27’th March, 2018.

Dhaka,Bangladesh.

EAST WEST UNIVERSITY.

SPRING 2018

CSE464 (Advanced Database Management System)

Cricket Batsman Data Analysis 2018

https://lh4.googleusercontent.com/1ZmrL992tPTivJ5-mTl07hCi4bh8Qk4VzVKHJ99THYgU1FtD_4Mj3znRmiaNaq_LvEZX79LJrG8VQISLjcYeVFJZj77s62FJoQ2vlSPPQmMdRT85YfTdStftyyE729rudukTLqtU



*Submitted By*

*Nazia Zerin. 2013-2-60-059*

*Md. Mehedy Hasan. 2014-1-60-015*

*Birat Saha. 2014-2-60-086*

*Kevin Stephen Biswas. 2014-2-60-091*

# Introduction

We have the cricket some batsman data 2017 to 2018. We have data files to begin with. Matches - giving us details of each batsman played. We would work on these data to extract some analysis on each batsman’s performance, runs, strike rate, average and dismissal etc.

## **Objectives**

We have analyzed the dataset using different graphs and distribution. Form our observation we have come to different conclusions of the data. Bellow we have focused on some of this important conclusion about the data and performance of the batsman. We have tried to give descriptive explanation behind our conclusion.

**Dataset Information:**

Our dataset contains total 297 instances of recent test batting record of 6 batsmen.

Attributes of the dataset:

-Batsman

-Runs

-Mins [Minutes-the time batsman played in that match (in minutes)]

-BF [Ballfaced]

-4s [the number of 4s he took]

-6s [the number of 6s he took]

-SR [Strike Rate= runs/BF(given in percentage(%))]

-Pos [Position]

-Dismissal

-Inns [Innings=1-4 innings]

We have analyzed the dataset using different graphs and distribution. Form our observation we have come to different conclusions of the data. Bellow we have focused on some of this important conclusion about the data and performance of the batsman. We have tried to give descriptive explanation behind our conclusion.

**Performance Analysis of the Players:**

Comparison 1:

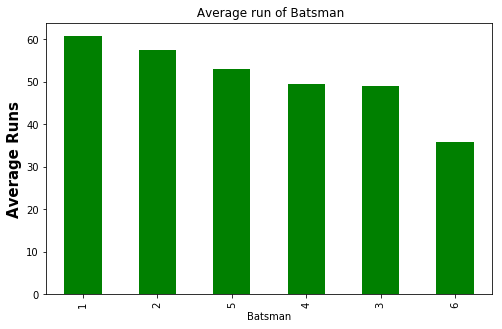


Figure – 1

This graph shows the average runs of each batsman. From the graph we can see that on average batsman 1 runs is best then chronologically batsman 2, 5, 4, 3, 6.

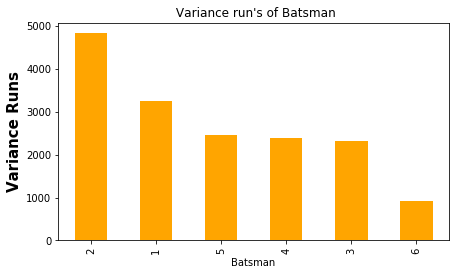


Figure - 2

This graph represents the variance of runs of each batsman’s. Variance represents the spread between the runs of a batsman. So, if a batsman have larger variance form their means it means he had matches in which he did runs which is farther away from his average runs. We can also say there are matches where he does a lot of runs (above the mean) and there are matches where he does very little runs (below the means). So this type of batsman doesn’t have consistency in their performance but if they play well in a match there are chances that he will have great contribution in winning the match.(e.g: batsman 2)

On the other hand, if a batsman’s variance in runs is small that means he scores runs very close to his average runs. This type of players give consistent performance in games. (e.g: batsman 6)

Comparison 2:

|  |  |
| --- | --- |
| Table-1: Batsman More than 100 run chase analysis | |
| Batsman | Chase more than Hundred (%) |
| 1 | 24.56 |
| 2 | 21.15 |
| 5 | 18.42 |
| 4 | 15.25 |
| 3 | 10.93 |
| 6 | 3.84 |

This table shows how many of the time each batsman have been able to chase down more than 100 runs. From this values we can see a positive correlation with the figure – 1. So, if the amount of time a batsman chases more than 100 increases then that batsman’s average run will also increase(e.g: batsman 1). On the other hand if the percentage of a batsman chasing more than 100 decreases the average run of the batsman will also decrease(e.g: batsman 6).

|  |  |
| --- | --- |
| Table-2: Batsman More than 50 run chase analysis | |
| Batsman | Chase more than Fifty (%) |
| 5 | 47.36 |
| 3 | 46.87 |
| 1 | 45.61 |
| 4 | 37.28 |
| 6 | 34.61 |
| 2 | 32.69 |

This table shows the percentage of times a batsman was able to chase down more than 50 runs.

|  |  |
| --- | --- |
| Table-3: batsman get 50 but can't make 100 run percentage analysis | |
| Batsman | Chase more than fifty but can’t make century (%) |
| 3 | 32.81 |
| 6 | 30.76 |
| 5 | 26.31 |
| 4 | 20.33 |
| 1 | 17.54 |
| 2 | 11.53 |

The higher the percentage of a batsman in table-3 means a batsman is more comfortable at taking half-century but he struggles when he tries to turn that half-century into century.

Comparison 3:

Batsman strike rate analysis for (less than 50 run) (50-to-100 run) (more than 100 run).

|  |  |  |
| --- | --- | --- |
| Table-4: Batsman Strike Rate Analysis | | |
| Batsman 1 | Batsman 2 | Batsman 3 |
| C:\Users\Kevin Stephen Biswas\AppData\Local\Microsoft\Windows\INetCache\Content.Word\3.png | E:\CSE-464 (Advance DataBase)\Project\4.png | E:\CSE-464 (Advance DataBase)\Project\5.png |
| Batsman 4 | Batsman 5 | Batsman 6 |
| E:\CSE-464 (Advance DataBase)\Project\6.png | E:\CSE-464 (Advance DataBase)\Project\7.png | E:\CSE-464 (Advance DataBase)\Project\8.png |

From the Strike Rate pie chart of the batsman we have found a positive correlation between a batsmen doing Century and his Average runs.

|  |  |  |
| --- | --- | --- |
| Table-5: Positive Correlation between Century and Average Runs | | |
| Batsman | Century (%) | Average Runs |
| 1 | 19.7 | 60.87 |
| 2 | 17.5 | 57.55 |
| 5 | 15.6 | 53.02 |
| 4 | 13.2 | 49.40 |
| 3 | 9.9 | 49.07 |
| 6 | 3.7 | 35.76 |

So, from this table we can see that there is positive correlation between Century of a batsman and his average runs. It means if a batsman does more century then his average runs also increases. If the batsman have less century then his average runs is also decreases.

Comparison 4:

Dismissal after doing 50 runs.

|  |  |  |
| --- | --- | --- |
| Table-6: Dismissal after doing 50 runs Analysis | | |
| Batsman | Pie Chart | Swarm Plot |
| 1 | E:\CSE-464 (Advance DataBase)\Project\9.png | E:\CSE-464 (Advance DataBase)\Project\10.png |
| 2 | E:\CSE-464 (Advance DataBase)\Project\11.png | E:\CSE-464 (Advance DataBase)\Project\12.png |
| 3 | E:\CSE-464 (Advance DataBase)\Project\13.png | E:\CSE-464 (Advance DataBase)\Project\14.png |
| 4 | E:\CSE-464 (Advance DataBase)\Project\15.png | E:\CSE-464 (Advance DataBase)\Project\16.png |
| 5 | E:\CSE-464 (Advance DataBase)\Project\17.png | E:\CSE-464 (Advance DataBase)\Project\18.png |
| 6 | E:\CSE-464 (Advance DataBase)\Project\19.png | E:\CSE-464 (Advance DataBase)\Project\20.png |

From the graph we can see that the Run Out percentage of a batsman have negative correlation with the variance of runs. Meaning:

If Run Out percentage increases, Variance decreases

And if Run Out percentage decreases, variance increases.

So, if a batsman have less run out in this carrier. Because it means he is a more consistent player.

Comparison 5:

Batsman Position

|  |  |  |
| --- | --- | --- |
| Table-7: Pie chart showing Average of Batsman at different batting position | | |
| Batsman 1 | Batsman 2 | Batsman 3 |
| E:\CSE-464 (Advance DataBase)\Project\21.png | E:\CSE-464 (Advance DataBase)\Project\22.png | E:\CSE-464 (Advance DataBase)\Project\23.png |
| Batsman 4 | Batsman 5 | Batsman 6 |
| E:\CSE-464 (Advance DataBase)\Project\24.png | E:\CSE-464 (Advance DataBase)\Project\25.png | E:\CSE-464 (Advance DataBase)\Project\26.png |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table-8: Total number of matches played in a Position | | | | | |
|  | Position 1 | Position 2 | Position 3 | Position 4 | Position 5 |
| Batsman 1 |  |  | 25 | 30 |  |
| Batsman 2 |  |  | 2 | 46 | 4 |
| Batsman 3 |  | 1 | 22 | 37 | ***4***\* |
| Batsman 4 | 16 | 42 | 1 |  |  |
| Batsman 5 |  |  | 38 |  |  |
| Batsman 6 | 25 |  |  |  | 1 |

From the above table-7 and table-8 it’s clear that there is a relation between the no. of matches a batsman played in a specific position and his average runs in that specific position. In most cases longer a batsman played in a specific position the better his average runs in that position became. In other words the more experience a batsman had playing in a position the better his performance became in that position. The only exception to this was for Batsman 3 playing in position 4. Where despite the batsman having few matches in that position he scored surprisingly well.

Prediction of career average of these batsmen based on sample data using t-distribution:

|  |  |  |
| --- | --- | --- |
| Table-9: T-distribution of 3 batsman(with 95% confidence level) | | |
|  | Confidence Interval Lower Level | Confidence Interval Upper Level |
| Batsman1 | 38.28 | 76.84 |
| Batsman2 | 45.79 | 75.97 |
| Batsman5 | 40.99 | 65.06 |

T-distribution helps us to make assumption of the population data from the sample data. In other words we can make an assumption how our sample data would behave if it was a data of complete population. In this case instead of having a batsman couple of test matches record if we had his all test matches record how the average run would actually look like.

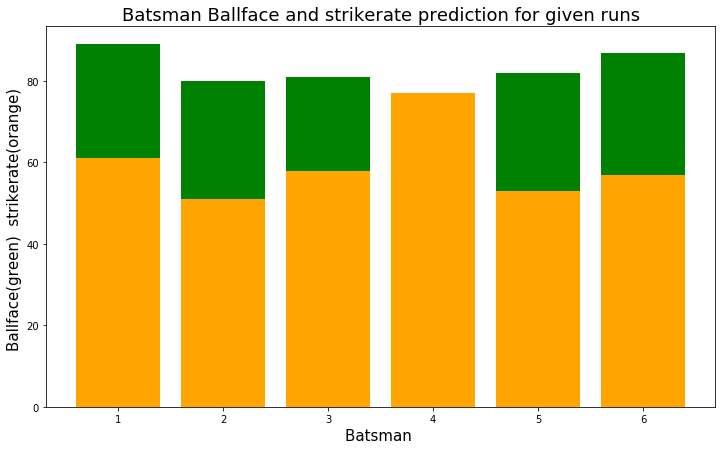
In table-9 there are confidence interval [Lower level, Upper level] is given with 95% confidence interval. Example:

For Batsman5 we predict that, if we had repeated the t-distribution 100 times, 95 of those times the average run of the batsman5 would fall between 36.57-to-69.49 runs.

**Prediction on Required Balls or Strike Rate a Batsman needs for a Specific Run:**

We have used the linear Regression algorithm to make for each batsman how many ball he would need to face or how much strike rate he need to have in order to do 50 in his next match. We can find the required BallFace and StrikeRate for any Runs but here we have chosen half-century(50).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Table-10: BallFace and StrikeRate Prediction for 50 Runs | | | | | | |
|  | Batsman1 | Batsman2 | Batsman3 | Batsman4 | Batsman5 | Batsman6 |
| BallFace | 89 | 80 | 81 | 66 | 82 | 87 |
| StrikeRate | 61 | 51 | 58 | 77 | 53 | 57 |



We may also be able to predict the Actual Batsman solely from his performance in table-11.

|  |  |
| --- | --- |
| Table-11: Who is 6 batsman’s actual Identity(name) | |
| Name | Player Name |
| Batsman 1 | Virat Kohli(IND) |
| Batsman 2 | AN Cook(ENG) |
| Batsman 3 | Q De cock(SA) |
| Batsman 4 | Joe Root |
| Batsman 5 | David Warner(AUS) |
| Batsman 6 | Tamim Iqbal(BAN) |

Conclusion:

The analysis of the Test Scores of 6 batsman have made us see this dataset in different light. We have found couple of interesting hidden relationship which have noticeable influences on a batsman’s performance. We have prediction how well a batsman will play in a future match and how much run we can expect form him. Also we have predicted how much ball a batsman will need to face or how much strike rate does he need to maintain in his next match in order to score a specific run. Finally we have identified who this 6 mystery batsman’s are in real life from their performance.