartes' rule of signs. This rule states that if a polynomial is ordered by descending value of the exponent, then the number of positive roots of the polynomial is either equal to the number of sign differences between consecutive non-zero coefficients or is less than it by an even number. The polynomial $f(\beta)$ has only one sign change in the coefficients hence has only one positive root. To prove that the positive root is less than 1, note that $f(0) = -p_4 < 0$ and $f(1) = \mu > 0$ so the single root must be between 0 and 1. To prove that all of the negative solutions have magnitude less that the positive ones, consider that

$$f(\beta) - f(-\beta) = -(1-\mu)(2p_1\beta^3 + 2p_3\beta) < 0 \text{ for } \beta > 0$$

This is a strict inequality, due to a realistic restriction that $p_1 \neq 0$. Then by letting β_1 and β_2 to be the positive and a negative root of $f(\beta)$ respectively, we obtain

$$f(\beta_1) - f(-\beta_1) = -f(-\beta_1) < 0 \text{ or } f(-\beta_1) > 0\,.$$

Since $f(0) < 0$, the negative root must be greater than $-\beta_1$, so $|\beta_2| < |\beta_1|$.

Having gone to all this trouble, reference [2] found that the convergence was actually very fast in most cases where realistic values of the parameters were chosen and is even quite fast for unrealistic cases (See Figure 1). In fact, the combination of 4 solutions in (9) can be considered to have converged to a single geometric distribution by the time 10 runs have been scored. This fact is not only useful at the start of the innings but also indicates that, should any of the parameters change during the innings, the solution would settle to a new steady state quite quickly.

As a final note, the exact solution (9) is a sum of geometric progressions so the mean can be easily expressed mathematically but it involves the solutions of the quadratic polynomial equation.

References

[1] B. Kachoyan, Batting is life and death, *Parabola* **51 (2)** (2015).

[2] B. Kachoyan and M. West, Deriving an exact batting survival function in cricket, *Australasian Conference on Mathematics and Computers in Sport (ANZIAM Mathsport 2018)*, 2018. ISBN: 978-0-646-99402-4