

Survival analysis for batsmen in international cricket

Levasen Reddy

19213612

STK795 Research Report

Submitted in partial fulfillment of the degree

BCom(Hons) Statistics

Supervisor: Dr Paul J van Staden

Department of Statistics, University of Pretoria



4 November 2022

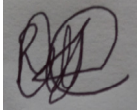
Abstract

This research project will look at developing survival analysis for batsmen in international cricket by using the product limit estimator and derivations of the Kaplan-Meier estimator that makes use of survival functions as well as hazard functions. This report will focus on Twenty20 cricket and its development over the past 20 years with regards to the players and their survival functions and make statistical interpretations on data gathered and inferences on models constructed.

Keywords: Kaplan-Meier Estimator, hazard function, Product limit estimator, censored, uncensored, Hypothesis test

Declaration

I, *Levasen Reddy* declare that this essay, submitted in partial fulfilment of the degree BCom(Hons) Data Science and Statistics, at the University of Pretoria, is my own work and has not been previously submitted at this or any other tertiary institution.



Levasen Reddy



Dr Paul J van Staden

4 November 2022

Date

Acknowledgements

This research was approved by the Ethics Committee, Faculty of Natural and Agricultural Sciences of the University of Pretoria, Reference Number: NAS116/2019

I would like to thank the Tomorrow Trust organization for providing me with a full bursary for my post graduate studies.

I would like to thank Tamryn Govender with assisting in the editing of my document.

Contents

1	Introduction	7
1.1	Cricket	7
1.2	Aims and Objectives	7
1.3	Outline of report	8
2	Literature Review	9
2.1	Survival analysis	9
2.2	Kaplan-Meier estimate	9
2.3	Other Performance Measures	10
3	Methodology	13
3.1	Exposition of the Kaplan-Meier estimate	13
3.2	Adaptation of Kaplan-Meier estimator in batting scorecards	18
4	Application	19
4.1	Standard survival analysis	20
4.2	Survival analysis of batting innings	21
4.3	Survival analysis of batting position	23
4.4	Nelson Aalen - Hazard function	25
4.5	Hypothesis testing	27
5	Conclusion	30
	Appendix A	32

List of Figures

1	Survival function	17
2	Hazard plot	17
3	Aaron Finch Survival Analysis	20
4	Dinesh Chandimal and Nicholas Pooran Survival Analysis	21
5	1st innings vs 2nd innings of different batters	22
6	Survival analysis of Jos Buttler's batting positions	23
7	Survival analysis of Quinton de Kock batting positions	24
8	Survival analysis of Shakib al Hasan batting positions	24
9	Hazard function with 1 batsmen	25
10	Hazard function with 2 batsmen	26

11	Hazard function with 3 batsmen	26
12	Hypothesis testing for Shakib Al Hasan for his first innings vs second innings	27
13	Hypothesis test output for Shakib Al Hasan	28
14	Hypothesis testing for Mohammed Nabi's batting position	28
15	Hypothesis testing output for Mohammed Nabi's	28

List of Tables

1	Test and ODI career records of Rahul Dravid.	11
2	Test,ODI and T20 career records of Faf du Plessis.	11
3	ODI and T20 career records of David Miller.	11
4	Group 1 data.	16
5	Group 2 data.	16
6	Variables Summary.	19

1 Introduction

1.1 Cricket

Cricket is one of the oldest team sports still in existence since its inception in the 16th century [Van Staden, 2017]. It is a game whereby the primary interaction occurs between a bat and ball. The game consists of 11 players on each team. The objective of cricket is that the game is divided into two innings or rounds. In the first innings, team A will bat and score as many runs until all batsmen are out or their overs¹ finish. Thereafter in the 2nd innings team B will attempt to chase down the score made by team A whilst team A tries to defend it. The primary ways of scoring runs is that two batters will run between wickets. Each time they run between wickets 1 run is scored. The second method is to score boundaries. A cricket field is encompassed by a rope or some boundary line. Four runs is scored if the ball crosses the boundary line and six runs is scored if the ball crosses the boundary line in the air without touching any part of the playing field. One of the various ways a batter can get out in cricket is to be bowled, whereby the bowler hits the stumps/wickets of the batter. A second way for a batter to be dismissed is to be caught out by hitting the ball in the air and a fielder catching it. A run out, whereby the fielders hit the stumps when the batters are trying to take a run is another form of being dismissal. There are 3 main popular types of cricket:

1. First class / Test cricket(unlimited overs): Consists of four innings, maximum of two innings per side which is played over four days(first class) or five days(Test). Test cricket is the format used in international cricket. It is the longest format of cricket with unlimited overs and broken up into three sessions of play a day.
2. One day cricket(limited overs): Is the format that consists of 40 overs or 50 overs per team and has a maximum of one innings per side. One day international(ODI) is the format used for international games.
3. Twenty20 cricket(limited overs): Consists of 20 overs a side and maximum one innings per side .Twenty20 International (T20I) cricket is played between countries

There are other formats of cricket around the world such as the T10 league and the Big Bash league amongst others that follow different rules. The T10 league consists of only 10 overs a side whereas the Big Bash league in Australia has something called the x-factor rule which allows for any team member to be substituted during the game, but these rules and leagues do not form part of the main focus of this study.

1.2 Aims and Objectives

The main aim of this report is to study and/or create a batting distribution of batters through a survival function. We will look at the career of several batters as well as their performance using the survival function and their success in the format of T20 cricket.

¹An over consists of six balls

The objectives we will try to achieve in this paper is:

1. To create a detailed exposition of the Kaplan-Meier estimator: we will be doing this by explaining the background of the Kaplan-Meier estimator and how it has been linked or manipulated to be used in different survival functions.
2. Adapting the Kaplan-Meier estimator and using it specifically in our batting survival function: we used the Python programming language (Python Software Foundation, <https://www.python.org/>) to do our analysis.
3. Estimation and interpretation of the batting survival functions and the batting hazard functions of various cricket players from different countries (Specifically players from the top ten ranked international teams)

1.3 Outline of report

Section 1 contains the introduction of the research report. Section 2 will have the literature review along with concepts related to the research project. Section 3 discusses the main theoretical concepts related to our survival analysis. Section 4 looks at the application segment of our research report. Section 5 is going to be our conclusion of the research project and our recommendations for further studies.

2 Literature Review

In the game of cricket, a batter's success is determined by the number of runs his/her scores. The more runs batters score the higher their rate of success is. But what determines how successful a batter has been is traditionally based on the batter's batting average. The batting average is the total runs scored divided by the number of completed innings. But these averages can be easily inflated since batters that do not get out in a match will only have their averages based on when they do get out which could possibly be in the next match [Kimber and Hansford, 1993]. Therefore, by looking at these averages that are seemingly inflated we do not necessarily see the success of a batter's career. Another performance measure that is indicative of success as a batter mentioned by [Van Staden, 2017] is the strike rate(SR) which is the number of runs scored per 100 balls faced.

2.1 Survival analysis

The survival function in layman's terms is just the probability given that some object of interest will survive beyond a certain time. For example we could investigate the probability that a light bulb will burn for longer than 1000 hours.

In the game of cricket statistics play a pivotal role in deciding whether a particular player would be suitable for a team, especially now since the game is heavily involved on the monetary side. The team's management is always looking for players that could bolster up their side. In one scenario where our survival function would become necessary is in first-class cricket in a country like England, where countries scout international players but are only limited to having a few international players in their squads. A survival function would provide them with a good indicator of the type of player they are looking at, especially when it comes to batters spending time at the crease and amassing large amounts of runs since conditions are quite conducive to fast bowling in England which many times will result in low-scoring games.

2.2 Kaplan-Meier estimate

[Kachoyan and West, 2016] discussed and used estimators such as the Kaplan-Meier estimator in deriving survival functions for batsmen. We can adapt estimators from Kaplan-Meier known as the Product Limit Estimator (PLE) for survival analysis [Kaplan and Meier, 1958]. This estimator was initially used in the medical industry. It was used to compute a survival function from a lifetime of data. In the medical industry it was used to measure the fraction of patients living after a certain amount of time after receiving treatment. The PLE is based on the conditional probability that an individual dies in the time interval from t_i to t_{i+1} , given survival up to time t_i is estimated as d_i/n_i where d_i is the number who die at time t_i , and n_i is the number alive just before time t_i , including those who will die at time t_i . In our case the Kaplan-Meier estimator can be adapted to view a

batter's innings as a lifespan, so you are born when you start your innings, 'live' for a certain amount of runs you score and then 'die' when you are dismissed or out. The PLE is particularly effective as it allows for censored and uncensored data in the context of cricket with the data being censored data, when the batter remains not out at the end of the innings. This could be for a few reasons such as the batter being the only one not dismissed or the team reaching the target to win (in that case both batters are not out). [Kachoyan and West, 2018]. Uncensored data is when the batters innings results in them being dismissed.

Another adaptation of the PLE is the batting hazard function also know as the failure rate. The failure rate $h(t)$ is defined from the survival function, $S(t)$, and is given by $h(t) = -S'(t)/S(t)$. The hazard function is a conditional probability of the failure density function $S'(t)$, conditioned that failure has not occurred at time t . If our constructed failure distribution is exponential, then the failure rate is constant which leads to the probability distribution being independent of the history, which is also known as the memoryless property. The exponential distribution tends to follow a geometric distribution as our sample size increases and the changes in probability is small. This can provide an outlook by ensuring independence on ball-on-ball analysis faced by batters, the geometric distribution shows that both the probability that an innings ends with each ball faced and that the probability the batter makes a scoring shot with each ball faced is constant. So this implies that even if batting follows a memoryless distribution and no matter what score you are on you have the same probability of getting out.

[Danaher, 1989] discussed at how a PLE can be applied to create a meaningful batting average estimate. It is also important that we take into account uncensored and censored data as this can affect the overall reliability of the estimate. If we look at Equation 1 it refers to a non-parametric product limit estimator whereby:

- a_i is the ranked distinct uncensored scores
- d_i is the number of uncensored scores equal to a_i
- r_i is the number of censored and uncensored scores greater than or equal to a_i

$$T_{PLE} = \sum_{i=1}^m (a_i - a_{i-1}) \prod_{j=0}^i (1 - d_j/r_j) \quad (1)$$

If a batsmen is dismissed for 0 runs $a_i = 0$ but if the batter remains not out and on zero then the T_{PLE} will increase slightly in an affect to reward the batter for surviving. So we can see the practical application of the PLE already in use in terms of a batting average.

2.3 Other Performance Measures

Van Staden developed a measure to determine the successfulness of a batter. This measure is known as the survival rate [Van Staden, 2009]. In essence the survival rate is the number of total balls faced divided by the number of completed innings. The greater the survival function the longer periods of time a batter can bat which is an important statistic when it comes to the format of first class cricket which is usually played over 4 or 5 days

and batting over long periods of time is required. Also during different circumstances like when the ball is new and swinging or when it gets about 40 overs old and starts reverse swinging. This survival rate shows the batting mental of the batter and their ability to apply himself/herself. If we refer to Table 1 we see the career summary statistics of Rahul Dravid. Rahul Dravid is a retired Indian cricketer and he had the ability to occupy the batting crease for many hours and he currently holds the record for facing the most balls in Test cricket. If we examine his survival rate it would be $31258/286 = 109.29$. Therefore, he would essentially face more than 100 balls per completed batting innings on average which is approximately 18 overs just by himself, assuming that on average he would face three out of six balls per over.

Table 2 shows the career batting summary of South African international player Faf du Plessis. By analysing his batting records in particular his T20I statistics. In T20 cricket the general trend is that batsmen who have very high strike rates do not necessarily have high survival rates and those who have high survival rates do not particularly score runs quickly. In terms of Faf du Plessis's strike rate he scores 134 runs per 100 balls faced and his survival rate is $1137/50 = 22.74$ balls which equates to almost 8 overs faced. If we compare this to the T20I batting summary statistics of David Miller (Table 3) who has a similar average to that of Faf du Plessis, we see he has a higher strike rate (140.62) than Faf du Plessis but a much lower survival rate of $1270/83 = 15.30$.

Table 1: Test and ODI career records of Rahul Dravid.

Format	Matches	Innings	Not Outs	Runs	High Score	Average	Balls Faced	Strike Rate
Test	164	286	32	13288	270	52.31	31258	42.51
ODI	344	318	40	10889	153	39.16	15285	71.23

Table 2: Test, ODI and T20 career records of Faf du Plessis.

Format	Matches	Innings	Nut Outs	Runs	High Score	Average	Balls Faced	Strike Rate
Test	69	118	14	4163	199	40.02	8986	46.32
ODI	143	136	20	5507	185	47.47	6215	88.60
T20I	50	50	7	1528	119	35.53	1137	134.38

Table 3: ODI and T20 career records of David Miller.

Format	Matches	Innings	Not Outs	Runs	High Score	Average	Balls Faced	Strike Rate
ODI	143	122	36	3503	139	40.73	3467	101.03
T20I	95	83	27	1786	101	31.89	1270	140.62

Along with the batting performance measures there are also equally important bowling performance measures [Van Staden, 2017]. The bowling strike rate is given as the number of balls bowled divided by number of wickets taken, bowling average which takes the number of runs conceded by the bowler and divides it by the number of wickets taken. The economy rate of the bowler is based on the number of runs conceded by the bowler per over bowled.

Further statistical work based on bowling and batting metrics have been done to an extent of quantifying an individuals performance inside of a team sport. [Shah, 2017] developed an index in terms of batsmen-bowler match ups. He suggests that a batsman's performance should be judged according to the 'quality' of runs scored rather than the quantity. So in essence certain bowlers statistically are better than others in terms of their wicket taking ability and skill sets which is determined by their bowler rankings. Batsmen should be judged differently based on their performance against the quality of different bowlers. One of the indices mentioned by Shah is the performance index of a batsmen (PIB) which takes the batting average of batsmen x against bowler y divided by the career bowling average of bowler y multiply by a hundred. What this entails is that if bowler y is a good/quality bowler then their career bowling average will be low and in the event that the PIB index increases it indicates that the batsmen scored his runs against a quality bowler.

3 Methodology

3.1 Exposition of the Kaplan-Meier estimate

The Kaplan-Meier estimator is a non parametric estimator, which implies that the data does not follow any known distribution. The whole point of using the Kaplan-Meier estimator is to develop survival curves and functions from them to be used for inferences and analysis. A survival function in its essence shows you the probability of some event or situation surviving beyond a certain time and how the probability of survival changes the longer time goes on. The data type used in the Kaplan-Meier estimator is discrete data.

We will explain the different components and factors of the Kaplan-Meier estimator by means of an example in which we look at death from a cancer after exposure to a particular carcinogen which was measured in two groups of rats. The reason we chose this example to explain the Kaplan-Meier estimator is that the estimator had originally been developed for the medical industry.

The survival rate can be expressed as the survivor function $S(t)$:

$$S(t) = (\text{no. of rats surviving longer than time } t) / (\text{total sample size of rats})$$

The Kaplan-Meier estimator is the same as the product limit estimator (PLE), both can be used interchangeably, which can be expressed as:

$$\hat{S}(t) = \prod_{t_i < t} \left[1 - \frac{d_i}{n_i} \right] \quad (2)$$

where:

1. t_i is the value at point i
2. d_i is the number of failures/deaths up to point i
3. n_i is the number of values just before t_i

The Kaplan-Meier estimator allows you to create survival functions as well as a instantaneous hazard function or hazard rate. What the hazard rate $h(t)$ quantifies is the number of times one would expect to see a failure within a given time period. The hazard function is viewed as a rate rather than a probability since it uses units of time as $\frac{1}{t}$. We used the cumulative hazard function $\hat{H}(t)$ in the analysis. The benefit of using cumulative survival/hazard functions is that they are designed to illustrate the event of a risk and also adverse effects as mentioned by [Ludbrook and Royse, 2008].

$$\hat{H}(t) = -\ln(\hat{S}_t) - \ln(\hat{S}_t) = \sum_{i=1}^t (-\ln(1 - \hat{h}_i)) \approx \sum_{i=1}^t \hat{h}_i = \hat{H}(t) \quad (3)$$

The PLE follows certain assumptions:

1. Any censored² data has the same survival chances as those who continue to follow
2. Survival probabilities are the same for subjects that join early or late in the study
3. The event happens at the specified time

We use the variance of the survival analysis as a part of our analysis. We use the equation developed by [Greenwood et al., 1926] to solve the variance of our survival function and as a by-product of the survival function we are able to determine the variance of the hazard function

The variance of the survival function is:

$$var(\hat{S}_t) = \hat{S}_t^2 \sum_{t_i < t} \frac{d_i}{n_i(n_i - d_i)} \quad (4)$$

The variance of the hazard function is:

$$var(\hat{H}_t) = \frac{var(\hat{S}_t)}{\hat{S}_t^2} \quad (5)$$

An important statistical metric of the Kaplan-Meier estimator is the mean and median of the estimator. The median is calculated by taking the smallest survival time where the survival function is less than 0.5. Once we have worked out the point estimates of our data we are able to accompany them by making use of confidence intervals developed by [Brookmeyer and Crowley, 1982], where they specifically look at the confidence interval of the median in survival functions.

The equation for the confidence interval of the median developed by [Brookmeyer and Crowley, 1982] is:

$$R_\alpha = \left\{ m | \hat{S}(m) - \frac{1}{2} \right\}^2 \leq c_\alpha (\hat{S}(m))^2 \sum_{x_i \leq m} \frac{d_i}{N_x(X_i)[N_x(X_i + d_i)]} \quad (6)$$

where:

- m = the median value
- c_α is such that $p(\chi(1) > c_\alpha) = \alpha$
- $N_x(X_i)$ the number of patients with observed survival times larger than X_i
- d_i is the number of observed deaths

²Refers to incomplete data whereby the event has been recorded properly or in our context the rat has survived at the end of the study implying no event

We make use of a non parametric test called the sign test, more specifically a generalised sign test, where we test censored data. This is done by taking a sample size of size n and we use the hypothesis-testing problem with

H_0 : median of survival function, $S^0(t) = M$

H_1 : median of survival function $S^0(t) \neq M$

where:

- $S^0(t)$ is the survival function.
- M is the median

We use a specific hypothesis test tailored for survival analysis when testing for survival curves and whether they are identical or not. This hypothesis test is called the log-rank test [Bland and Altman, 2004]. The log-rank test follows the basic assumptions of the Kaplan-Meier estimator such as the survival probabilities are the same for subjects used whether at the beginning or towards the end in the study. The log-rank test is effective in testing a difference between groups but is very unlikely to test whether survival curves intersect or cross each other.

We will now look at a practical example of the application of the Kaplan-Meier estimator along with the survival and hazard functions that accompany it.

The data used in this particular example is based on 2 sample groups of rats where we investigate the death of rats from a cancer after exposure to a particular carcinogen. The first group of rats followed a different pre-treatment routine compared to the second group.

According to group 1 and group 2 seen in tables 4 and 5 ³ we see several variables but the most important variables of note is S =survival probability, $SE(S)$ = the standard error of the survival function probabilities, H = Hazard rate and $SE(H)$ = Standard error of the hazard rate ⁴.

³https://www.statsdirect.com/help/survival_analysis/kaplan_meier.htm

⁴https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_survival/bs704_survival4.html

Table 4: Group 1 data.

Group: 1 (Group Surv = 2)

<u>Time</u>	<u>At risk</u>	<u>Dead</u>	<u>Censored</u>	<u>S</u>	<u>SE(S)</u>	<u>H</u>	<u>SE(H)</u>
142	22	1	0	0.954545	0.044409	0.04652	0.046524
157	21	1	0	0.909091	0.061291	0.09531	0.06742
163	20	1	0	0.863636	0.073165	0.146603	0.084717
198	19	1	0	0.818182	0.08223	0.200671	0.100504
204	18	0	1	0.818182	0.08223	0.200671	0.100504
205	17	1	0	0.770053	0.090387	0.261295	0.117378
232	16	3	0	0.625668	0.105069	0.468935	0.16793
233	13	4	0	0.433155	0.108192	0.836659	0.249777
239	9	1	0	0.385027	0.106338	0.954442	0.276184
240	8	1	0	0.336898	0.103365	1.087974	0.306814
261	7	1	0	0.28877	0.099172	1.242125	0.34343
280	6	2	0	0.192513	0.086369	1.64759	0.44864
295	4	2	0	0.096257	0.064663	2.340737	0.671772
323	2	1	0	0.048128	0.046941	3.033884	0.975335
344	1	0	1	0.048128	0.046941	3.033884	0.975335

Table 5: Group 2 data.

Group: 2 (Group Surv = 1)

<u>Time</u>	<u>At risk</u>	<u>Dead</u>	<u>Censored</u>	<u>S</u>	<u>SE(S)</u>	<u>H</u>	<u>SE(H)</u>
143	19	1	0	0.947368	0.051228	0.054067	0.054074
165	18	1	0	0.894737	0.070406	0.111226	0.078689
188	17	2	0	0.789474	0.093529	0.236389	0.11847
190	15	1	0	0.736842	0.101023	0.305382	0.137102
192	14	1	0	0.684211	0.106639	0.37949	0.155857
206	13	1	0	0.631579	0.110665	0.459532	0.175219
208	12	1	0	0.578947	0.113269	0.546544	0.195646
212	11	1	0	0.526316	0.114549	0.641854	0.217643
216	10	1	1	0.473684	0.114549	0.747214	0.241825
220	8	1	0	0.414474	0.114515	0.880746	0.276291
227	7	1	0	0.355263	0.112426	1.034896	0.316459
230	6	1	0	0.296053	0.108162	1.217218	0.365349
235	5	1	0	0.236842	0.10145	1.440362	0.428345
244	4	0	1	0.236842	0.10145	1.440362	0.428345
246	3	1	0	0.157895	0.093431	1.845827	0.591732
265	2	1	0	0.078947	0.072792	2.538974	0.922034
303	1	1	0	0	*	infinity	*

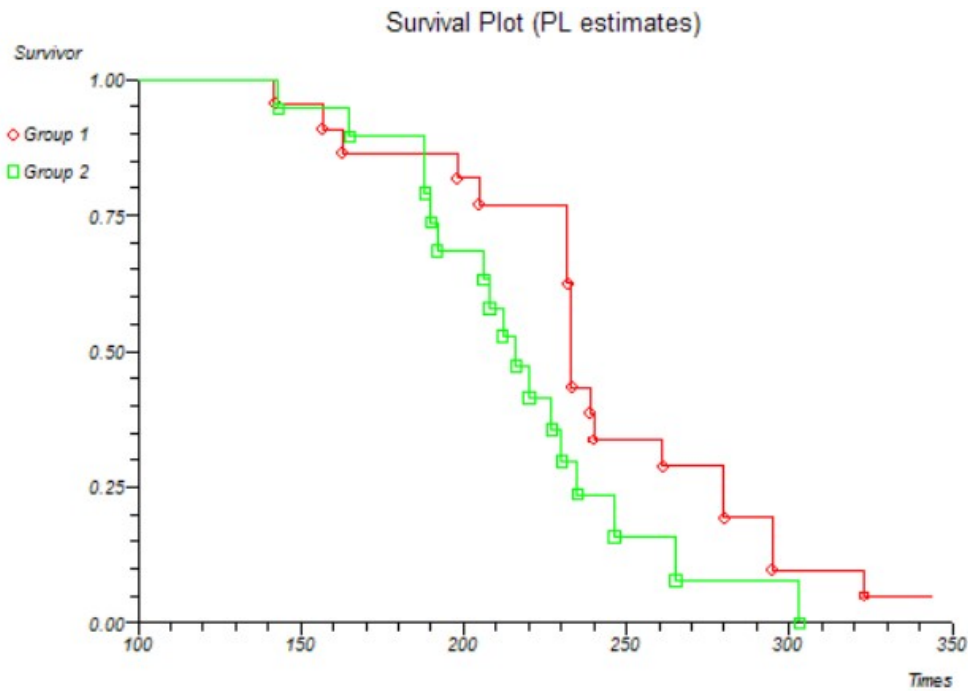


Figure 1: Survival function

By analysing Figure 1 we can see how the survival function works by viewing how the survival chances of the rats for the 2 groups decrease the longer time goes on. We can also observe from the graph that group 1 has a slightly better survival function in terms of the rats of group 1 surviving for longer periods of time.

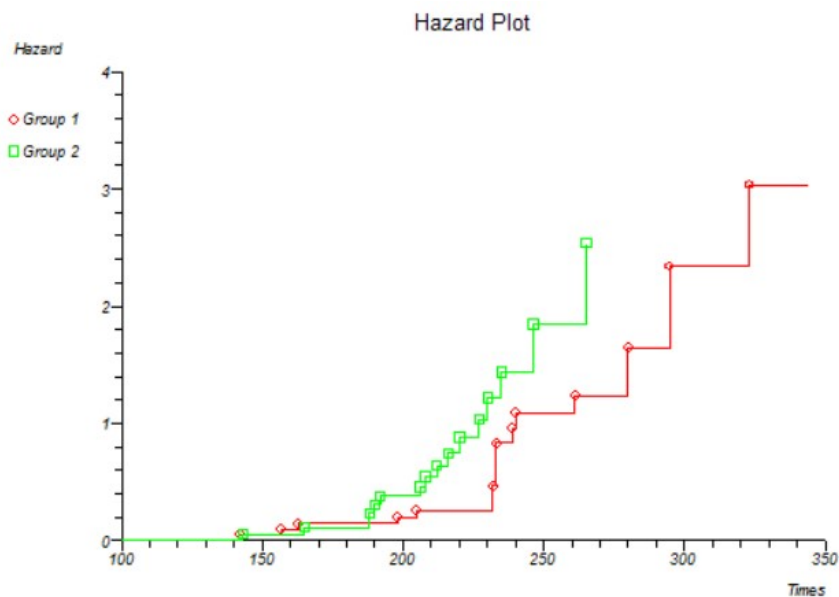


Figure 2: Hazard plot

The hazard function models the periods which have the highest chance of an event occurring and due to the analysis of the survival function our decision to say that the group 1 of rats has a better or higher chance of survival is further reinforced by the hazard plot according to Figure 2 that shows us this.

3.2 Adaptation of Kaplan-Meier estimator in batting scorecards

The Kaplan-Meier estimator can be adapted into a sporting context and specifically for our investigation of developing a survival function of batsmen in international cricket, in particular T20 cricket. We are able to do this by viewing a batter's innings as a lifespan, so your time starts the moment the batter steps out to bat and start the innings, the batter 'live' for a certain amount of runs scored and then 'die' when dismissed. The PLE is particularly effective as it allows for censored and uncensored data which we get in cricket with the data being censored data, when the batter remains not out at the end of the innings [Kachoyan and West, 2018]. Uncensored data is when the batter's innings results in them being dismissed.

In terms of developing the survival function for a batter we use the following equation.

$$\hat{S}(t) = \prod_{t_i < t} \left[1 - \frac{d_i}{n_i} \right] \quad (7)$$

where:

1. t_i is the number of runs scored at a certain time point
2. d_i is the number of runs scored at the point the batter gets dismissed
3. n_i is the number of runs just before t_i

The equations of the hazard function and cumulative hazard function are as follows:

$$h(t) = -\frac{d}{dt} \ln(S(t)) \quad (8)$$

$$H(t) = -\ln(S(t)) \quad (9)$$

In terms of using the log-rank test in a cricket context a crucial assumption used is that the survival probability is the same for a player whether its early in the career of innings played or even after a significant amount of time/innings played.

4 Application

The world of cricket is on the rise in terms of a commercial view point with the inception of short forms of cricket such as the Hundred in England and more recently in South African the Mzansi Super League. Our application looks at the short form analysis of cricket players from different countries and their survival analysis in different aspects.

The data being analysed is based on ten cricketers listed in table , each of the players come from the top ten ranked T20 cricket playing nations as on 31 July 2022. The data collect was also up to 31 July 2022.

Cricket, being an unpredictable sport based on the fact that we do not know how a player will actually perform on a particular day as well as factors like players going through a 'purple patch' which means that they play exceptionally better than they usually do, whether it be in batting or bowling. We decided to assume that these factors are negligible in our survival analysis in order to present an unbiased view of the analysis.

We look at three different applications of the survival analysis namely the standard survival curve, the survival analysis for the first innings against second innings and the survival analysis of the batting position of batters. The hazard function is discussed in this section along with some hypothesis testing of the curves.

Table 6: Variables Summary.

Player Name	Initials	Country
Virat Kohli	VK	India
Babar Azam	BA	Pakistan
Dinesh Chandimal	DC	Sri Lanka
Jos Buttler	JB	England
Kane Williamson	KW	New Zealand
Nicholas Pooran	NP	West Indies
Quinton de Kock	QDK	South Africa
Shakib al Hasan	SAH	Bangladesh
Aaron Finch	AF	Australia
Mohammed Nabi	MN	Afganistan

4.1 Standard survival analysis

In this section we analyze the complete survival curve of batters. We looked for batters that fall in to the top and middle order batting positions in a team, top order batters bat in positions 1-3 and middle order batters fall into positions 4-7.

Our sample sizes considered are batters that have played more than 40 T20 international games. Our reason for choosing this sample size is because batters with less games played often give us very rigid survival curves which makes the analysis very difficult and unreliable to get an appropriate result due to the lack of games played .

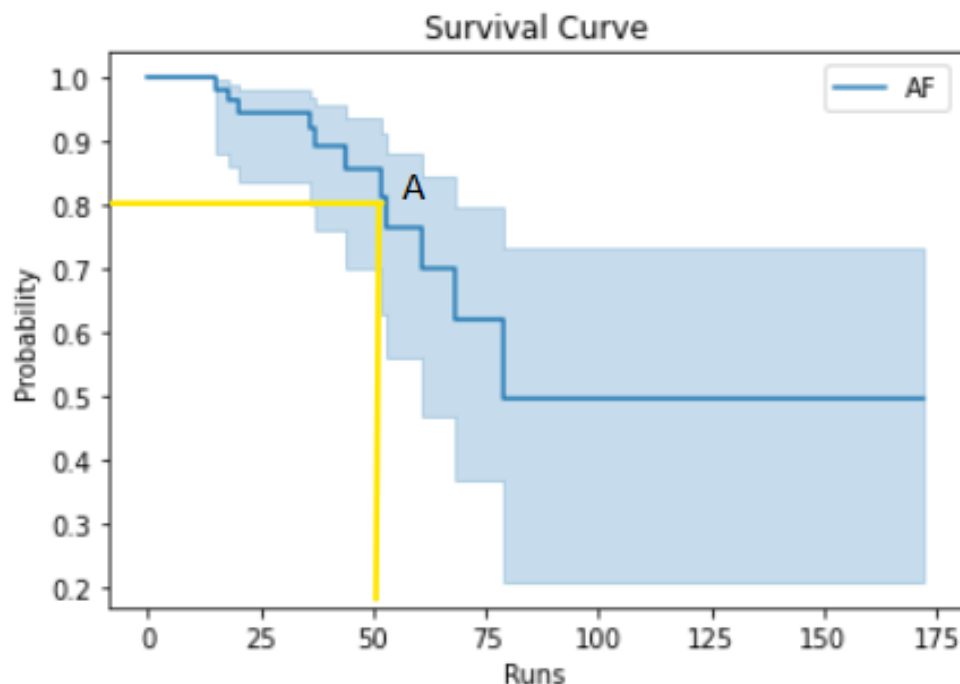


Figure 3: Aaron Finch Survival Analysis

Figure 3 shows us the full survival analysis of the T20 career of Aaron Finch from Australia. The shaded parts are the 95% confidence interval of his probability of not being dismissed. If we look carefully at point A on the graph we can see that he currently has a probability of 80% to reach a score of around 50 runs scored before he even starts his batting innings.

Furthermore his probability of getting out in a game never nears zero owing to the fact that when he scored his highest T20 score of around 175 runs he ended up being not out when he reached that milestone. So since the survival curve views that innings as censored data it is unable to lower his probability of being dismissed passed 175 runs because he has not been dismissed after reaching his highest score.

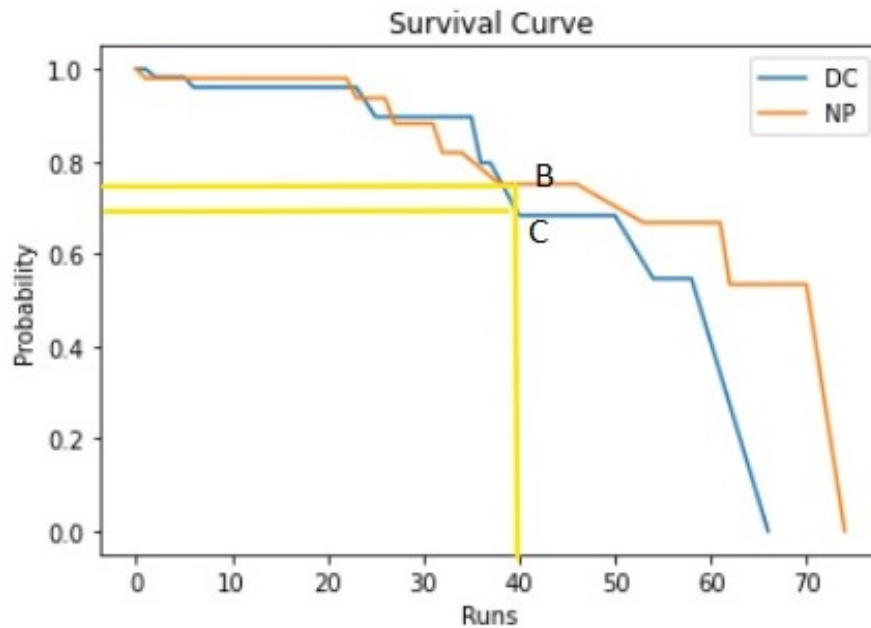


Figure 4: Dinesh Chandimal and Nicholas Pooran Survival Analysis

We now look at a multiple survival analysis in Figure 4 between Sri Lankan player Dinesh Chandimal and West Indian Nicholas Pooran. According to their survival analysis the probability of them not being dismissed is different when they reach 40 runs scored. Nicholas Pooran has a probability of 0.75 or a 75% chance of not being dismissed when he scores 40 runs whereas Dinesh Chandimal has a probability of around 0.70 or 70% of not being dismissed

We can view this as a relevant comparison as both players predominantly bat in the middle order for their national teams. We can assume that because they both bat in the middle order they usually come in to bat at around the same time in their batting innings.

4.2 Survival analysis of batting innings

In this section we try to analyze the survival functions of different batsmen in terms of the innings that they bat in, i.e first or second innings. The reason we use this comparison is to see which innings batters prefer to bat in or are able to make more of an impact in a game.

Cricket teams or scouts can use this type of analysis to put together a team that fits a particular profile for a team. If a batsman has a high survival rate batting first this implies when setting a score their probability of getting out is low which leads to them batting for over long periods of time. This can be seen as an anchor role, where the batter playing the anchor role does not necessarily score runs quickly but is able to stay in whilst other batters bat around him.

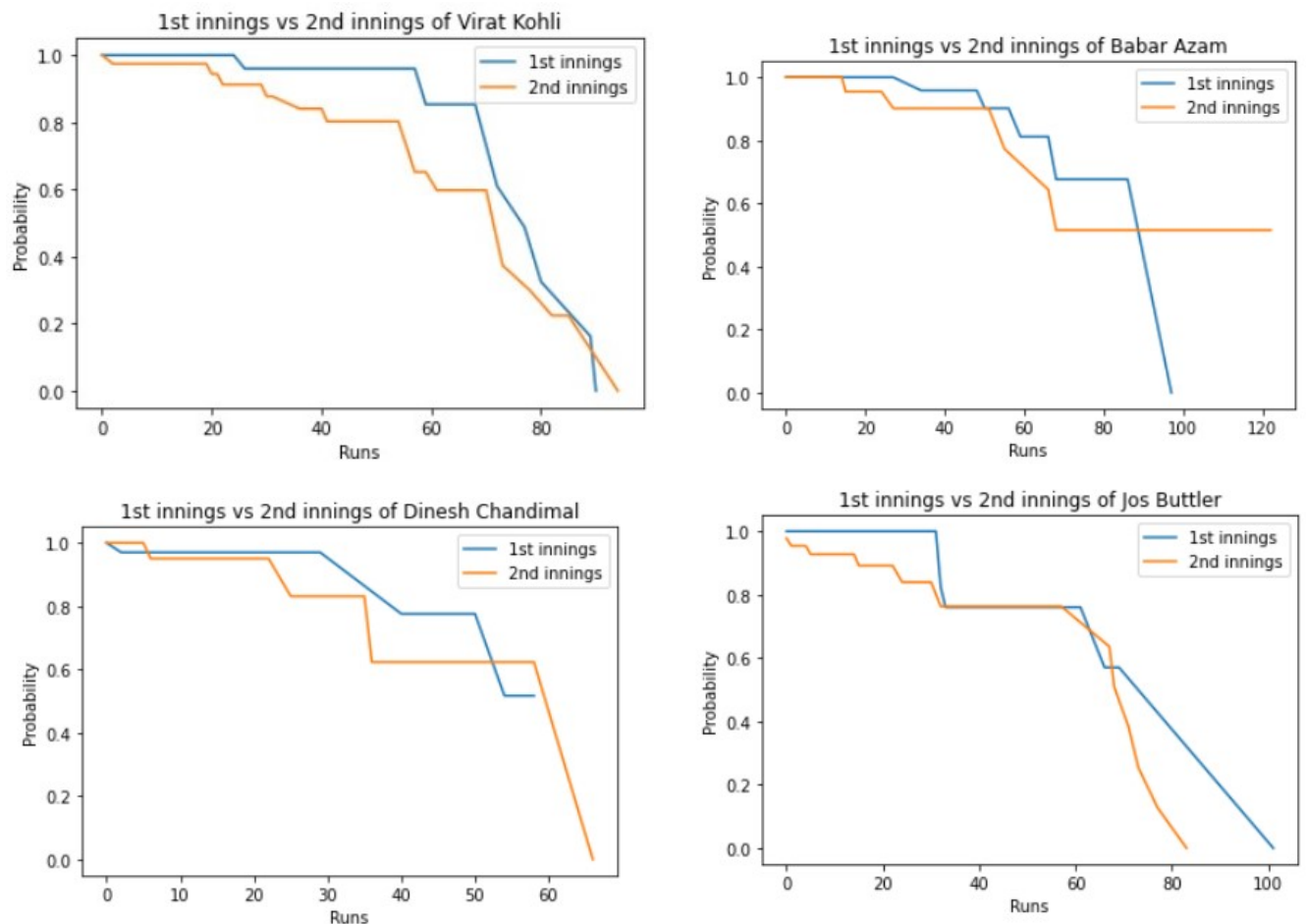


Figure 5: 1st innings vs 2nd innings of different batters

We look at the survival curves of four different batters in Figure 5 based on their first and second innings and we from the onset notice a clear preference or trend based on these survival curves.

We notice that most of the above mentioned batters have a higher survival rate when batting first. This does not necessarily mean that these batters all serve as an anchor role in their team. For example, Jos Buttler has a strike rate of around 143.58 in T20 international cricket which is exceptionally high as it indicates for every 100 balls faced he scores approximately 144 runs.

We can relate these observations to some psychology of cricket as when setting a score or batting first there is not that additional pressure of having a score to worry about chasing down, or factors such as the run rate required to reach the total score or the pressure of dealing with a relatively quiet over in terms of not a lot of runs being scored. So due to some of these factors of batting in the second innings we can understand that most batters prefer batting in the first innings whereas compared to the second innings.

The exception in these graphs is that of Babar Azam as he has a much better second innings survival function compared to his first innings. The reason his second innings survival curve is better than his first innings survival

curve is down to the fact that he scored his highest T20 score in his second innings and remained not out. This heavily influenced his survival curve as from 0 to around 90 runs scored his survival curve for his first innings was better than that of his second innings.

4.3 Survival analysis of batting position

This section makes use of another interpretation of the survival analysis in terms of the batting position of batsmen. Batters usually fluctuate around several batting positions for reasons such as form. In the event that they are out of form they usually drop down the batting order. This usually happens until they find a position that suits them appropriately in terms of their role in the team and as well as their ability. The batting position also depends on match circumstances.

When looking at the different batting positions of the batsmen we chose positions where they had played more than 20 innings in a particular batting position in order to get a readable interpretation of the survival functions.

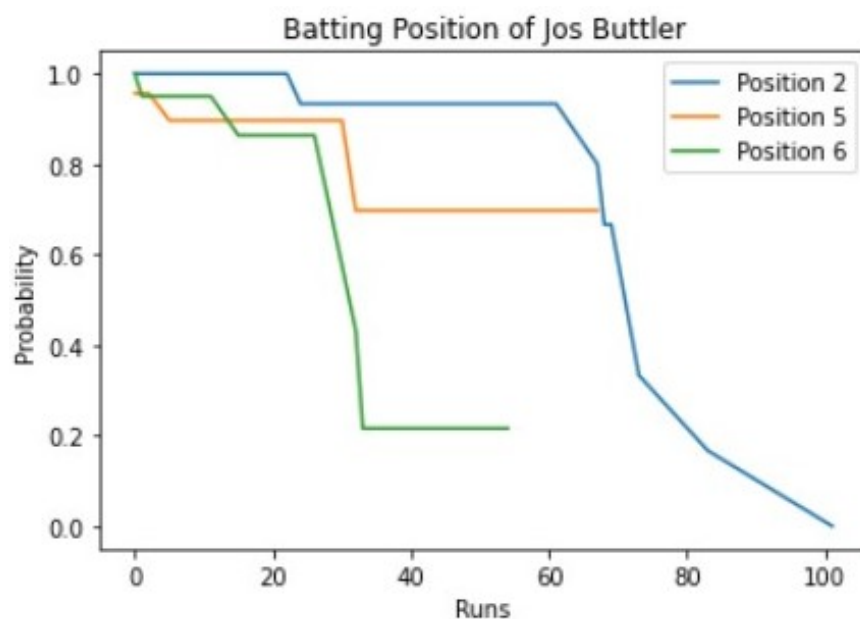


Figure 6: Survival analysis of Jos Buttler's batting positions

The first batsmen we look at is Jos Buttler in Figure 6. Jos Buttler started his career as a middle order batsman and now plays predominantly as a top order batsman and this decision appears to be justified according to his survival function.

When he opens the batting he has a better survival rate and scores more runs compared to when he bats at position 5 and 6. We can make the assumption that when he bats at position 5 and 6 he comes in to bat when there are very few overs left so he has a lot less time to get himself into his innings and probably has to play a lot more aggressively which causes him to get dismissed more often for few runs as when compared to when he

opens the batting.

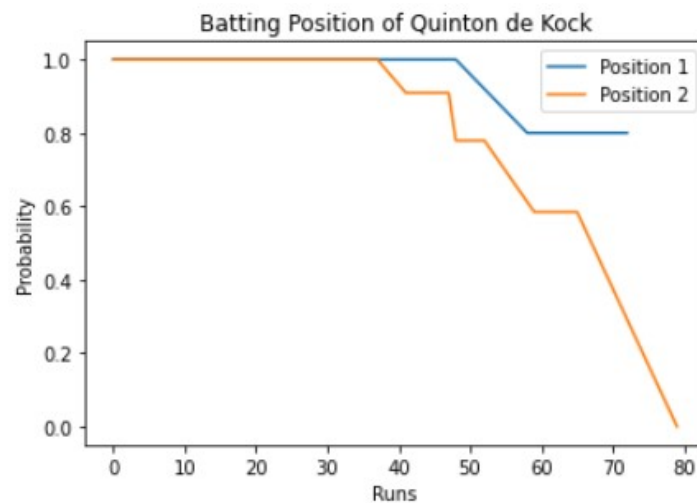


Figure 7: Survival analysis of Quinton de Kock batting positions

Figure 7 looks at South Africa's Quinton de Kock's positional survival functions. He bats mainly in the top order at position 1 or 2 and according to his survival function he has a better survival function when he bats at position 1 compared to position 2. Although he has an extremely good survival function when he faces the first ball he ends up scoring more runs but with a higher probability of getting out when he does not face the first ball.

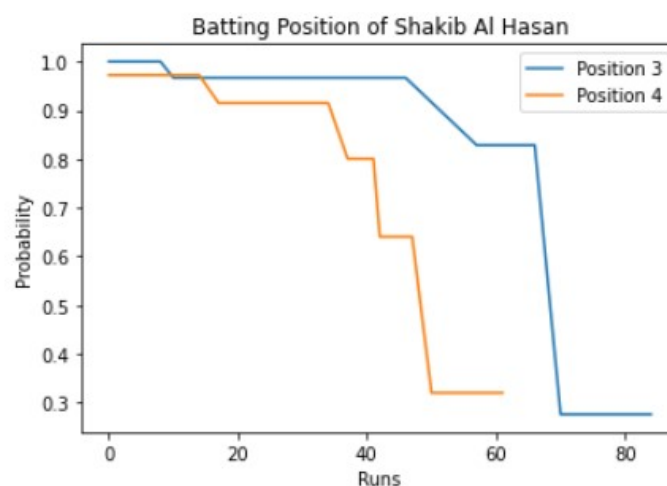


Figure 8: Survival analysis of Shakib al Hasan batting positions

Figure 8 looks at Shakib Al Hasan's positional survival functions. He has 2 positions that he bats in the majority of the time which is position 3 and 4. According to his survival function he has a better survival curve when he bats at position 3 compared to position 4. We can deduce that when he bats higher up the order he is able to take more time to get himself in hence have a better survival curve whilst also scoring more runs when he bats higher.

4.4 Nelson Aalen - Hazard function

We use a Nelson Aalen estimate in this section to develop a cumulative hazard function for our data. The hazard rate is conditional on a batsman actually surviving until a certain amount of runs scored. The hazard rate is a function of runs scored being $\frac{1}{Runs}$.

If we take the reciprocal of the hazard rate being a function of time it gives us the expected number of failures for that period which is what we are interested in as these hazard functions will tell us how many times do we expect a batsmen to fail before they reach or score a certain amount of runs.

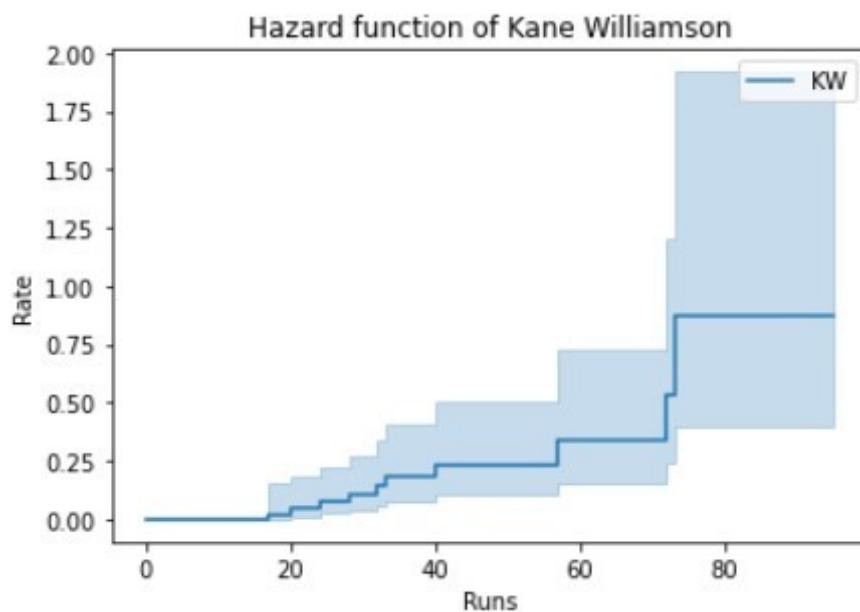


Figure 9: Hazard function with 1 batsmen

The hazard function of Kane Williamson is shown in Figure 9. This curve tells us that we expect Kane Williamson in 1 out of every 2 innings played to score approximately 70 runs which is where rate is equal to 1, when he does score around 70 runs his confidence interval for the failure ranges between 0.5 and 2 given a 95% survival interval.

Scoring 70 runs in 1 out of every 2 innings is a phenomenal return for any batsmen in T20 cricket but this is not particularly accurate as the estimate also takes censored data into account meaning he remains not out at the end of an innings. So taking that into account also affects the value and interpretation of the estimate.

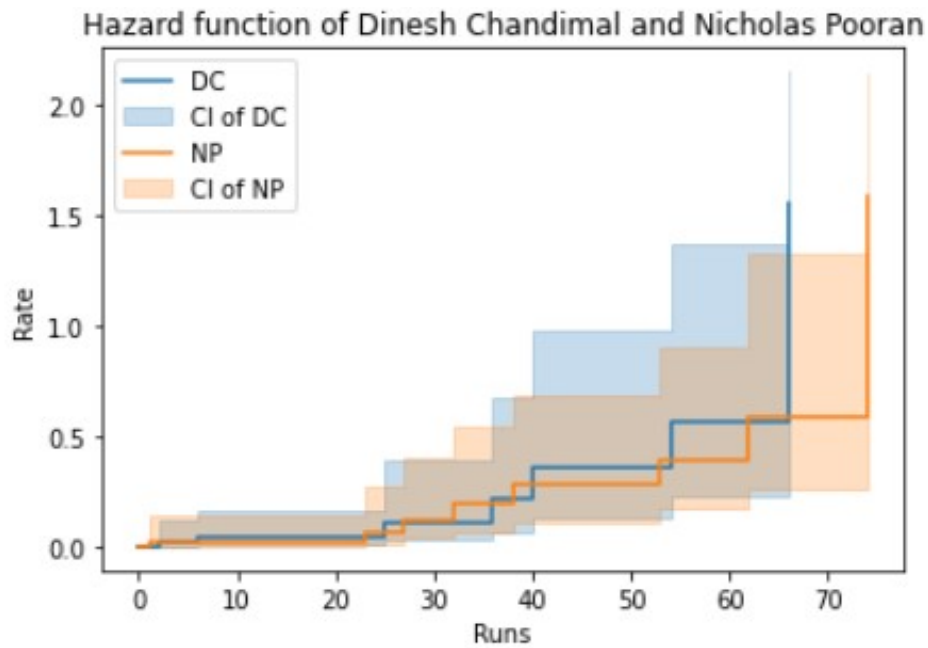


Figure 10: Hazard function with 2 batsmen

Figure 10 analyzes the hazard functions of Dinesh Chandimal and Nicholas Pooran with a 95% confidence interval. Their hazard functions remain very similar in structure barring the main difference being the highest number of runs scored which was by Nicholas Pooran which increased the length of his hazard function.

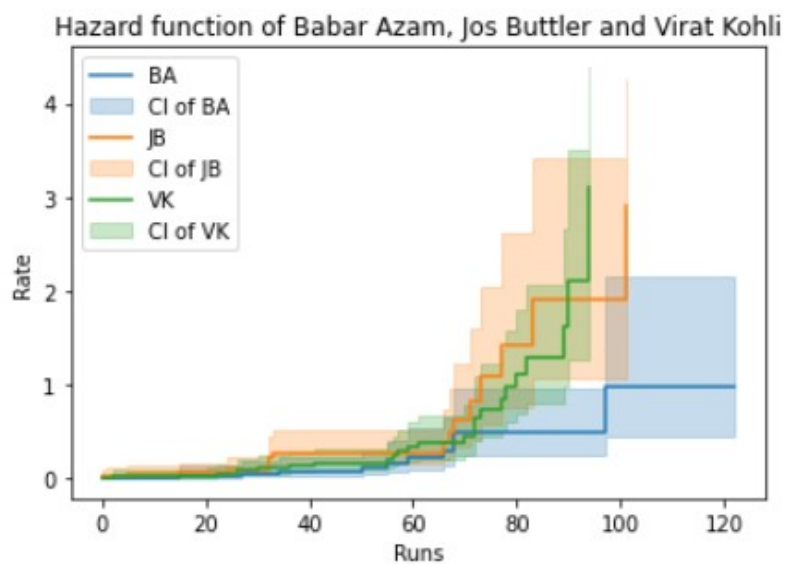


Figure 11: Hazard function with 3 batsmen

The cumulative hazard functions of Babar Azam, Jos Buttler and Virat Kohli are analyzed in Figure 11 with their 95% confidence intervals. It is of interest to analyze these 3 batsmen in particular as they are seen as the main players for each of their national teams.

Referring to their cumulative hazard functions, Babar Azam appears to be the player with the best hazard function. His hazard function depicts a relatively long and horizontal curve indicating he scores a lot of runs in each innings with a relatively low failure rate which just creeps over 1 compared to that of Jos Buttler and Virat Kohli.

4.5 Hypothesis testing

This section discusses the hypothesis testing of the different survival curves of the batsmen. We make use of the log rank test designed specifically for survival analysis that compares survival distributions of two samples. We test the null hypothesis (survival curves are the same) against the alternative hypothesis (survival curves are different). The hypothesis test statistic used is a chi-square test statistic.

In Figure 12 we look at the first innings versus second innings hypothesis test of Shakib Al Hasan to test whether the two survival curves are the same .

H_0 = The 2 populations have an identical distribution

H_A = The 2 populations do not have an identical distribution

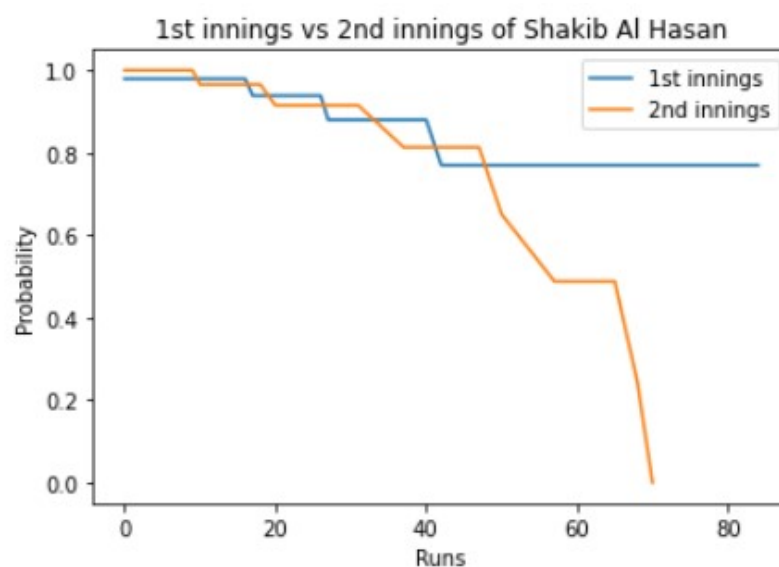


Figure 12: Hypothesis testing for Shakib Al Hasan for his first innings vs second innings

We used the lifelines package in python to determine the log rank test.

t_0	-1
null_distribution	chi squared
degrees_of_freedom	1
event_observed_b	1 0.0 4 0.0 6 0.0 7 0.0 8 ...
test_name	logrank_test
test_statistic	p -log2(p)
0	44.38 <0.005 35.11

Figure 13: Hypothesis test output for Shakib Al Hasan

According to the output produced in Python given in Figure 13 we see that $\chi^2 = 44.38$ and p-value < 0.005

So at a 1% level of significance we reject our null hypothesis and can conclude that the two distributions are not identical

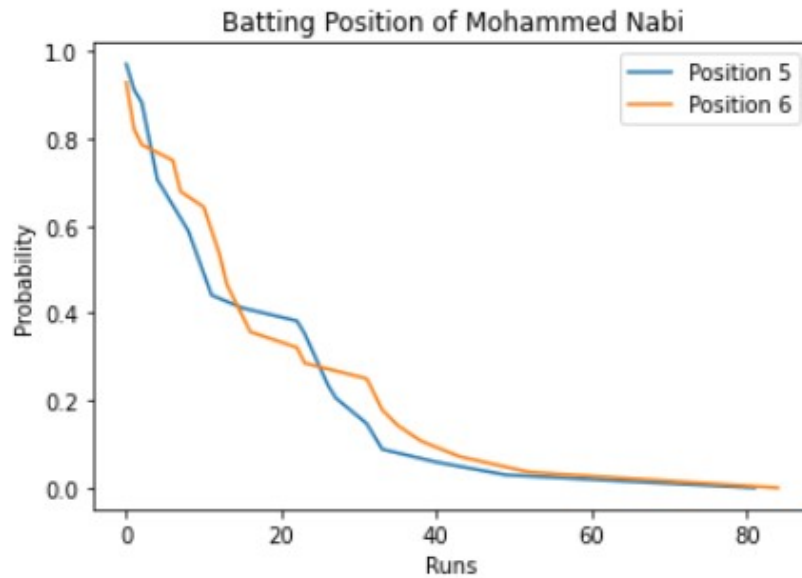


Figure 14: Hypothesis testing for Mohammed Nabi's batting position

Figure 14 analyses the batting position of Mohammed Nabi and looks at his survival function when he bats at positions 5 and 6. We conduct our hypothesis test to test whether the survival distributions are identical.

H_0 = The 2 populations have an identical distribution

H_A = The 2 populations do not have an identical distribution

t_0	-1
null_distribution	chi squared
degrees_of_freedom	1
event_observed_b	1 1.0 2 1.0 3 1.0 6 1.0 9 ...
test_name	logrank_test
test_statistic	p -log2(p)
0	0.35 0.55 0.85

Figure 15: Hypothesis testing output for Mohammed Nabi's

According to the output in Figure 15 produced by Python we can observe that $\chi^2 = 0.35$ with p-value = 0.55 we do not reject our null hypothesis and we can conclude that the 2 distributions are not significantly different.

5 Conclusion

In summary we have adapted the i.e Kaplan Meier estimator to determine a survival analysis for batsmen in international cricket. This analysis can be used for all types of players provided we have access to their runs scored and whether they remained out or not out.

We have showed that we can adapt the Kaplan-Meier estimator to not only the career of a batsman but to also show within different aspects of the cricket game how effective they can be in "surviving". Some of these aspects include which batting positions they survive the longest in or which innings they prefer batting in terms of surviving.

Recommendations for further research

We have a few recommendations for future research do be done in terms of survival function in cricket.

1. To approximate the survival function to other known distributions such as a Weibull distribution, since it meets the requirements of being approximated .
2. To adapt the Kaplan-Meier estimator as a function of strike rate. Strike rate is a measure of how quickly batsmen score their runs or can be viewed as runs scored per 100 balls faced. So constucting a survival analysis in terms of strike rate we can investigate which types of batsmen score their runs quickly with a high survival rate which is vital especially in T20 cricket

References

- [Bland and Altman, 2004] Bland, J. M. and Altman, D. G. (2004). The logrank test. *BMJ*, 328(7447):1073.
- [Brookmeyer and Crowley, 1982] Brookmeyer, R. and Crowley, J. (1982). A confidence interval for the median survival time. *Biometrics*, 38:29–41.
- [Danaher, 1989] Danaher, P. J. (1989). Estimating a cricketer's batting average using the product limit estimator. *The New Zealand Statistician*, 24(1):2–5.
- [Greenwood et al., 1926] Greenwood, M. et al. (1926). A report on the natural duration of cancer. *Reports on Public Health and Medical Subjects*, 33:1–26.
- [Kachoyan and West, 2016] Kachoyan, B. and West, M. (2016). Cricket as life and death. In *13th Australasian Conference on Mathematics and Computers in Sport, Melbourne*.
- [Kachoyan and West, 2018] Kachoyan, B. and West, M. (2018). Deriving an exact batting survival function in cricket. In *14th Australasian Conference on Mathematics and Computers in Sport (ANZIAM MathSport 2018). University of Sunshine Coast, ANZIAM Mathsport*, pages 160–172.
- [Kaplan and Meier, 1958] Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.
- [Kimber and Hansford, 1993] Kimber, A. C. and Hansford, A. R. (1993). A statistical analysis of batting in cricket. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 156(3):443–455.
- [Ludbrook and Royse, 2008] Ludbrook, J. and Royse, A. G. (2008). Analysing clinical studies: Principles, practice and pitfalls of Kaplan–Meier plots. *ANZ Journal of Surgery*, 78(3):204–210.
- [Shah, 2017] Shah, P. (2017). New performance measure in cricket. *ISOR Journal of Sports and Physical Education*, 4(3):28–30.
- [Van Staden, 2009] Van Staden, P. J. (2009). Comparison of cricketers' bowling and batting performances using graphical displays. *Current Science*, 96(6):764–766.
- [Van Staden, 2017] Van Staden, P. J. (2017). Cricket. *The Oxford Anthology of Statistics in Sports*., 1.

Appendix A

The link to the Python code and output used can be found and the following github link [GITHUB-code](#)