New Cricket Statistics Inspired by Sabermetrics

Weighted Scoring Rate (wSR), Runs Created (RC, RC+), and Created Against Average (CAA, CAA+)

Introduction

One of the many areas where baseball statistics have an advantage over cricket is in statistics which provide an overall picture of a player's performance. For example, weighted runs created plus (wRC+) provides a complete picture of a baseball hitter's performance [4]. It is also very easy to understand. Weighted runs created plus is scaled so that 100 is the league average. Every point above or below 100 is a 1% increase or decrease in performance. So, a hitter with a wRC+ of 120 is 20% better than average. You can already tell that Francisco Lindor (wRC+ of 130) is a better hitter this year than Josh Donaldson (wRC+ of 104). Cricket does not have such a statistic. When comparing batters people most commonly point to their averages. Firstly, this doesn't measure all aspects of what makes a batter good, especially for limited overs formats. Secondly, you need cricket knowledge to understand what a good average is.

In this project we are directly inspired by wRC+ and create an equivalent statistic for T20 cricket, runs created plus (RC+). This statistic combines aspects of average and strike rate to provide an overall picture of batting performance. We will also create an analogous bowling statistic, created against average plus (CAA+). This statistic combines aspects of bowling average and economy. Both statistics are scaled in a way that makes them easy to understand. In order to arrive at these statistics we also create weighted scoring rate (wSR).

Dataset

The dataset consists of ball by ball information from Cricsheet [1]. Cricsheet data is available under the Open Data Commons Attribution License. The dataset consists of international and club T20 matches that meet the following conditions:

- The match was played during or since 2015. The scoring rate in T20 matches has increased over time. Including only recent matches gives more accurate results.
- International matches must be between full ICC member nations. Matches with associate nations often contain extreme scores which could skew the data.
- Club matches are from the IPL, PSL, Big Bash, Super Smash, or T20 Blast.
- The full match must have been completed. Abandoned matches and matches resolved with Duckworth-Lewis-Stern are not included. This is because partial matches provide misleading data to the run expectancy matrix described in the next section.

- Male and female matches are considered separately. Throughout this project we will use the male dataset as an example.

The dataset consists of a total of 1905 male and 492 female matches.

In some overs a miscount means the number of balls is different from 6. In the case of an overcount we ignore the balls bowled after the 6th delivery. There are times when a 7 has been scored on a ball. We count this as a 6. These are very rare occurrences and should not affect the results.

It is important to note that throughout this project when we refer to player's statistics we mean within this dataset. They may be different to statistics available online.

Run Expectancy Matrix

In this section we will model a classic problem. How many additional runs is a T20 team expected to score given the state of the game? In T20 cricket the primary factors affecting the number of additional runs to be scored in the innings are the number of balls that have been bowled and the number of wickets that have been taken. The simplest approach would be: For each combination of balls remaining and wickets taken, compute the total number of additional runs scored and divide by the number of occurrences. This is how run expectancy matrices are created for baseball. The difference is that a baseball innings has 24 states while a T20 innings has 1200. Furthermore, while every state occurs somewhat frequently in baseball, some states in T20 hardly occur or have never occurred at all. This means there is not enough data to build a matrix in the same way. An alternative solution is given by the win and score predictor (WASP) model [2]. The solution uses dynamic programming.

Let b be the number of *legal* deliveries already bowled. Let w be the number of wickets that have been taken. Each state of the game can be represented by a pair (b, w). We want to find the expected number of additional runs scored from a given state (b, w). Call this value $V(b, w)$. The number of additional runs expected to be scored is the number of runs expected to be scored on the next ball plus the number additional runs expected to be scored from the next state. The next state could either be $(b + 1, w)$ or $(b + 1, w + 1)$. Let $r(b, w)$ be the number of runs expected to be scored on the next ball when the state is (b, w). Similarly, let $q(b, w)$ be the probability of a wicket on the next ball. We calculate r and q by by diving the total number of runs and wickets on the relevant ball by the number of occurrences. Now we have the following formula for $V(b, w)$:

$$V(b, w) = r(b, w) + q(b, q) \times V(b + 1, w + 1) + (1 - q(b, w)) \times V(b + 1, w).$$

If there are no balls remaining or all 10 wickets have been taken then it is not possible for any additional runs to be scored. This means $V(120, w) = 0$ for all w and $V(b, 10) = 0$ for all b. Now we can create the model by working backwards. Figure 1 demonstrates this.

The advantage of this model is that for each state (b, w), the value of $V(b, w)$ depends only slightly on the values for $r(b, w)$ and $q(b, w)$ and primarily depends on the values of V computed earlier. This means that even in rare scenarios the inputs are dominated by points with sufficient data. Figure 2 shows how the value of $V(b, w)$ depends on the previously computed points, shown in blue. Figure 2: While only V(b+1,w) and V(b+1,w+1) appear directly in the equation, all previously computed points (in the blue area) influence the value of V(b,w).

The values of $V(b, w)$ for $b = 0, \ldots, 119$ and $w = 0, \ldots, 9$ are the entries for our run expectancy matrix.

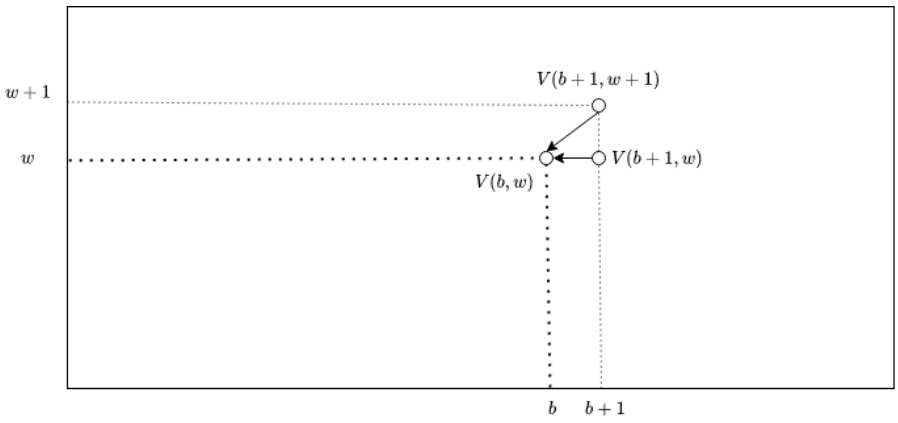

Figure 1: Values of V are computed from previous values.

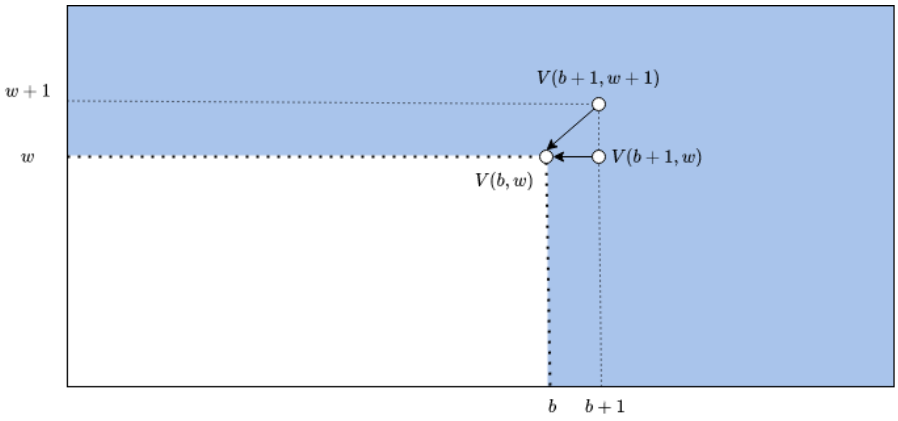

Figure 2: While only $V(b+1,w)$ and $V(b+1,w+1)$ appear directly in the equation, all previously computed points (in the blue area) influence the value of $V(b,w)$.

Weighted Scoring Rate (wSR)

Calculation

The goal of wSR is to work out how the total runs your team is expected to score in the innings changes after each outcome of a legal delivery. Either a $0, 1, 2, \ldots, 6$ or wicket. To do this we use the run expectancy matrix. This method is inspired by baseball's wOBA statistic [3].

Suppose we want to find the expected runs added from a 2. We find all occurrences of a 2 in our dataset. We then find the change in expected total (Δ) with the following formula

$$\Delta = V(b+1,w) - V(b,w) + 2$$

where (b,w) is the state of the game the 2 occurred on. We apply this formula to all occurrences of a 2 in our dataset and divide by the number of occurrences to obtain the expected number of runs added from a 2. We say that x runs are added to mean that the team's expected total runs scored in the innings has increased by x. In our dataset 2s increased expected totals by a total of 37811 runs and 38673 2s were scored. This gives scoring a 2 an expected runs added of 28935/31267 =

0.93. We repeat this for the other possible outcomes with the general formula

$$\Delta = V(b+1, w') - V(b, w) + \text{runs}.$$

where w' is the number of wickets after the delivery and will either be w or $w+1$. We obtain the following.

Outcome	Expected runs added
0	-0.91
1	0
2	0.93
3	2.01
4	2.98
5	4.19
6	4.87
Wicket	-7.03

Note that the weights given here have been rounded to 2dp for ease of presentation. The unrounded weighted are used to calculate wSR in the spreadsheets provided at the end of the article. This means you may get slightly different results if you perform the calculations yourself.

The weights quantify the idea that if you give opportunities for your teammates to do well then you should be assigned some of the positive outcome. Similarly, if you restrict your teammate's opportunities then some of the negative outcome should be attributed to you. To compute a player's wSR we simply multiply the weight for each outcome but the number of times the player achieved that outcome. Finally, we divide by the number of balls the player has faced to create a rate stat. Let's compute David Warner's wSR as an example. In our dataset he has scored 715 0s, 893 1s, 222 2s, 14 3s, 319 4s, one 5, and 103 6s. He has also been out 65 times and faced 2332 balls. His wSR is computed as follows:

$$\begin{aligned} \text{wSR} &= \frac{-0.91 \times 715 + 0.93 \times 222 + 2.01 \times 14 + 2.98 \times 319 + 4.19 + 4.87 \times 103 - 7.03 \times 65}{2332} \\ &= 0.25 \end{aligned}$$

Interpretation

Weighted scoring rate describes how many runs a player adds to their team's expected total, above average, per ball. To get the number of runs a player adds per ball we take their wSR and subtract the dataset wSR. However, the dataset wSR is almost 0 (in our dataset it is 0.0013 to 4 decimal places). This makes sense. The dataset wSR is the amount we expect the final team total to change by after each ball. If the dataset wSR was not 0 then we would expect the final total to change every ball. But then we should have expected a different final total to begin with.

All this is to say that a player's wSR is the rate at which they add runs to the expected team total. David Warner has a strike rate of 141 and a wSR of 0.25. This means that per 100 balls he scores 141 runs and adds 25 runs to his team's total.

While wSR is the number of runs added per ball, it is recommended to present the statistic with a rate of per 100 balls. This is just like strike rate. We will present wSR as per 100 balls throughout the rest of the project. The per ball rate will still be used in calculations.

Purpose

Weighted scoring rate combines aspects of average and strike rate into a single statistic that measures a batters overall contribution to the team total. This idea is explored further in the section on correlations with existing statistics.

Currently we look at player's average and strike rate and use both statistics to form an opinion on their overall ability as a batter. While wSR will provide a similar conclusion in most circumstances, it provides a consistent way of combining the statistics.

Another advantage of wSR is it provides a useful interpretation for the combination. It is one thing to say that David Warner's average/strike rate of 50.45/141 is clearly better than Martin Guptill's 30.45/138. It is more useful to say that Warner adds 25 runs to his team's total per 100 balls while Guptill only adds 13.

Context

The wSR percentiles for male players who have faced at least 200 balls is as follows.

Percentile	wSR Above
Top 75%	-0.11
Top 50%	-0.01
Top 25%	0.09
Top 10%	0.16
Top 1%	0.33

The male player with the best wSR is Finn Allen with an average of 31.16, a strike rate of 179, and a wSR of 45. The female player with the best wSR is Sophie Devine with an average of 42.52, a strike rate of 134, and a wSR of 38.

Runs Created (RC and RC+)

Calculation of RC

In the previous section we saw how wSR describes the number of runs a player adds to their teams total per ball. Now we want to consider the number of runs a player has created. This is the number of runs they added above or below average plus the average. The average number of runs per ball is simply the dataset strike rate (note we do not multiply by 100 in calculations). This leads to the following equation. Here a subscript DS indicates the statistic is for the dataset.

$$\text{RC} = (\text{wSR} - \text{wSR}_{DS} + \text{SR}_{DS}) \times \text{balls faced}$$

We have previously discussed how the dataset wSR is almost 0. You may omit this term if you like. Note that runs created is a counting stat rather than a rate stat. A player's RC depends on the number of balls faced.

Interpretation of RC

A player's RC is the overall number of runs they create for the team. This is different to the number of runs scored. Scoring quickly or rarely going out adds more runs to your team's expected total than you actually score. If a player scores a quick 50 before getting out then the players coming in after them have more balls to face and so can score more runs. So, some of the runs the later players score are attributed to the first player. On the other hand, scoring slowly or going out frequently costs your team runs overall. It is possible for a player's RC to be less than their runs.

Calculation of RC+

To compute RC+ we need to find the ratio between a player's batting performance and the average batting performance. We can do so by finding the ratio between a player's RC per ball and the RC per ball of an average batsman. From above we know that RC per ball has the following equation.

$$\text{RC/ball} = \text{wSR} - \text{wSR}_{DS} + \text{SR}_{DS}$$

The RC of an average batsman is the dataset RC when the dataset is restricted to batsmen, batting allrounders, allrounders, and wicketkeeper batsmen. Bowling allrounders and bowlers are excluded. We also require a minimum of 250 balls faced.

$$\begin{aligned}\text{RC}_{batters}\text{/ball} &= \text{wSR}_{batters} - \text{wSR}_{batters} + \text{SR}_{batters} \\ &= \text{SR}_{batters}\end{aligned}$$

Now we have our equation for RC+

$$\text{RC+} = \frac{\text{wSR} - \textit{dataset}\ \text{wSR} + \textit{dataset}\ \text{SR}}{\textit{batters}\ \text{SR}} \times 100$$

Interpretation of RC+

The main strength of RC+ is how easy it is to interpret. The average batsman has a RC+ of 100 and every point above of below 100 is a 1% increase or decrease. Tom Latham's RC+ of 110 means he creates 10% more runs than the average batsman. Craig Overton's RC+ of 95 means he creates 5% less runs than the average batsman.

Purpose of RC+

Runs created plus is intended to give a single overall evaluation of a batter's performance. The job of a batsman is to create runs and RC+ measures how well they do that. Note that in the equation for RC+ the only variable is the player's wSR. Why don't we create wSR+? Firstly, the dataset wSR is so close to 0 that the scaling does not work. Secondly, wSR+ would overvalue increases in wSR. Adding the dataset SR constant factor dampens the effect of a wSR increase.

Context for RC+

The RC+ percentiles for male players who have faced at least 200 balls is as follows. Note that RC+ measures performance relative to the average batsmen while a number of bowlers are included in the computation of the table.

Percentile	RC+ Above
Top 75%	90
Top 50%	97
Top 25%	105
Top 10%	110
Top 1%	123

Finn Allen has the best RC+ among male players of 131. Sophie Devine has the best RC+ among female players of 130.

Created Against Average (CAA and CAA+)

Calculation of CAA

Created against average (CAA) measures how many runs are created against a bowler per 4 overs bowled (the maximum and typical T20 spell).

We will calculate the number of runs that have been created against a bowler. To do this we start by computing the wSR against the bowler. This is computed in the same way as for batters expect with the runs scored against the bowler. We divide by the number of balls bowled.

We now calculate RC against the bowler in almost the same way as we compute RC for batsmen. The only difference is we add the total extras from the bowler to the RC. This is because extras do not take a wicket (except in extremely rare cases such as stumped off a wide) and do not change the number of deliveries remaining. This means $V(b, w)$ doesn't change. The formula for runs added is: State after minus state before plus runs. So, we have the value of extras is $V(b, w) - V(b, w) + \text{extras} = \text{extras}$. Thus, the following is the formula for RC against.

$$\text{RC against} = (\text{wSR against} - \text{wSR}_{DS} + \text{SR}_{DS}) \times \text{balls bowled} + \text{extras}$$

To find CAA we use the following equation.

$$\text{CAA} = \frac{24 \times \text{RC against}}{\text{balls bowled}}$$

Interpretation of CAA

Created against average measures the average number of runs created against a bowler per 4 overs bowled (per typical match). As CAA is based on wSR against, it combines aspects of a bowler's average and economy. The role of a bowler in a T20 is to avoid conceding runs. Wickets do not matter expect that they lead to less runs conceded. In this way CAA is a complete measure of a bowler's performance. It is intended to be the bowling equivalent of wSR. Created against average is entirely dependent on wSR against but has a nicer interpretation.

Similar to a batsman's RC, the runs created against a bowler may be different from the actual number of runs conceded. If a bowler bowls a maiden then while the actual team total hasn't changed, the expected final total has decreased. In this sense a maiden is worth negative runs rather than 0. The bowler has runs taken off their RC against as a result.

Context for CAA

The CAA percentiles for male players who have bowled at least 200 balls is as follows.

Percentile	CAA Below
Top 75%	35.2
Top 50%	32.9
Top 25%	30.8
Top 10%	29.1
Top 1%	25

Among male players Nasum Ahmed has the best CAA of 22.18. Among female players Frances Mackay has the best CAA of 16.46.

Calculation of CAA+

Computing CAA+ uses similar reasoning as RC+. We first want to find the dataset CAA. It is easy to see that the dataset CAA is 24 times the dataset strike rate + the average number of extras per 24 balls. A lower CAA is better but we want a higher CAA+ to be better. For this reason the dataset CAA is the numerator. If you believe that all bowling stats should be presented so that lower is better then you may flip the numerator and denominator to obtain CAA-.

$$\text{CAA+} = \frac{24 \times \text{SR}_{DS} + \text{extras per 24 balls}_{DS}}{\text{CAA}}$$

Interpretation of CAA+

The interpretation of CAA+ is the same as RC+. A CAA+ of 100 indicates an average bowler and every point above or below is a 1% difference from average. Created against average plus provides an easy way to assess and compare a player's complete bowling performance.

Context for CAA+

The CAA+ percentiles for male players who have bowled at least 200 balls is as follows.

Percentile	CAA+ Above
Top 75%	92
Top 50%	99
Top 25%	106
Top 10%	112
Top 1%	130

Nasum Ahmed has the best CAA+ among male players of 147. Among female players Frances Mackay has the best CAA+ of 163.

Correlations with existing statistics

We will consider the correlation between wSR and CAA with existing statistics. We want to find a strong correlation but not too strong. A weak correlation would suggest these new statistics are meaningless. We also don't want to find perfect correlations. Otherwise, the new statistics would be redundant.

We have the following relationships.

Correlation Between	Pearson's r	Strength
Average and wSR	0.66	Strong
Strike rate and wSR	0.79	Strong
Bowling average and CAA	0.61	Strong
Economy and CAA	0.81	Very Strong

All correlations are strong or very strong without being too strong. This suggests that wSR and CAA are meaningful but not redundant. Furthermore, the aim of wSR is to combine aspects of average and strike rate, and the aim of CAA is to combine aspects of bowling average and economy. The strong correlations suggest the statistics achieve their aims.

Further work

One advantage of RC and CAA is that they are context independent. This makes is easy to get an overall view of player's performance and compare it to other. However, adding some context to the statistics could improve them. The first such context is the stage of the game. The value of outcomes change during power play overs. This could be taken into account. The model also assumes players bat the same in the first innings as the second. This is not a terrible assumption. Strike rate and average are not significantly different between the first and second innings ($p = 0.66$ and $p = 0.79$ respectively). However, the model could be improved by considering the number of runs left to chase. Previous attempts have added runs to chase as a dimension to the model [5]. The extra dimension adds too many points and there is not enough data, even for relatively common states. A more simplistic approach might be to multiply runs scored by a 'pace factor'. For example, the pace factor could be the number of runs left to chase divided by the number of

additional runs expected. If the team is behind in the chase then the pace factor is greater than 1 and runs are worth more. Similarly, if the team is ahead of the chase then the pace factor is less than 1 and runs are less valuable. Refinements to this idea would be required.

Another context that should be taken into account is the ground and conditions. The ground will have a static affect on the ease with which batters score runs. In smaller grounds it is easier to score boundaries. Conditions will have a dynamic affect on scoring that will vary between games. One option for modeling the affect of the ground is to examine how park factor is calculated in baseball sabermetrics [6]. The conditions may need to be modeled from pitch report or weather data. These factors could be integrated with RC+ and CAA+ so that players who frequently play in favorable conditions do not get inflated statistics.

In this project we have computed the statistics across many competitions and many years. It is also interesting to restrict to certain tournaments. For example, we may want to see which batter had the best performance in IPL 2022. To do this we can compute RC+ relative to the season. In the IPL 2022 season Dinesh Karthik had the highest RC+ of 148. This means he created 48% more runs than the average batter in the IPL 2022 tournament. Kane Williamson was paid more than twice as much yet had a RC+ of 74 meaning he was 26% worse than the average batsman. More in depth tournament by tournament analysis could be completed using the new statistics.

Finally, these statistics could be applied to Women's T20s. The weights for Women's T20 are as follows.

Outcome	Expected runs added
0	-0.7
1	0.19
2	1.13
3	2.17
4	3.16
5	4.4
6	5.04
Wicket	-6.34

A problem is that there is only a quarter as much data for Women's T20s. This makes some analysis, especially the additional analysis suggest earlier in this section, less accurate. Overcoming these challenges is an opportunity for future work.

Conclusion

We have introduced new statistics for analysing T20 performance. The batting statistics wSR and RC+ combine aspects of a player's average and strike rate to provide a complete picture of their batting performance. Similarly, the bowling statistics CAA and CAA+ combine aspects of a bowler's average and economy to provide a complete overview of their bowling performance. The 'plus' stats RC+ and CAA+ are easy to understand and provide a way to compare the performance of players. While more can be done to refine these statistics, they are already a great description of a player's overall performance.

Data

A link to CSV files containing the new statistics for both women's and men's T20s is given below. Note that players who have faced or bowled few balls can have strange results. It is recommended to impose a minimum number of balls faced or balls bowled before working with the statistics.

CSV files

References

[1] Cricsheet ball by ball information, https://cricsheet.org/matches
[2] Seamus Hogan, Cricket and the Wasp (2012),
http://offsettingbehaviour.blogspot.com/2012/11/cricket-and-wasp-shameless-self.html
[3] Piper Slowinski, wOBA (2010), https://library.fangraphs.com/offense/woba/
[4] Piper Slowinski, wRC and wRC+ (2010), https://library.fangraphs.com/offense/wrc/
[5] Pranav Sodhani, WASP Second Innings (2016),
https://github.com/sodhanipranav/WASP/blob/master/Second%20Innings.pdf
[6] Neil Weinberg, The Beginners Guide to Understanding Park Factors (2015),
https://library.fangraphs.com/the-beginners-guide-to-understanding-park-factors/