A Bayesian stochastic model for team performance evaluation in T20 cricket: Effectiveness of Power Hitting & Consistency

Srisai

July 20, 2016

"Dealing with comparisons in cricket is harder and more complex than in most other sports... Averages can be a guide, but are not conclusive since pitches and conditions have changed." - Sir Donald Bradman, taken from Harold Larwood by Duncan Hamilton

As much as Cricket is a game of the minds, its a game of numbers (read: stats), also permutations and combinations. Its not all that hard to imagine how the myraid pitch, weather, playing and match conditions. It might not come as a surprise that it can rival the possible moves in a game of chess. An astute follower of the game (the author's dad) observed that "Cricket is the game of chess played by a team, generating useful, among plenty of pointless stat." (Howastat?)

Introduction

In T20I Cricket, there are an endless number of questions that are not amenable to experimentation or direct analysis but could be easily addressed via simulation. For example, on average, would India benefit more from increasing their Strike Rate of their middle order or consistency of the top-order. If they do, what percentage of time would India be expected to win over Australia while setting, in Sharjah?

As the game of cricket has developed over the last five centuries, more and more detailed statistics have become available. These are important indicators and can be used to predict a game's outcome, or at least give a possible indication of an expected result. In fact, the New Zealand coach, Mike Hesson, quite eloquently answered is T20 the format where data is most effective at yielding that advantage, while addressing in general, the use of data in the modern game.

While Hesson's interview sets up the premise for the analysis, its predicated by the more recent article by Karthikeya Date, on Why hitting is more optimal than batting in T20. While the results of the Date's analysis reveal the importance of power hitting, this analysis aims to investigate the impact of power hitting, contrasted with improving consistency, their effects on the outcomes of a n-match bilateral series between 2 teams of choice by the use of a Bayesian framework to predict the outcome of the series.

This analysis currently investigates the outcomes of games from a batsman point of view. This restricts the variables being considered for analysis to be strictly batsman centric.

Why Bayesian

The most frequently used statistical methods are known as frequentist (or classical) methods. These methods assume that unknown parameters are fixed constants, and they define probability by using limiting relative frequencies. It follows from these assumptions that probabilities are objective and that you cannot make probabilistic statements about parameters because they are fixed. This approach is best suited to falsify a hypothesis, but the current analysis needs a framework best suited to (re)allocate the credibility of a statement or degree of belief in one of the teams winning, as the 'variables' (like power hitting or consistency) involved in the game changes.

So we seek a natural and principled way of combining prior information with data, within a solid decision theoretical framework. This can be achieved by incorporating past information about a parameter to form a

prior distribution for future analysis and with the availability of new observations, the previous posterior distribution can be used as a prior. All inferences logically follow from Bayes' theorem. Thus the analysis employs a Bayesian framework to estimate the series outcomes.

Assumptions

Years of playing/following experience at various capacities has enabled the author to uphold assumptions deemed reasonable-in-context and necessary for the analysis. As stated previously, the analysis would entirely batting centric.

The most important assumption is that the score, in itself, is the sole and just reflection of all the factors relevant to describing the contest and its outcome.

Players

It's assumed that the player is available for all the matches in the simulated series, without absence due to injury, personal reasons and the like. A key assumption is that the previous performances of players are good predictors of current/future performances. Where possible (= data available) player performance against the specific opposition, either setting or chasing will be used as the player's data. Shuold this data be small (less than 2 data points), less restrictions would be applied. This implies that there is a possibility that a player with no previous matches against India, might have his data comprising of performances against other teams. An example is Carlos Brathwaite, who has played 1 game against India. In the simulations, his performance against India would be predicted based his performances against other teams. Its also assumed that the player's data captures his effectiveness agains pace and/or spin.

In the pre-computer days, Elderton (1945) and Wood (1945) fit the geometric distribution to individual runs scored based on results from test cricket. Kimber & Hansford (1993) argue against the geometric distribution and obtain probabilities for selected ranges of individual scores in test cricket using product-limit estimators. More recently, Dyte (1998) simulates batting outcomes between a specified test batsman and bowler using career batting and bowling averages as the key inputs without regard to the state of the match (e.g., the score, the number of wickets lost, the number of overs completed). Bailey & Clarke (2004, 2006) investigate the impact of various factors on the outcome of ODI cricket matches. Some of the more prominent factors include home ground advantage, team quality (class) and current form. Their analysis is based on the modelling of runs using the normal distribution. In light of these studies, we assume that the players performance is normally distributed and the predicted score in a match can be estimated from his sample using the gaussian parameters. There is no reason to believe otherwise.

Playing XI

Each of the teams in the simulation would have the same playing XI. Under no circumstance during a match that would a player be rendered unavailable for batting or bowling. The 3 teams being investigated are India, West Indies and Australia. The each of the 3 teams is assumed to retain the same playing XI for the simulations, as their last game in the recently concluded World T20.

Playing Conditions

No distinction is made between host countries. Hence the host country and the venue ground is assumed to have no effect on the player performance or outcome of the game. This is a overriding pragmatic requirement given that the T20 is a relatively new format and most international players havent yet played in most of the host countries, let alone individual grounds.

All opposition are treated on the merit of their data. No distinction is made on the quality of opposition. So its just as likely for Kohli to score a fifty against Afghanistan as it is to score against Australia.

The restrictive nature of the data also limit the effect ofsSetting and chasing a total to have no effect on the outcome of the game. So a team is equally likely to score certain total, irrespective ot setting or chasing. Another assumption that is assumed to have been captured in the predicted score is the quality of fielding of the teams!

The toss and the decision to set/chase collectively is decided by the team batting first, which is chosen at random. The weather and pitch conditions is assumed to be captured in the past scores and performances. And since the predicted scores are dependent on past scores, the pitch and weather conditions are assumed to be captured in the predctions.

Match Outcomes

The teams are allowed to complete their full innings (while chasing) of 20 overs. The team with higher runs is declared to be the winner of the game. Team with 4 or more wins in the 7 match series is declared the series winner.

Number of series simulated: Monte Carlo method

Each 7 match series is simulated 1000 times. The repeated random sampling is used to obtain numerical results that reliable mimics the dynamics of a cricket. This helps in generating the probability distribution and the likelihood required for the subsequent Bayesian analysis. 1000 is deemed a good enough number of simulations of a series, given that there has been a grand total 0f ~500 T20I till date.

Choice of Prior

A prior distribution of belief is required for the Bayesian Analysis. This is chosen by the author's prior cricketing experince. Its assumed that the probability of win or loss is symmetrically distributed, with the highest prior for the series going 4-3 to either sides.

Prior series win probability for India = 0.5
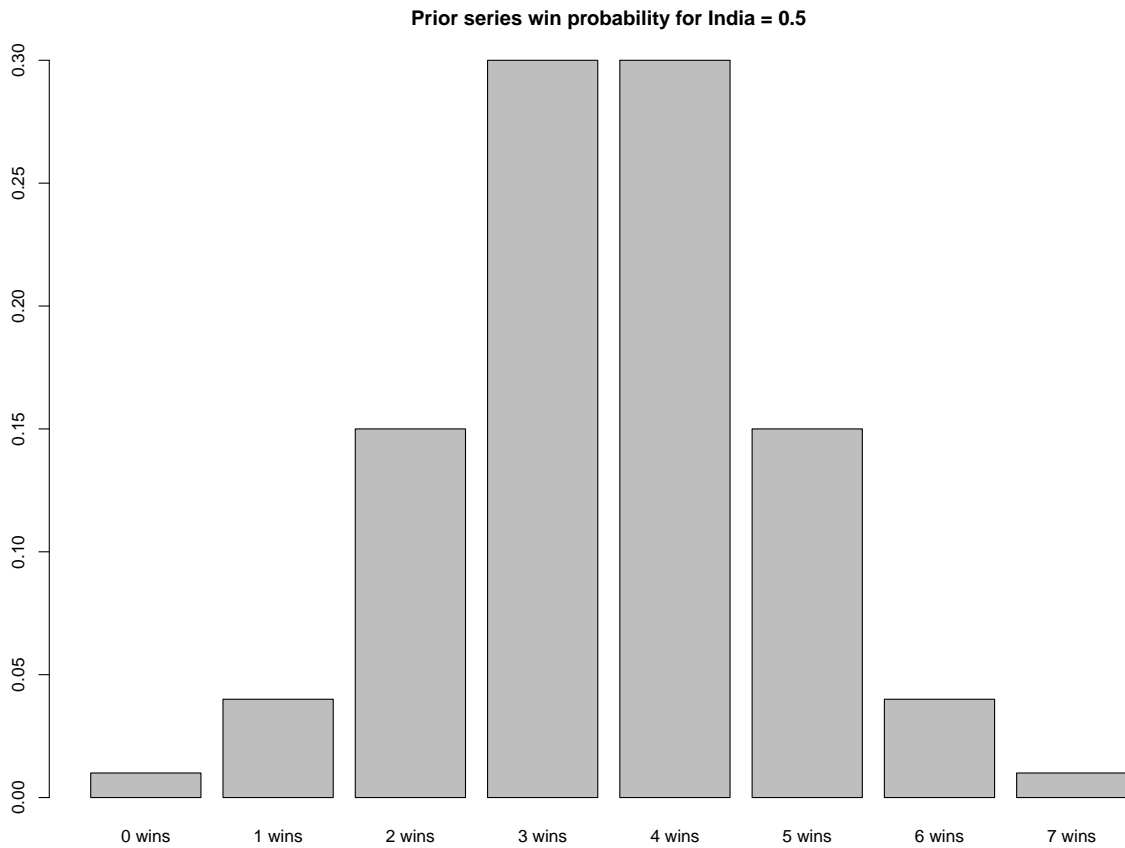
This shows that there is an equal chance of either teams winning a 7 match series. Since there is no evidence to suggest that one team is inherently better than the other, this prior is fair enough.

Effects of assumptions

Despite these many simplifications and assumptions, the proposed simulator appears to do a reasonable job at producing realistic scores and results. This simulation allows the investigation of complex questions involving T20I cricket matches, some of which are probability of winning a series and the effect of power hitters.

The data

Of the teams participated in the World T20, Australia, India and West Indies, 3 of the most successful T20I teams of recent times. This can, in part, be attributed to the largely successful domestic leagues like the BBL, IPL and CPL.

Though simulations for Australia Vs. India and West Indies are available, all subsequent analysis will ignore Australia and feature India-West Indies games.

All the data for the individual players has been scrapped from ESPNCricinfo's Statsguru. As mentioned earlier, the teams from the last game of their T20 World Cup has been chosen for the simulations.

This is the playing XI for India.

```
##                  India
## 1        Rohit Sharma
```

```
## 2         Ajinkya Rahane
## 3           Virat Kohli
## 4             MS Dhoni
## 5          Suresh Raina
## 6         Manish Pandey
## 7         Hardik Pandya
## 8        Ravindra Jadeja
## 9   Ravichandran Ashwin
## 10          Ashish Nehra
## 11       Jasprit Bumrah
```

This for West Indies.

```
##            West.Indies
## 1     Johnson Charles
## 2          Chris Gayle
## 3      Marlon Samuels
## 4       Lendl Simmons
## 5         Dwayne Bravo
## 6        Andre Russell
## 7         Darren Sammy
## 8    Carlos Brathwaite
## 9         Denesh Ramdin
## 10        Samuel Badree
## 11       Sulieman Benn
```

Scoring model and baseline simulation

As stated before, the players past batting/bowling performance is used to predict his performance in the simulation. Its assumed that the past perfromance is normally distributed with the mean around his career average. This, along with prior standard deviation is used to predict the performance in the current match, and the series. Same philosophy is applied to the bowlers too.

This rules are applied this to a 1000 7-match series between India and West Indies. The results examined below.

Performance

Its imperative that the outcomes of the model are examined thoroughly before using these results for further calculation.
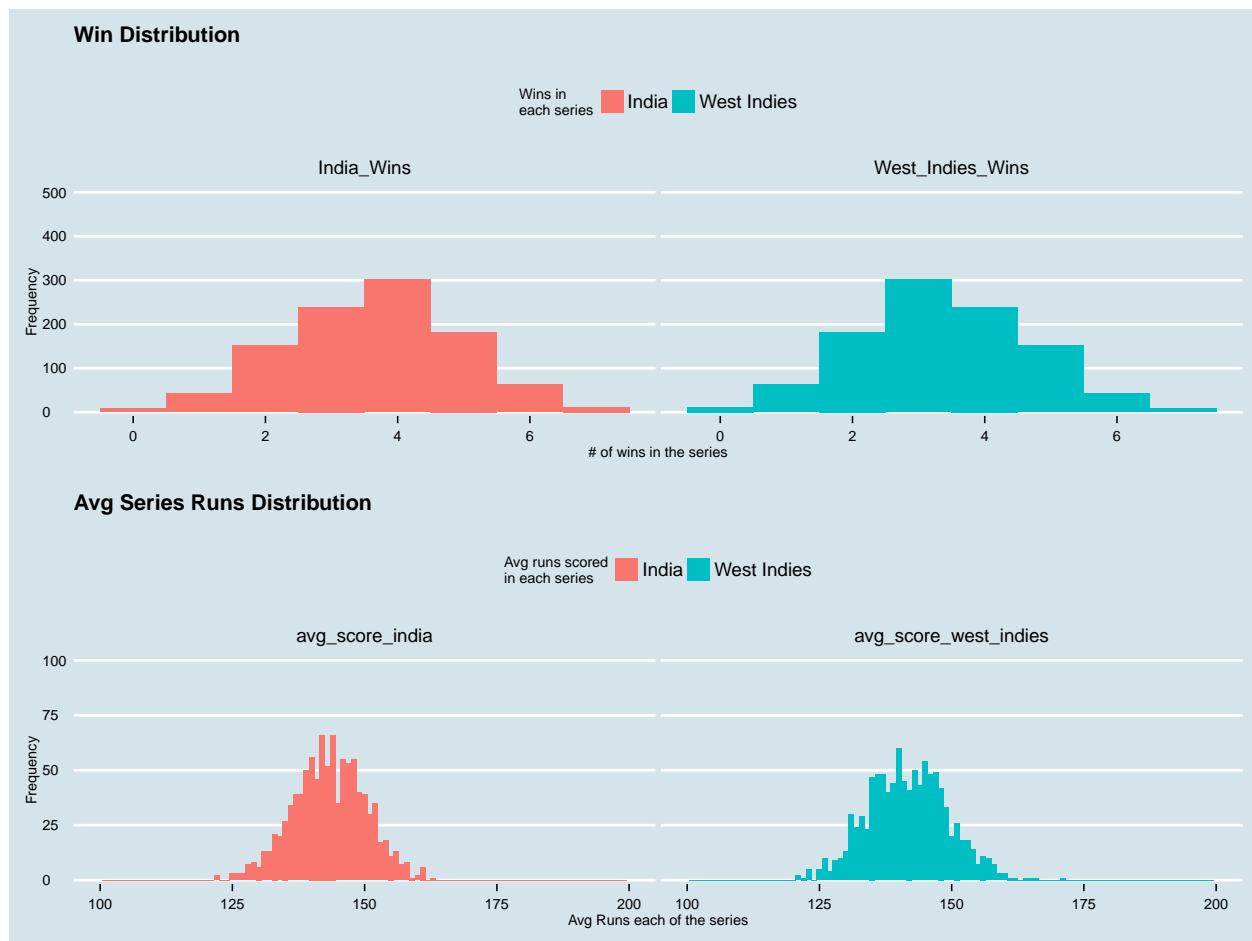
The first plot sugggests that the number of wins in the series is almost normally distributed, with a small right skew for India and left skew for te West Indies. This indicates that India has won more series than the West Indies.

The second plot shows the average runs by each team in each series. The distribution of this looks similar to what one might expect in a modern day T20I tournament.
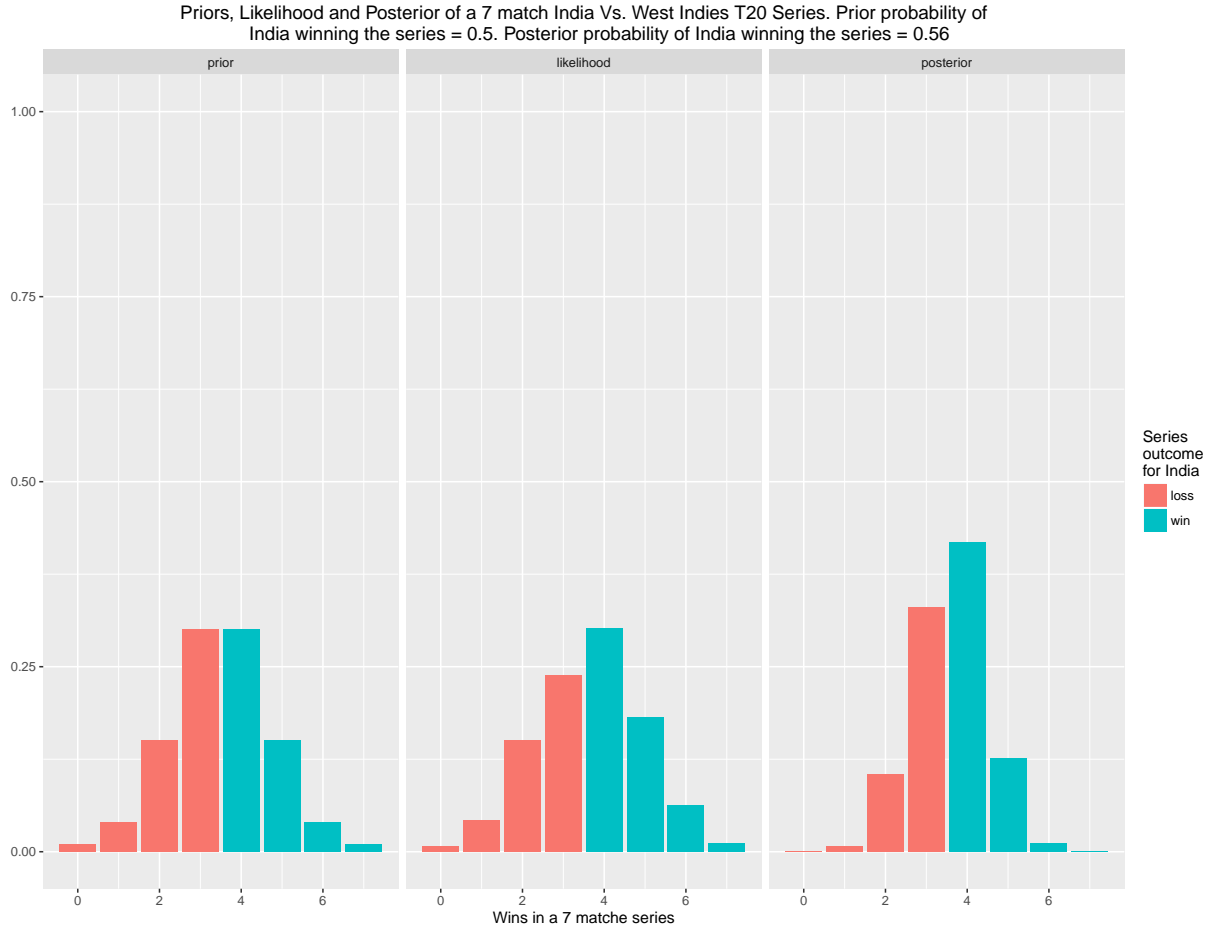
India Vs West Indies: Prior, Likelihood and Posterior

The results thusfar provides the likelihood of each senario and using this in tandem with the prior, the posterior probability of the series outcome can be evaluated.

```
## [1] TRUE
```

```
## [1] TRUE
```

```
## [1] TRUE
```

```
## [1] 1
```

Priors, Likelihood and Posterior of a 7 match India Vs. West Indies T20 Series. Prior probability of India winning the series = 0.5. Posterior probability of India winning the series = 0.56

The above plots shows the proportion of 1000 7-match series India wins. This comes close to 55%. This may be restrictively generalized to state India have a 55% chance of winning a 7 match T20I series against the West Indies, where there is a 0.45 chance of West Indies winning the weries. This is not dissimilar to the initial assumption of a 50-50 chance of either teams wining the series, but this change is effected due to subjective change in belief given the players performance.

It may be of interest to note the difference between the prior and posterior. This change is driven by the likelihood. The difference between the prior and the posterior is expected to increase with an increase in the number of series simulated (effect of sample).
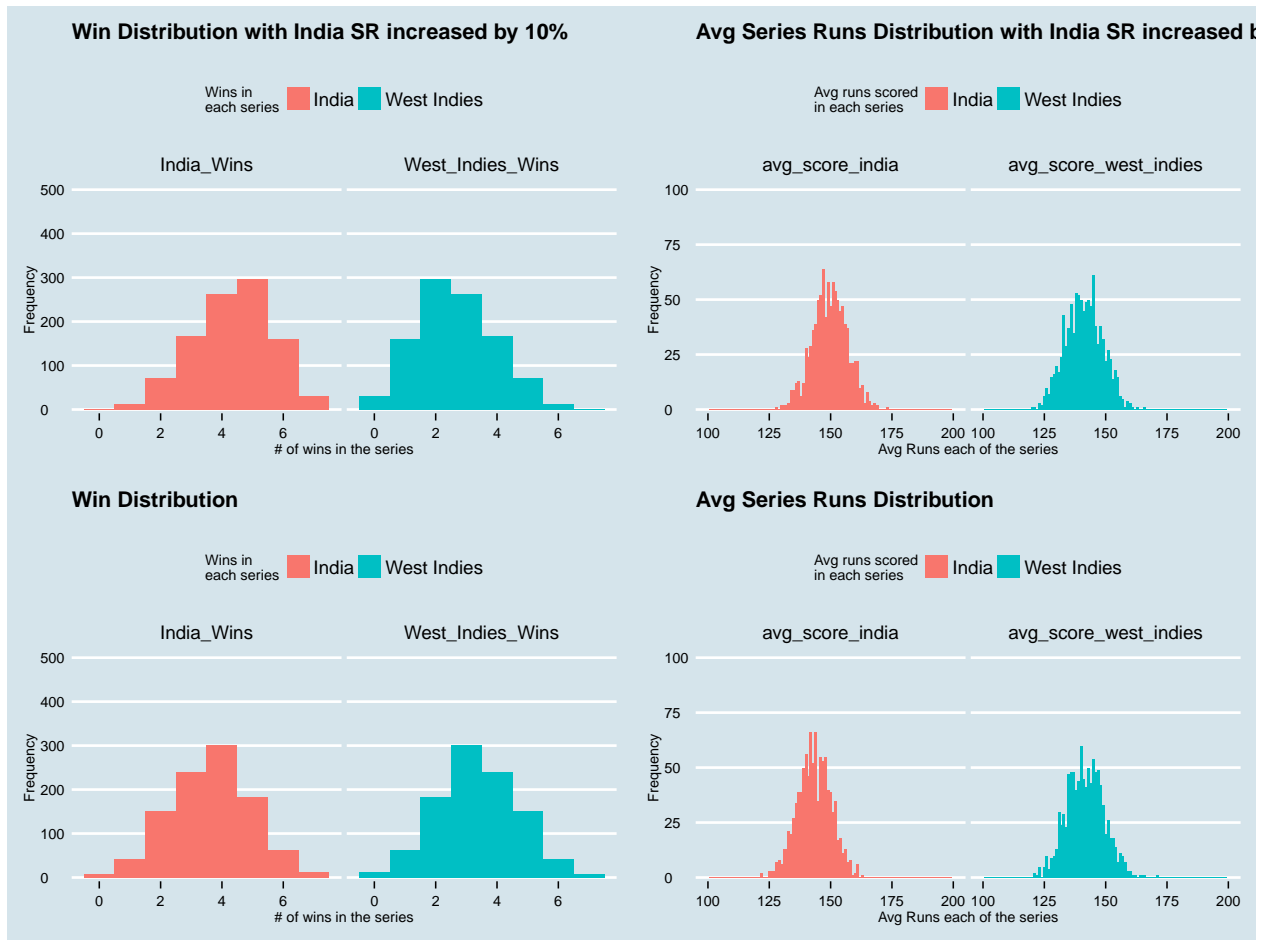
Effect of Hitting Vs Consistency

This provides a robust bayesian simulation framework. With this as the benchmark, the subtle and difficult-to-measure-in-realtime effects/influences of consistency and/or power hitting can be examined, in terms of the posterior win probability. I.e. How does the chances of winning the series changes with change in one of the parameters of interest like power hitting and consistency. The objective is evaluate, if possible, which of these parameters, when changed, results in favourable outcome for the team of choice.
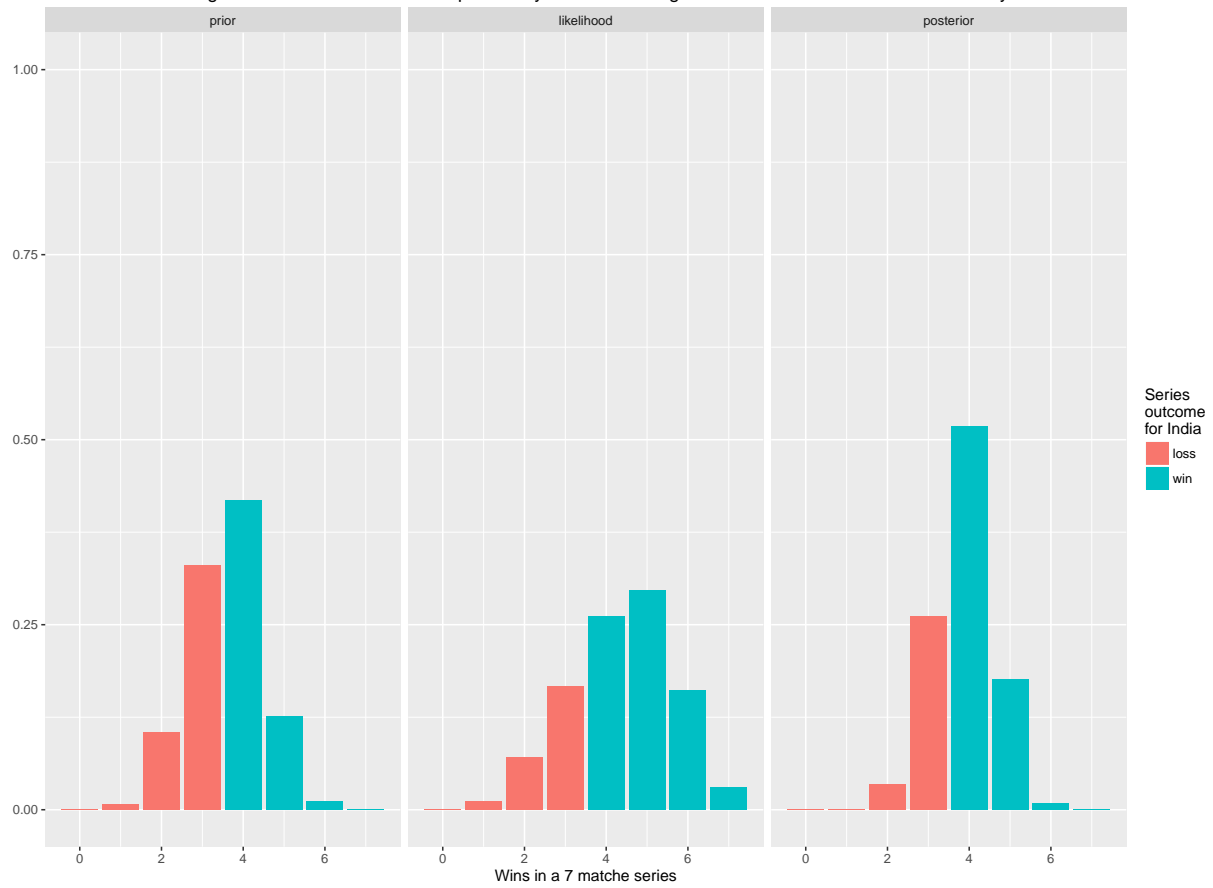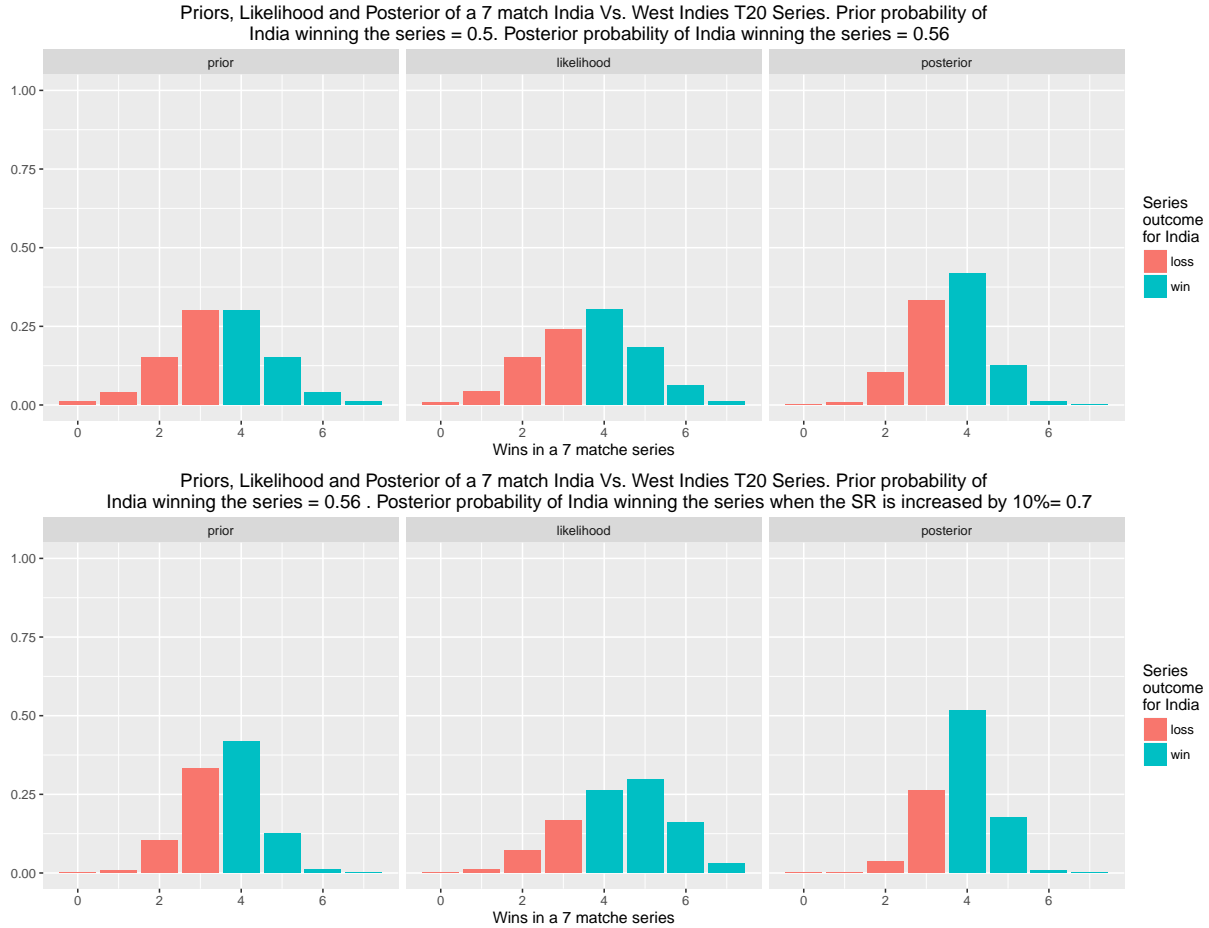
Power Hitting

Power hitting is represented through the strike rate. To systematically evaluate its effects, the strike rates of Indian players will be increased in steps of 2%, until a maximum of 10% over the career strike rate. Only the results of the 10% increased strike rate shall be presented here for the sake of bravity.

The prior used in this is the posterior obtained from the last series simulation.

Win Distribution with India SR increased by 10%

Wins in each series   India   West Indies

India_Wins   West_Indies_Wins

Frequency

# of wins in the series

Avg Series Runs Distribution with India SR increased b

Avg runs scored in each series   India   West Indies

avg_score_india   avg_score_west_indies

Frequency

Avg Runs each of the series

Win Distribution

Wins in each series   India   West Indies

India_Wins   West_Indies_Wins

Frequency

# of wins in the series

Avg Series Runs Distribution

Avg runs scored in each series   India   West Indies

avg_score_india   avg_score_west_indies

Frequency

Avg Runs each of the series

Priors, Likelihood and Posterior of a 7 match India Vs. West Indies T20 Series. Prior probability of India winning the series = 0.56 . Posterior probability of India winning the series when the SR is increased by 10%= 0.7

Priors, Likelihood and Posterior of a 7 match India Vs. West Indies T20 Series. Prior probability of India winning the series = 0.5. Posterior probability of India winning the series = 0.56

Priors, Likelihood and Posterior of a 7 match India Vs. West Indies T20 Series. Prior probability of India winning the series = 0.56 . Posterior probability of India winning the series when the SR is increased by 10%= 0.7

It may be observed from the firt set of histogram of India wins are more right skewed than before, implying India are winning more games (hence series). Whilst the distribution of the runs scored by West Indies hasnt change, there is a marked right migration in that of India- India are making more runs than before. So India are winning more. This is reflected in the series win % going up to 71 from earlier 55.

This categorically shows that power hitting significantly increases the odds of winning a series.
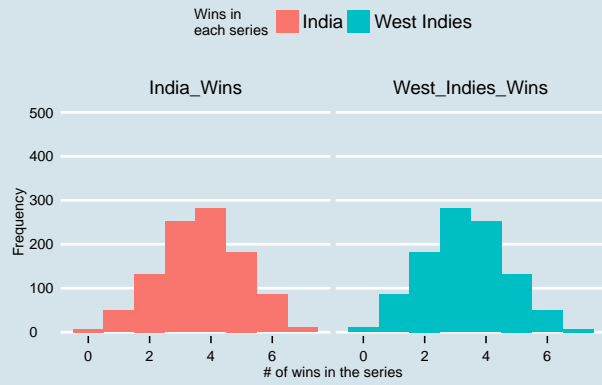
Consistency

The definition of power hitting was straight forward. But the definition of consistency is more involved. One might expect a batsman to be consistent in terms of runs scored. This may be represented by a decrease in the standard deviation, while the average remains unchanged. But this doesnt take into account, the possibility of a batsman slowing his scoring rate. Thus the definition of consistency should encompass the runs scored along with the rates at which is it scored. To achieve this, this analysis employs the consistency definition of lower standard deviation in balls faced and strike rate.
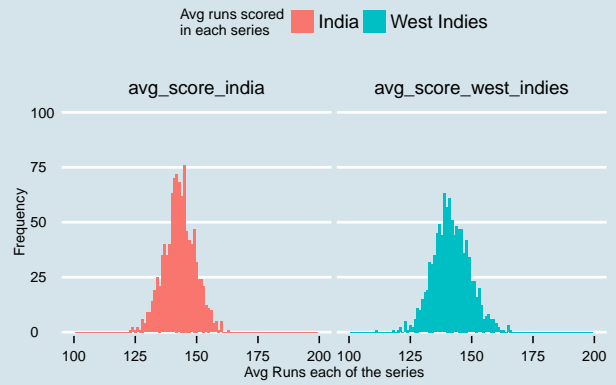
As before, to systematically evaluate its effects, the standard deviation of balls faced and strike rate of Indian players will be decreased in steps of 2%, until a minimum of 10% lower than the career deviations. Only the results of the 10% decreased deviations shall be presented here for the sake of bravity.

The prior used in this is the posterior obtained from the baseline series simulation.
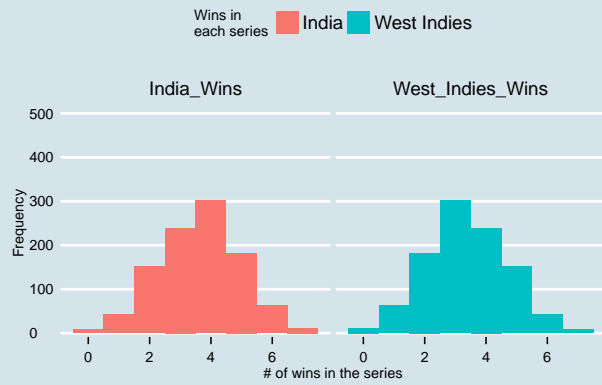
Priors, Likelihood and Posterior of a 7 match India Vs. West Indies T20 Series. Prior probability of India winning the series = 0.56 . Posterior probability of India winning the series when the consistency is increased by 10%= 0.59

Priors, Likelihood and Posterior of a 7 match India Vs. West Indies T20 Series. Prior probability of India winning the series = 0.5. Posterior probability of India winning the series = 0.56
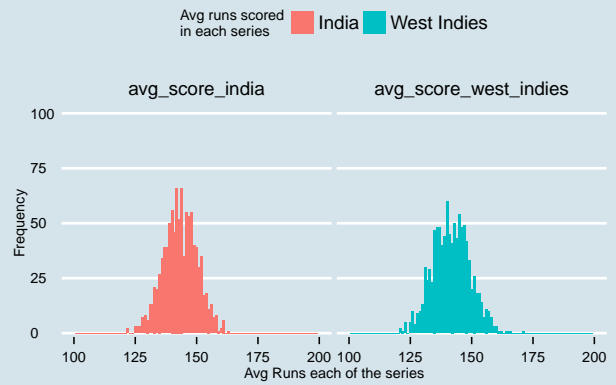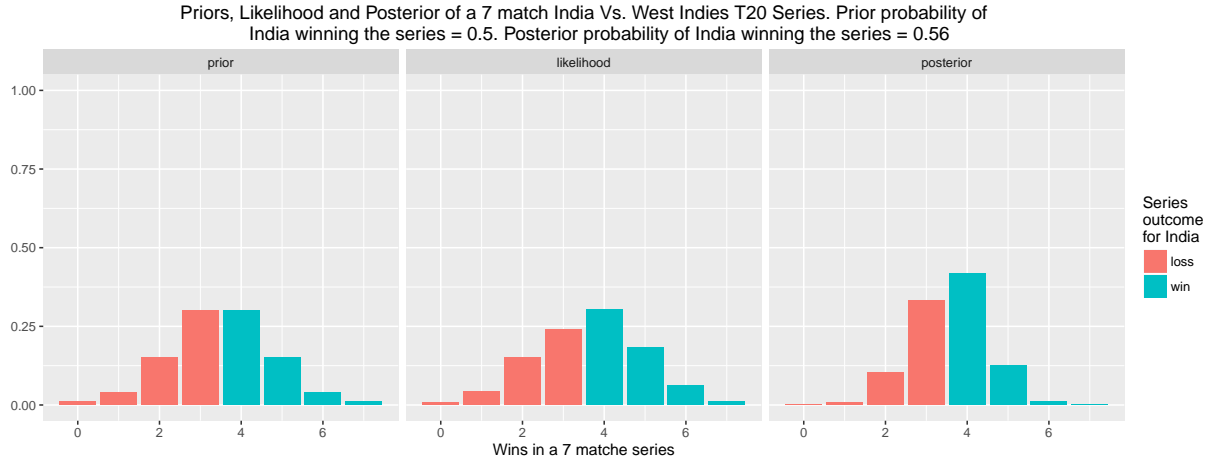

Priors, Likelihood and Posterior of a 7 match India Vs. West Indies T20 Series. Prior probability of India winning the series = 0.56 . Posterior probability of India winning the series when the consistency is increased by 10%= 0.59
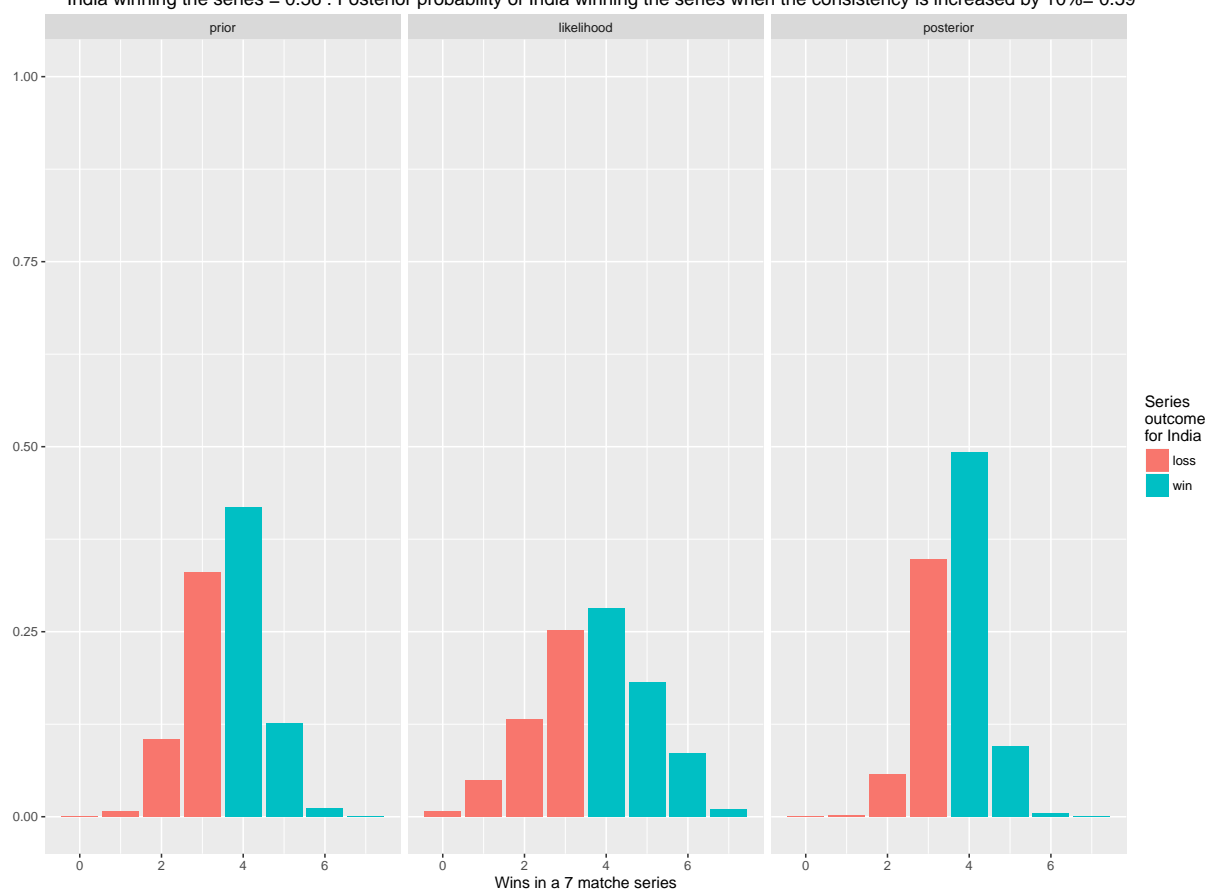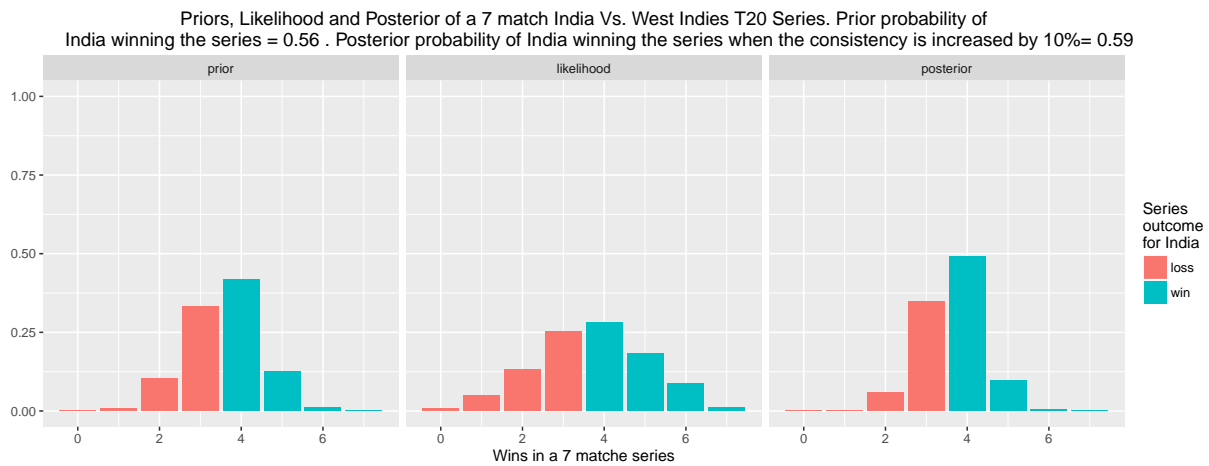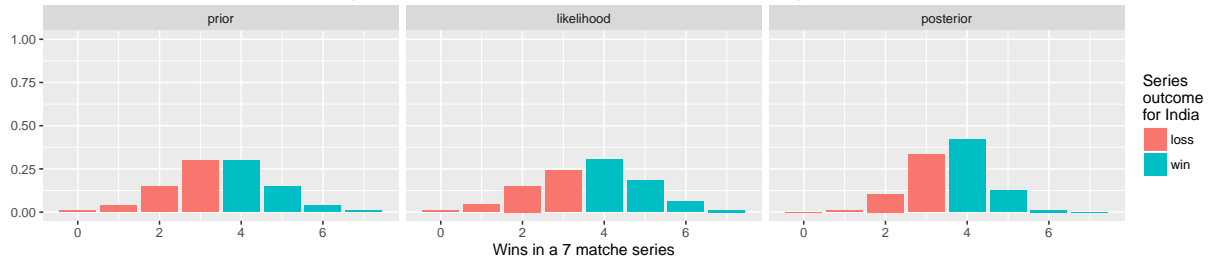
The effect of consistency on win % seems subtle compared to that of power hitting. It was noticed that with increased strike rate the histogram of runs scored shifted right, thus increasing the win % significantly. With the increase in consistency, there isnt much observable deviation from the baseline series. We do notice that the histogram of runs has got a bit 'tighter', implying narrower range of runs scored. This makes intuitive sense- as the consistency of all batsmen increases, do does the scores of the team.

This doesnt increase the win % from a statistical point of view. Same view holds good from a cricketing point of view as well.
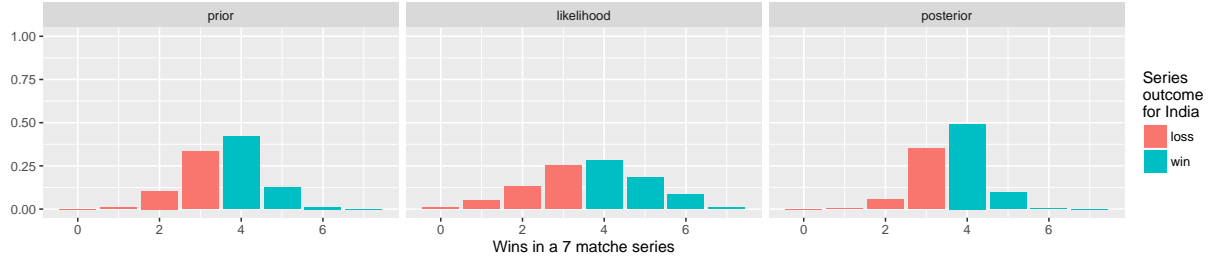
Power Hitting Vs. Consistency: Summary

The analysis clearly shows that in the T20 format, power hitting clearly beats consistency. This makes intuitive cricketing sense too. From cricketing wisdom and this analysis, it may be generalized that shorter the format, more the effect of power hitters on the outcome of games.
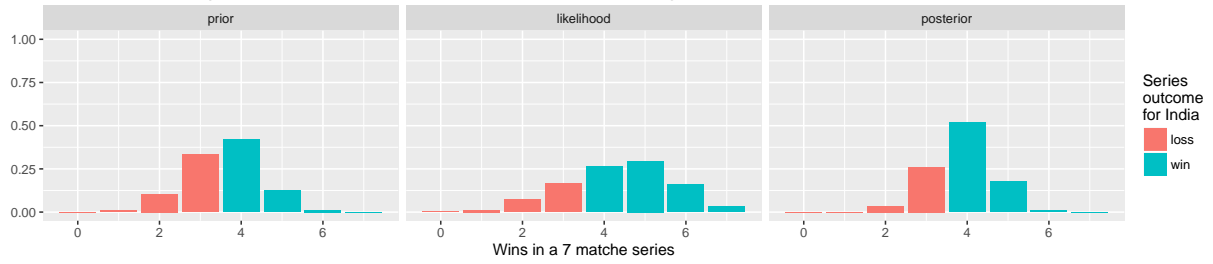
Priors, Likelihood and Posterior of a 7 match India Vs. West Indies T20 Series. Prior probability of
India winning the series = 0.5. Posterior probability of India winning the series = 0.56

Priors, Likelihood and Posterior of a 7 match India Vs. West Indies T20 Series. Prior probability of
India winning the series = 0.56 . Posterior probability of India winning the series when the consistency is increased by 10%= 0.59
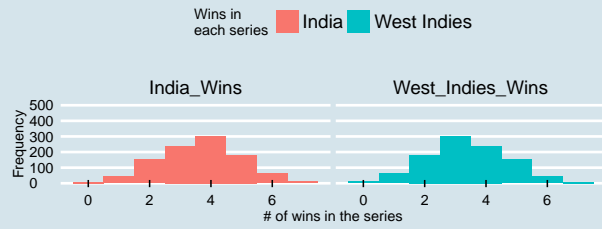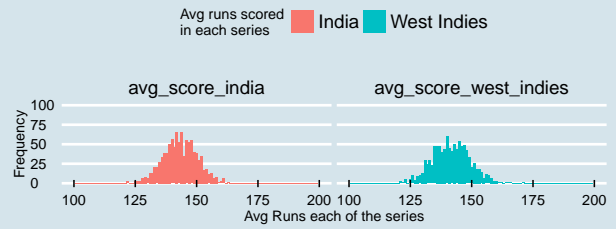

Priors, Likelihood and Posterior of a 7 match India Vs. West Indies T20 Series. Prior probability of
India winning the series = 0.56 . Posterior probability of India winning the series when the SR is increased by 10%= 0.7
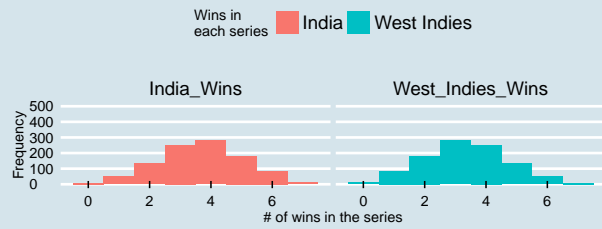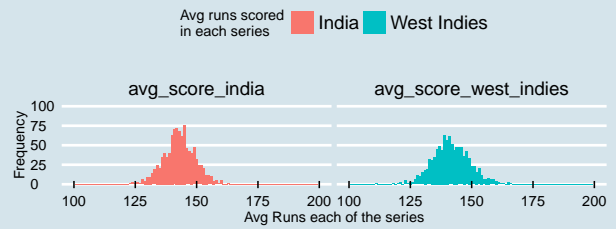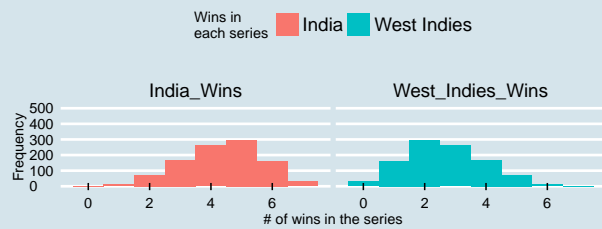
Win Distribution

Wins in each series ■ India ■ West Indies

India_Wins / West_Indies_Wins

Frequency: 500 400 300 200 100 0

# of wins in the series: 0 2 4 6 / 0 2 4 6

Avg Series Runs Distribution

Avg runs scored in each series ■ India ■ West Indies

avg_score_india / avg_score_west_indies

Frequency: 100 75 50 25 0

Avg Runs each of the series: 100 125 150 175 200 / 100 125 150 175 200

Win Distribution with India consistency increased by 1

Wins in each series ■ India ■ West Indies

India_Wins / West_Indies_Wins

Frequency: 500 400 300 200 100 0

# of wins in the series: 0 2 4 6 / 0 2 4 6

Avg Series Runs Distribution with India consttency inc

Avg runs scored in each series ■ India ■ West Indies

avg_score_india / avg_score_west_indies

Frequency: 100 75 50 25 0

Avg Runs each of the series: 100 125 150 175 200 / 100 125 150 175 200

Win Distribution with India SR increased by 10%

Wins in each series ■ India ■ West Indies

India_Wins / West_Indies_Wins

Frequency: 500 400 300 200 100 0

# of wins in the series: 0 2 4 6 / 0 2 4 6

Avg Series Runs Distribution with India SR increased b

Avg runs scored in each series ■ India ■ West Indies

avg_score_india / avg_score_west_indies

Frequency: 100 75 50 25 0

Avg Runs each of the series: 100 125 150 175 200 / 100 125 150 175 200

15

Effect of increasing strike rates and consistencies of Indian players on win probability of 7 match T20I series again

This settles the discussion: Hitting trumps Consistency in T20I. Atleast for India agains West Indies, with the same playing XI of their last world T20I matches respectively.

Index and Cumulative Index

The preceeding analysis have shown the importance of power hitting over consistency in a T20 setting. But this isnt to say that consistency isnt important, because it is. The following section attempts to provide a unified 'index' of batting performance, that would be a better measure than the average (there are plenty of literature available that points out the drawbacks in using average as a measure of batsman performance - one example shown here. While most of these literature deals with the issues of not-outs skewing batting averages, this index ignores the effect of not-outs) and/or strike rate.

There have been talks in the cricketing fraternity to use a "impact index", which is a fancy term for the product of average and strike rate. Though this achieves the unification of average and strike rate, it doesnt capture much about the regularity of it happening. So its a partial good measure at best.

On the other hand, if we use this measure:

index = (average x mean.strike.rate) / (stdev(runs) x stdev(strike.rate))

This captures both the impact, while giving higher weight to consistent performance (w.r.t the average). A tweak can be made to this index to identify the MVPs of batting, by including the career T20I runs:

career_index = log(runs) x (average x mean.strike.rate) / (stdev(runs) x stdev(strike.rate))

```
##              player matches runs   avg   S.R index cum.index
```

16

```
## 1          virat.kohli   40 1641 41.02 126.30  6.04   9911.64
## 2         david.warner   61 1633 26.77 120.88  2.10   3429.30
## 3          suresh.raina   51 1203 23.59 115.69  2.75   3308.25
## 4          rohit.sharma   52 1292 24.85 111.29  2.22   2868.24
## 5         shane.watson   56 1462 26.11 118.85  1.95   2850.90
## 6       marlon.samuels   44 1133 25.75 105.05  2.38   2696.54
## 7              ms.dhoni   62 1069 17.24 131.71  2.50   2672.50
## 8          chris.gayle   47 1519 32.32 121.24  1.74   2643.06
## 9           aaron.finch   28  974 34.79 125.90  2.59   2522.66
## 10         dwayne.bravo   53 1054 19.89 107.86  2.32   2445.28
## 11        lendl.simmons   34  843 24.79  94.22  1.92   1618.56
## 12        glenn.maxwell   30  611 20.37 133.58  2.09   1276.99
## 13        usman.khawaja    7  199 28.43 145.53  5.66   1126.34
## 14      johnson.charles   28  580 20.71  94.08  1.72    997.60
## 15       ajinkya.rahane   18  364 20.22 100.78  2.32    844.48
## 16         darren.sammy   48  534 11.12 136.13  1.55    827.70
## 17         steven.smith   25  431 17.24 103.84  1.81    780.11
## 18        denesh.ramdin   37  421 11.38 109.48  1.65    694.65
## 19        andre.russell   32  310  9.69 105.67  1.37    424.70
## 20       james.faulkner   14  141 10.07 123.41  2.93    413.13
## 21 ravichandran.ashwin   10  112 11.20 118.41  2.11    236.32
## 22       ravindra.jadeja   15  103  6.87  84.72  1.76    181.28
## 23        hardik.pandya    7   78 11.14 101.30  0.86     67.08
## 24  nathan.coulter.nile    7   51  7.29  88.74  1.20     61.20
## 25    carlos.brathwaite    5   59 11.80 140.50  1.02     60.18
## 26        samuel.badree    6   26  4.33  78.66  2.29     59.54
## 27        sulieman.benn    7   37  5.29 100.10  1.25     46.25
## 28        manish.pandey    4   67 16.75  59.28  0.62     41.54
## 29         peter.nevill    4   22  5.50 257.50  1.80     39.60
## 30          adam.zampa    2    7  3.50 150.00  3.50     24.50
## 31         ashish.nehra    5   28  5.60  40.16  0.32      8.96
## 32       jasprit.bumrah    1    0  0.00   0.00   NaN       NaN
```

This shows the MVPs of batting among the 3 teams combined. The top of list comprises of usual suspects-top performers who have had a phenomenal run for a long time, with remarkable consistency. Using the career runs in the index widens the gap between a old, but good performer against a new but extraordinary performer. It gives more weightage to the old, but good performer, over the new and extraordinary performer.

Conclusion

A stochastic Bayesian framework for team performance evaluation in T20 cricket was concieved and implemented. This was tested against the authors cricketing judgements. Upon satisfactory performance of the framework, it was used to examine the effects of hard hitting and consistency on the outcomes of a 7 match T20 series between 2 teams of choice. Power hitting is predicted to be the superior contributor to higher win percentages. The author also proposes a unified career batting index to express the batsman's impact. The current top performers in world T20 is atop the list- lending credibility to the proposed metric.