Survival ability of Indian and overseas batsmen on the cricket pitch in Indian premier league

Abstract
Twenty20 format of cricket is a fast track ball game compared to the other formats of cricket viz. Test and One-day International (50-over a side). The Indian Premier League (IPL) is a national franchise based Twenty20 cricket tournament initiated by Board of Control for Cricket in India (BCCI). In this format, each batsman tries to score maximum runs in minimum balls. This fact increases the probability of dismissal of a batsman. As fall of wickets leads to the loss of resources of the batting side, so it has an impact on the result of the game. This study tries to examine the survival ability of Indian and overseas batsmen in IPL 2012 season using a probabilistic model. The proposed model can be used to forecast the survival rate of the batsmen on the pitch in other format of cricket also, while the game is in progress. The findings of the study can be used to arrange the batting order of a team in Twenty20 cricket based on the match situation.

Keywords: cricket, probability, survival analysis, sport

Introduction
Cricket is an outdoor game played between two teams of eleven (11) players each in a circular ground. It is administered by certain rules and regulations, where the interaction between bat and ball takes place on a 22-yard hard surface in the middle of a circular ground called the cricket pitch. Unlike other sports, there are different versions of cricket. The different versions of cricket can be broadly classified as unlimited overs cricket (Test matches) and limited overs cricket (One-day and Twenty20). Indian Premier League (IPL) is a national franchise based Twenty20 format of cricket league initiated by Board of Control for Cricket in India (BCCI) in 2008. In IPL, each team faces only twenty (20) overs in a match, therefore, within these limited overs every batsman tries to score maximum runs in minimum balls. In the process of scoring runs quickly, the batsmen are exposed to the risk of losing their wicket. This fact increases the probability of dismissal of a batsman. As fall of wickets leads to the loss of resources of the batting side, so it has an impact on the result of the game. However, it does not mean that stability of a batsman on the pitch would help a team to win the match. Evidently, he should have scored runs as quickly as possible. Thus, in Twenty20 cricket, one can at least measure how much time a batsman can survive or how many balls a batsman can face on the cricket pitch while batting. Therefore, the study makes an attempt to measure as well as compare the standing capability of Indian and overseas batsmen on the cricket pitch in IPL 2012 using survival analysis.

Survival analysis is defined as a set of methods for analyzing data where the outcome variable is the time until the occurrence of a particular event of interest. The event could be death due to cancer, occurrence of a disease, relief from a severe back pain, etc. Let us take an example to explain mathematical definition of survival function. Suppose the actual survival time of an individual (say) \( t \) which can be regarded as the value of a variable \( T \) (i.e. associated with the survival time). It can take any non-negative value. The different values that \( T \) can take have a probability distribution, so the variable \( T \) can be considered as a random variable. Now for the random variable \( T \), the probability distribution function of \( T \) can be defined as \( F(t) \) and it is given by

\[
F(t) = P(T < t) = \int_{0}^{t} f(x)dx
\]

(1)

Which represents the probability that the survival time is less than some value \( t \). Now the survival function is defined as the probability that the survival time is greater than or equal to \( t \). Usually, it is denoted by \( S(t) \) and given by

\[
S(t) = P(T \geq t) = 1 - P(T < t) = 1 - F(t) = 1 - \int_{0}^{t} f(x)dx
\]

(2)

Therefore, the survival function can be used to represent the probability that an individual survives from the time origin to some time beyond \( t \). The survival time or time to an event of interest can be measured in days, weeks, years, etc. in which the objects or subjects are followed over a specified period of time to pinpoint the event of interest occurs. Though its uses in medical, clinical trial, actuarial science, etc. are hefty, but still the application of survival analysis in sport (especially in cricket) is limited. A few studies have found in this regard are explicitly mentioned here. Danaher\(^4\) applied the survival analysis to find an estimate of a cricketer’s unknown batting average\(^3\) based on the product limit estimator. Similar product-limit estimation technique was adopted by Kimber and Hansford\(^6\) for assessing the batting performance of cricket batsmen based on runs scored. The product limit estimator (PLE) is a non-parametric estimator originally proposed by Kaplan and Meier\(^4\) and it is defined as -

\[
PLE = \prod_{i=1}^{n} \left(1 - \frac{d_i}{n_i} \right), t_1, t_2, t_3, \ldots t_n
\]

(3)

Where \( t_i \) be the observed times of \( n \) samples until the event of interest occurred from a given population. However, sometimes lack of information arises when observations have some information...
available for the event of interest but the information is not complete. This incomplete information is termed as censoring. If there is no censoring, \( n_i \) is the number of survivors prior to time \( t_i \). But if there is a censoring, \( n_i \) is the number of survivors minus the number of censored cases in the sample. To measure the survival capability of cricketers on the cricket pitch, censoring can be perceived in those situations when a batsman remains not-out in limited overs cricket. It can be termed as so-called right censoring data.

Following the work of Kimber and Hansford,\(^4\) product limit estimator was used by Das\(^5\) to estimate the adjusted batting average of some selected cricketers. He argues, it has been revealed from the past information that batsmen have a variable risk of getting dismissed based on their current score in the innings. Thus, he proposed to model the batsmen’s scores using generalized geometric distribution. A similar problem was also addressed by van Staden\(^6\) developing a new batting criterion named as ‘survival rate’. It is defined as the number of balls faced in all innings divided by the number of completed innings. Symbolically,

\[
SV = \frac{\text{Number of balls faced}}{\text{Number of completed innings}} = \frac{1}{n} \left( \sum_{i=1}^{n} b_i + \sum_{i=n+1}^{n+m} b_i \right)
\]

where \( b_i \) (\( i = 1, 2, \ldots, n \)) represents the number of balls faced by the batsman in \( n \) completed innings and \( b_i \) (\( i = n+1, n+2, \ldots, n+m \)) represents the number of balls faced by the batsman in \((n+m)\) not-out innings. Now on the basis of equation (2) and the batting average developed by Maini and Narayanan (2007) which is based on average exposure-to-risk (AV exposure), van Staden \( \text{et al.} \) (2010) proposed a new meaningful batting average. It is based upon exposure using survival rate and it is defined as

\[
AV_{\text{survival}} = \frac{SV}{AV_{\text{exposure}}}
\]

where

\[
AV_{\text{exposure}} = \frac{\sum_{i=1}^{n} r_i \times n_i + \sum_{i=n+1}^{n+m} r_i \times n_i}{\sum_{i=1}^{n} n_i + \sum_{i=n+1}^{n+m} n_i}
\]

Where \( r_1, r_2, \ldots, r_n \) and \( r_{n+1}, r_{n+2}, \ldots, r_{n+m} \) denote the batsman’s exposure in \( n \) completed innings and \( m \) not-out innings respectively. Here \( r_i = 1, \) if the score in \( r_i \) innings is a completed score, which means that, the exposure is one for all completed innings.

Otherwise,

\[
r_i^* = \begin{cases} b_i^*/\bar{b}, & \text{if the score is a not – out score and } b_i^* < \bar{b} \\ 1, & \text{else} \end{cases}
\]

Where \( \bar{b} \) is the average number of balls faced by a batsman in his \((n+m)\) innings and it is defined as,

\[
\bar{b} = \frac{\text{Number of balls faced}}{\text{Number of innings}} = \frac{1}{n+m} \left( \sum_{i=1}^{n} b_i + \sum_{i=n+1}^{n+m} b_i \right)
\]

At this point, it is very crucial to clarify that the word ‘survival rate’ stands here to exemplify the capability of a batsman to remain present on the cricket pitch during a match. So that there will be no confusion with the measure defined by van Staden\(^6\) (c.f. equation (4)). When we mentioned the word “survival” in terms of batsman in cricket, it could be the end of the time spent by a batsman on the pitch before dismissal. As mentioned earlier, time to an event of interest can be measured in days, weeks, etc. Similarly, a batsman’s ability to stand on the pitch in the match(es) can be measured in terms of number of balls faced by him. It can be considered as an outcome variable in so-called survival analysis. Thus, if a batsman has been facing consistently colossal number of balls in the match(es) then he would have the higher probability of survival on the pitch.

**Methodology**

The approach of this study is different than earlier studies, as it does not focus on any single performance statistics of batsman like batting average, strike rate, etc. Instead, a non-parametric estimator established by Kaplan and Meier (1958) a long back is used to calculate survival probability of batsman on the pitch based on the information of number of balls faced as well as whether the player was out or not-out in the match(es). The methodology applied here is discussed below in details. Let \( O_i = 1 \), if a batsman out in the \( i^{th} \) match and 0 otherwise, \( n_i \) be the number of batsmen survives prior to \( b_i \) balls of a match, where \( b_i \) is the observed number of balls faced by the batsmen in a match. However, as mentioned earlier, to measure the survival capability of batsmen, censoring (more specifically right censoring) can be perceived in those situation of a match when a batsman remains not-out in limited overs cricket. Therefore, if any batsman remains not-out in a limited overs cricket then \( n_i \) be the number of batsmen survives minus the number of batsmen not-out in prior to \( b_i \) balls of a match. Now the Kaplan-Meier (KM) estimator in terms of the game of cricket is defined as

\[
S(b) = \prod_{b_i \leq b \leq b_i+1} \left( 1 - \frac{O_i}{n_i} \right)
\]

Where \( b_i \leq b_2 \leq \ldots \) and \( i = 1, 2, \ldots, 120 \)

Thus, the survival probabilities of the batsmen on the cricket pitch are computed using the above defined estimator. As we know that the precision of any estimate is reflected in the standard error of the estimate. Therefore, the standard error of KM estimate is computed as an essential aid to the interpretation of estimate. In this regard, Peto \( \text{et al.} \) (1977) proposed a formula to compute standard error of \( \hat{S}(b) \) and it is defined as

\[
se \{\hat{S}(b)\} = \frac{\hat{S}(b) \sqrt{1 - \hat{S}(b)}}{\sqrt{n_i}}
\]

The expression (5) is conservative for standard error of \( \hat{S}(b) \) because the standard errors will tend to be larger than they actually ought to be. Thus, the formula proposed by Greenwood’s (1926) for standard error of \( \hat{S}(b) \) is usually recommended. The Greenwood’s (1926) standard error formula can be defined as

\[
se \{\hat{S}(b)\} = \frac{\hat{S}(b) \sum O_i}{n_i \left(n_i - O_i\right)} \left(\frac{1}{\bar{b}_i} \right)
\]

For \( b_i \leq b \leq \bar{b}_i+1 \)

Now confidence interval for the corresponding value of the
survival function can be obtained based on the above standard error of KM estimate. Finally, based on the KM estimate, a survival plot is being depicted to identify the survival ability of Indian and overseas batsmen on the cricket pitch.

**Hypothesis testing**

Hypothesis testing is the modest way of examining survival ability of overseas and Indian batsmen on the cricket pitch. It allows assessing the extent to which whether an observed set of data are consistent with a particular hypothesis or not. Here the working hypothesis is that there is no difference between the survival ability of overseas and Indian batsmen on the cricket pitch. The well-known log-rank test is used to test the working hypothesis. This is the appropriate non-parametric test to use when the right censored data are non-informative, as the case is comparable here in case of not-out batsmen. In order to apply log-rank test, the survival ability of overseas and Indian batsmen computed separately. It compares observed and expected number of event of interest from both the groups. The groups in log-rank test are labeled as overseas (coded as “1”) and Indian (coded as “2”) players. Now suppose there are \( k \) different distinct balls, \( b_1 < b_2 < \ldots < b_k \) where batsmen are out across the two groups. Let \( O_{ij} \) be the individual batsman out in overseas group and \( L_{ii} \) be batsman in Indian group. Again, suppose that \( n_i \) be the total number batsmen out in overseas group and \( n_i \) be the total number batsmen out in Indian group. Consequently, there are \( O_i = O_{i1} + O_{i2} \) batsmen out in a tournament from total of \( n = n_1 + n_2 \) batsmen, at \( b_i \) ball. Now, under the null hypothesis, it is defined as

\[
W_L = \frac{U_L^2}{V_L} \sim \chi^2
\]

(7)

Where \( U_L = \sum_{i=1}^{k} (O_{ii} - E_{ii})^2 \) and \( V_L = Var(U_L) = \sum_{i=1}^{k} v_i \)

Since the different balls are independent of one another, the term \( V_L \) (i.e. variance of \( U_L \)) is sum of the variances of \( O_{ii} \) and it is given by

\[
v_i = \frac{n_1n_2O_i(n_i - O_i)}{n_i^2(n_i - 1)}
\]

(8)

The statistics \( W_L \) summarizes the extent to which the observed number of survival balls in the two groups of batsmen deviate from those expected number of survival balls in the cricket pitch, under the null hypothesis of no group differences. The larger the value of log-rank test statistics (i.e. \( W \)), greater the evidence against the null hypothesis.

**Data consideration and analysis**

In IPL 2012, each franchise has maximum of eight overseas players per squad. However, only four of them can be played in the playing XI for each match. There were 14 league matches usually played by each of the IPL team prior to knock-out stage of the tournament. Therefore, a lot of Indian as well as overseas players’ performance (i.e. number of balls faced per match and out or not-out) can be collected from the scorecard of the matches. All this scorecard information is collected from the website www.espncricinfo.com from 2012 season of the IPL. A total of 1123 players are considered for the study, out of which 459 are overseas players and 664 are Indian players. The results of the analyses based on the methodology discussed above are provided below. From the above table, it has been observed that there are 103 and 162 cases are censored (i.e. not-out cases) corresponding to overseas and Indian players respectively. Overall 23.6% not-out case is being found for a total of 1123 cases (Table 1) (Table 2) (Table 3) (Table 4).

**Table 1** Descriptive information of players in IPL 2012

<table>
<thead>
<tr>
<th>Player Type</th>
<th>Total N</th>
<th>N of Events</th>
<th>Censored N</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overseas</td>
<td>459</td>
<td>356</td>
<td>103</td>
<td>22.40%</td>
</tr>
<tr>
<td>Indian</td>
<td>664</td>
<td>502</td>
<td>162</td>
<td>24.40%</td>
</tr>
<tr>
<td>Total</td>
<td>1123</td>
<td>858</td>
<td>265</td>
<td>23.60%</td>
</tr>
</tbody>
</table>

**Table 2** Group means for survival time in terms of number of balls faced

<table>
<thead>
<tr>
<th>Player Type</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>95% Confidence Interval</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overseas</td>
<td>20.581</td>
<td>0.853</td>
<td>18.908</td>
<td>22.253</td>
<td></td>
</tr>
<tr>
<td>Indian</td>
<td>18.444</td>
<td>0.673</td>
<td>17.124</td>
<td>19.763</td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>19.356</td>
<td>0.534</td>
<td>18.31</td>
<td>20.403</td>
<td></td>
</tr>
</tbody>
</table>

**Table 3** Group medians for survival time in terms of number of balls faced

<table>
<thead>
<tr>
<th>Player Type</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>95% Confidence Interval</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overseas</td>
<td>4.289</td>
<td>0.038</td>
<td>4.190</td>
<td>4.347</td>
<td></td>
</tr>
<tr>
<td>Indian</td>
<td>4.289</td>
<td>0.038</td>
<td>4.190</td>
<td>4.347</td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>4.289</td>
<td>0.038</td>
<td>4.190</td>
<td>4.347</td>
<td></td>
</tr>
</tbody>
</table>

**Table 4** Hypothesis testing based on log-rank test

<table>
<thead>
<tr>
<th>Overall Comparisons</th>
<th>Chi-Square</th>
<th>Df</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log Rank (Mantel-Cox)</td>
<td>4.289</td>
<td>1</td>
<td>0.038</td>
</tr>
</tbody>
</table>

From the above survival plot, one can easily be observed that there is a very little difference between standing ability of overseas and Indian batsmen on the pitch in IPL 2012. Up to 18 balls there is no difference between overseas and Indian batsmen. However, after 20 balls and up to 57 balls, it has been observed that overseas batsmen have moderately more standing ability on the cricket pitch than Indian players in IPL 2012. The following log-rank test has also confirmed that there is a significant difference in terms of standing ability on the cricket pitch between overseas and Indian batsmen (as \( p \)-value is 0.038 < 0.05) in IPL 2012. Earlier researcher such as Stefani and Clarke, Clarke and Harville and Smith, Clarke and Norman, Clarke and Allsopp and Allsopp and Clarke have acknowledged that there is

Citation: Saikia H, Bhattacharjee D. Survival ability of Indian and overseas batsmen on the cricket pitch in Indian premier league. MOJ Sports Med. 2018;2(4):113–116. DOI: 10.15406/mojisn.2018.02.00057
a significant home advantage for home team in the game of cricket. Now, both survival plot as well as log-rank test, have confirmed that there is an advantage for overseas batsmen than Indian batsmen. Thus, the advantage for Indian batsmen playing in their home country or home ground is become a question mark in IPL 2012. This finding can be considered as future scope of this research. May be there was no significant home advantage in IPL 2012 for the Indian batsmen (Figure 1).

**Figure 1** Survival functions of Indian and overseas batsmen in IPL 2014.

**Conclusion**

This study tries to examine the survival ability of Indian and overseas batsmen on the cricket pitch using survival analysis in Indian Premier League. The study has identified that overseas batsmen have moderately more standing ability on the cricket pitch than Indian batsmen after they faced 20 number of balls in IPL 2012. However, up to 20 number of balls, there is no difference between overseas and Indian batsmen in terms of standing ability on the cricket pitch. Since overseas batsmen have moderate advantage than Indian batsmen in IPL 2012; therefore, the home advantage for Indian players in IPL 2012 can be considered as future scope of the study. The proposed survival model can be used to forecast the survival rate of the batsmen in other format of cricket also, while the game is in progress. It can also be used to arrange the batting order of a team in the game of cricket based on the match situation.

**Acknowledgements**

None.

**Conflict of interest**

Author declares that there is no conflict of interest.

**References**