©Journal of Sports Science and Medicine (2006) **5,** 503-508
http://www.jssm.org

Research article

The 8th Australasian Conference on Mathematics and Computers in Sport, 3-5 July 2006, Queensland, Australia

STOCHASTIC DOMINANCE AND ANALYSIS OF ODI BATTING PERFORMANCE: THE INDIAN CRICKET TEAM, 1989-2005

Uday Damodaran

XLRI Jamshedpur School of Management, Jamshedpur, Jharkand, India

Published (online): 15 December 2006

ABSTRACT

Relative to other team games, the contribution of individual team members to the overall team performance is more easily quantifiable in cricket. Viewing players as securities and the team as a portfolio, cricket thus lends itself better to the use of analytical methods usually employed in the analysis of securities and portfolios. This paper demonstrates the use of stochastic dominance rules, normally used in investment management, to analyze the One Day International (ODI) batting performance of Indian cricketers. The data used span the years 1989 to 2005. In dealing with cricketing data the existence of 'not out' scores poses a problem while processing the data. In this paper, using a Bayesian approach, the 'not-out' scores are first replaced with a conditional average. The conditional average that is used represents an estimate of the score that the player would have gone on to score, if the 'not out' innings had been completed. The data thus treated are then used in the stochastic dominance analysis. To use stochastic dominance rules we need to characterize the 'utility' of a batsman. The first derivative of the utility function, with respect to runs scored, of an ODI batsman can safely be assumed to be positive (more runs scored are preferred to less). However, the second derivative needs not be negative (no diminishing marginal utility for runs scored). This means that we cannot clearly specify whether the value attached to an additional run scored is lesser at higher levels of scores. Because of this, only first-order stochastic dominance is used to analyze the performance of the players under consideration. While this has its limitation (specifically, we cannot arrive at a complete utility value for each batsman), the approach does well in describing player performance. Moreover, the results have intuitive appeal.

KEY WORDS: Bayesian, utility function, batting average, conditional average, geometric distribution.

INTRODUCTION

As a game, cricket is a statistician's delight. Each game of cricket throws up a huge amount of performance related statistics. As other games have evolved and developed, they too have become richer in the use of performance statistics. For example, use of statistics like 'unforced errors' in lawn tennis or 'assists' in basketball is increasingly becoming popular. However in cricket these statistics have *always* been part and parcel of the game. Cricket is one of the few games in which a 'scorer' is required to continuously maintain statistical data on key game/player-specific performance statistics. It is one of the few games that have detailed 'scoring sheets'. These scoring sheets were maintained manually in the pre-digital age and are maintained electronically today.

In spite of this legacy and long history of maintaining statistical data, two aspects associated with cricketing data are striking. The first is the idiosyncrasy that has persisted in the treatment of the 'not out' scores of a player. The second is the lack of effort in exploiting the richness of data to improve the representation of player performance.

The batting average of player *i,* $R_{i,}$ is computed as:

$$R_i = \Sigma Rit / (n\text{-}k) \qquad (1)$$

Where: R_{it} is the number of runs scored by the *i* th player in the *t* th innings; *n* is the total number of innings in which the *i* th player has batted and *k* is the number of innings in which the *i* th player has remained 'not out'.

Equation (1) introduces an upward bias in the average. This bias is caused because the numerator is the total runs scored over all innings while the denominator excludes the innings in which the player has remained 'not out'. This bias cannot seemingly be avoided. Taking the denominator to be *n* instead of *n-k* would instead introduce a downward bias in the average. A similar problem arises while preparing the input data required for the stochastic dominance rules developed later in the paper. The input data that is required is the innings-by-innings runs scored by the player. What should be done with the scores for the innings in which the player has remained 'not out'? This paper first proposes a method to deal with this problem.

The second aspect of cricketing data is the scant attention that has been focused by researchers on certain aspects of cricket. A substantial portion of the work has focused on devising optimal playing strategies. The strategies studied have either focused on batting strategies (Clarke, 1988; Clarke and Norman, 1999; Preston and Thomas, 2000; Swartz et al., 2006) or on bowling strategies (Rajadhyaksha and Arapostathis, 2004). A fair amount of work has also focused on the problem of arriving at a fair result when a game has to be prematurely terminated due to weather conditions or other disturbances (Duckworth and Lewis, 1998; Preston and Thomas, 2002; Carter and Guthrie, 2004).

The third stream of work, on the understanding and development of player-specific performance statistics, (Kimber and Hansford, 1993; Lemmer, 2004; Lewis, 2005; Wood, 1945), has received little attention. Cricket, with its slow pace and non-continuous nature, is a very television-friendly game. It allows viewers the leisure of watching replays without impinging on real-time action. It thus allows for the presentation of a vast amount of descriptive statistics during the course of a game. In spite of this feature of the game and the long history of the game, cricket commentators sometimes seem to feel constrained by the inability of performance statistics to really describe player performance. Comments like "Statistics don't say everything" are very commonly heard. The attention devoted by researchers to this aspect of cricket, therefore, seems surprisingly scant in relation to its importance and relevance.

This paper seeks to develop methods to assess the performance of batsmen in cricket that (i) makes use of more information than current methods do and (ii) can be converted into visually appealing graphics for the television medium. The method is demonstrated using player statistics for the some of the key members of the Indian One Day International (ODI) cricket team between 1989 and 2005. The names of the players included in the study are listed in Table 1.

Table 1. Names of players included in the study.

Serial Number	Player
1	S. Tendulkar
2	A. Kumble
3	S. Ganguly
4	R. Dravid
5	A. Agarkar
6	V.V.S. Laxman
7	H. Singh
8	V. Sehwag
9	Z. Khan
10	Y. Singh
11	A. Nehra
12	M. Kaif
13	M. Karthik
14	S. Bangar

METHODS

The primary measure of a batsman's performance in cricket today is the player's batting average defined as in Equation (1). This measure suffers from the shortcoming that it is a one-dimensional number and does not capture the richness of the underlying data. Though cited very often, this measure fails to capture the various facets of a batsman. It does not provide answers to many questions that arise during the course of a game. These un-addressed concerns or questions feature often in the comments of cricket commentators. For example, commentators of the game are found to say "Player X is a dangerous player once he is set". Or "Player X has the ability to convert a good start to a big score". Or "Though his average does not reflect it, Player X is a more consistent performer than Player Y".

Adjusting the raw data

The raw data used in the development of any method for representing a batsman's performance are the innings-by-innings runs scored by the player. However, using this raw data poses a problem. In some of the innings the batsman would not have been dismissed. In such cases the score would not reflect the number of runs the player could potentially have gone on to score. The scores for these innings (the ‘not out’ situations) have thus to be replaced by a number that is a good estimate of the number of runs the player would have scored had he batted on.

In an early work Wood (1945) had provided empirical support to support the claim that a batsman’s scores follow a geometric distribution. Under this assumption, because of the memoryless property of the geometric distribution, a batsman’s chance of getting out is independent of the score he is on. This assumption can be used to arrive at an estimate of the number of runs a ‘not-out’ player would have scored had he batted on. However, the assumption of a geometric distribution for a batsman’s scores might not hold for all players. There may be some players who are ‘slow starters’ and who therefore do better as they progress. There may be other players who become more adventurous as their score increases. For such adventurous players their chances of getting out might increase as their score increases.

Kimber and Hansford (1993) did consider deviations from the geometric distribution, but their focus was on arriving at an optimal estimator for the population mean. On the other hand, we need a method to arrive at an estimate of the number of runs a ‘not out’ batsman would have gone on to score. A Bayesian approach has been adopted in this paper to arrive at this estimate. This is achieved in the following manner.

Assume that in his j th innings player i remains ‘not out’ on a score of R_{ij}. Define a binary variable G_{rik} such that:

$$G_{rik} = 0 \text{ if } R_{ik} < R_{ij} \text{ and } = 1 \text{ if } R_{ik} >= R_{ij} \text{ for } k = 1, 2, \ldots j-1 \qquad (2)$$

$$\text{Define } n_{ij} = \sum_{k=1}^{j-1} G_{rik} \qquad (3)$$

$$\text{Define } C_{ik} = 0 \text{ if } R_{ik} < R_{ij} \text{ and } = R_{ik} \text{ if } R_{ik} >= R_{ij} \text{ for } k = 1, 2, \ldots j-1 \qquad (4)$$

The estimate of the number of runs that the ‘not-out’ batsman would have gone on to score is then given by:

$$E_{ij} = \sum_{k=1}^{j-1} C_{ik} / n_{ij} \qquad (5)$$

In other words, the estimator used for the runs that the ‘not out’ batsman would have gone on to score is the conditional average of the batsman at that point of time, given that he has already scored a certain number of runs. In every instance of a ‘not out’, the batsman’s score in that innings j is replaced by the estimate E_{ij}. This approach has the advantage of handling deviations from the geometric distribution assumption. It is also information efficient, with the posterior values of the conditional average incorporating more information on the batsman’s performance. Table 2 gives an example of the computational procedure used for finding the replacement values for the first two ‘not outs’ in the career of one member of the Indian ODI team, Sachin Tendulkar.

Table 2. Replacing ‘Not Out’ scores with estimates of runs likely to have been scored by a batsman: Sachin Tendulkar’s first 15 ODI innings.

Sl. No.	Runs Scored	Whether Dismissed	Adjusted Score	Remarks
1	0	Y	0	
2	0	Y	0	
3	36	Y	36	
4	10	Y	10	
5	20	Y	20	
6	19	Y	19	
7	31	Y	31	
8	36	Y	36	
9	53	Y	53	
10	30	Y	30	
11	4	Y	4	
12	53	Y	53	
13	52	N	53	Average of Sl. Nos. 9, 12
14	22	Y	22	
15	11	N	35	Average of Sl. Nos.3,5,6,7,8 9,10,12, 13 (adjusted), 14

Stochastic Dominance

The adjusted raw data is now used to arrive at an analytical representation of the player’s batting performance. The approach adopted draws from methods normally used for the analysis of securities and portfolios in investment management.

The focus in investment management is on wealth creation. The problem of portfolio choice is that of selecting a portfolio that maximizes the utility for the investor. The utility function for the investor attaches a utility to various levels of wealth. The utility function can be constrained to have certain properties like non-satiation (more wealth is always preferred to less) or risk aversion (diminishing marginal utility for incremental units of wealth). In mathematical terms the first constraint requires the first derivative of the utility function to be positive. Again, in mathematical terms the second constraint

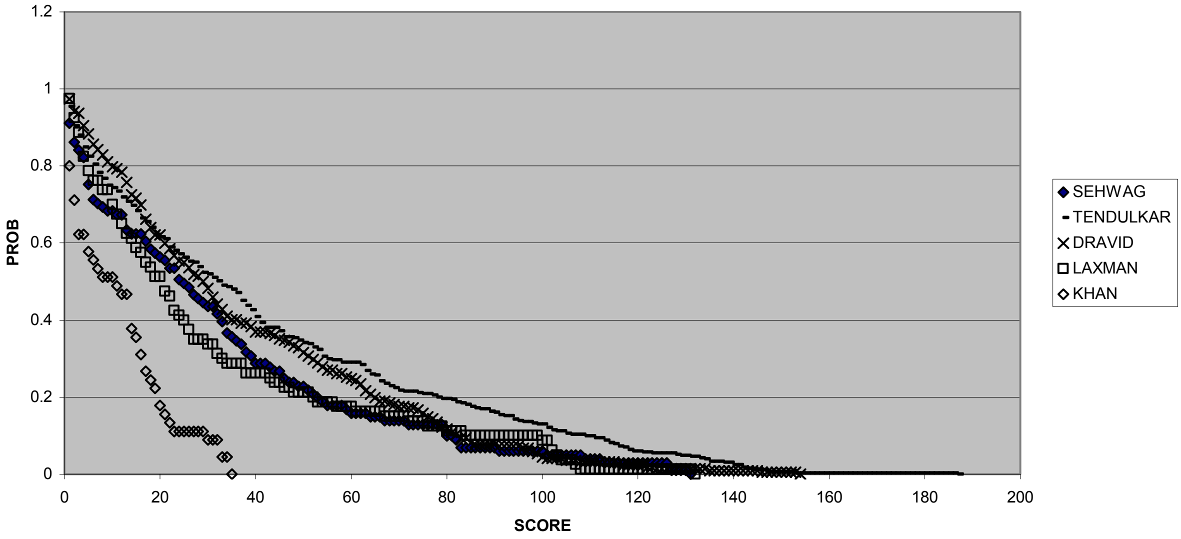

Figure 1. Sample stochastic dominance curves.

requires the second derivative of the utility function to be negative.

Consistent with some of the above-listed features of utility functions, the traditional approach to the portfolio selection problem has been the mean-variance approach. Amongst the alternative approaches to the portfolio selection problem suggested in the investment management literature is the set of stochastic dominance rules (Ali, 1975; Bawa, 1978; Levy, 1973). To use stochastic dominance rules we need to characterize the utility function of the investor. According to the first-order stochastic dominance rules a portfolio A is preferred to another portfolio B if, for any level of return, the cumulative probability of portfolio A giving a return lesser than the given level of return is never greater, and sometimes less, than the cumulative probability of portfolio B giving a return lesser than that given level of return. This rule is consistent with the assumption that in the investors' utility function more wealth is preferred to less. (Elton and Gruber, 1997)

Analogous to the portfolio selection problem, a similar approach is adopted in this paper to represent the batting performance of cricketers. Using this approach we can say that a batsman A's performance is better than another batsman B's if, for any level of score, the probability of batsman A getting a score greater than the given score is never lesser, and sometimes greater, than the probability of batsman B getting a score greater than that given score. This rule corresponds to the first-order stochastic dominance rules and assumes that more runs are always preferred to less.

The cumulative probability charts of various batsmen can now be charted with runs on the X-axis (with the origin as zero) and the probability of scoring more runs than the X-axis value of the score (that is one minus the cumulative probabilities of scoring lesser than the X-axis value of score) on the Y-axis. Visually this would mean that a batsman whose stochastic dominance curve envelops another's curve stochastically dominates the other batsman.

RESULTS

The method is demonstrated using data for the Indian ODI cricket team spanning the years 1989 (the year one of India's most highly rated players, Sachin Tendulkar, made his debut) to 2005. This period was chosen because this was a period during which the compositional changes in the Indian ODI team were very few. A sample batting performance stochastic dominance chart output for five Indian players is given in Figure 1.

Four of the five players represented are essentially specialist batsmen (Tendulkar, Dravid, Sehwag and Laxman) and one a specialist bowler (Khan). The results are interesting and have intuitive appeal. They are consistent with popular notions regarding the batsmen whose performances were studied. For example, the curve for Sachin Tendulkar, who is considered an icon of Indian cricket, almost completely envelops the curves for other players. And the curve for Rahul Dravid, who is referred to as 'the wall' because of his perceived consistency, does indeed dominate the curves for other players till the 20 run point. In other words, the chances of Rahul Dravid getting a score less than 20 is lesser than the chances for any other player in the Indian team getting a score lesser than

20. Finally, the curves for the specialist batsmen very clearly dominate the curves for the specialist bowlers, as should be the case.

DISCUSSION

The method that has been developed only provides an alternative approach to represent the batting performance of cricket players. This alternative approach is visually and intuitively appealing. The attempt in this paper is not to arrive at a model to rank the utility of players. Nor is the goal to develop a model to assist in team selection. The utility of a player goes far beyond the runs scored by him. Factors like tactical skills, passive support to the partner batsmen, etc. cannot be gauged by looking at the runs scored. Even if we use runs scored as the sole measure of utility, first-order stochastic dominance rules alone cannot be used to rank players in terms of their utility. And if we go on to second-order stochastic dominance rules the utility function might not have a negative second derivative. In other words, there could be potentially match winning situations in which a batsman who is batting on a very high score (say, 108) has to score one more run in order for the team to win the match. In this situation the incremental one run (from 108 to 109) might be much more valuable than the incremental one run the batsman scored while he was on a lower score (say, 23) during the same innings.

CONCLUSIONS

Within the limits of this study, the paper seeks to highlight the tremendous scope that exists to improve and develop on the measures currently used to describe the performances of cricket players in general, and batsmen in particular. The measures used today do not adequately capture the richness of the underlying data. Similar approaches can be adopted to represent the performances of bowlers too.

REFERENCES

Ali, M. (1975) Stochastic Dominance and Portfolio Analysis. *Journal of Financial Economics* **2**, 205-230

Bawa, V. (1978) Safety-first, stochastic dominance, and optimal portfolio choice. *Journal of Financial and Quantitative Analysis* **XIII**, 225-271

Carter, M. and Guthrie, G. (2004) Cricket interruptus: fairness and incentive in limited overs cricket matches. *Journal of the Operational Research Society* **55,** 822-829.

Clarke S.R. (1988) Dynamic Programming in one-day cricket- optimal scoring rates. *Journal of the Operational Research Society* **39,** 331-337

Clarke S.R. and Norman J.M. (1999) To run or not?: Some dynamic programming models in cricket. *Journal of the Operational Research Society* **50,** 536-545.

Duckworth, F.C. and Lewis, A.J. (1998) A fair method of resetting the target in interrupted one-day cricket matches. *Journal of the Operational Research Society* **49**, 220-227.

Elton, E.J. and Gruber, M.J. (1997) *Modern portfolio theory and investment analysis*. 5th edition. John Wiley and Sons.

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**, 443-455.

Lemmer, H.H. (2004) A measure for the batting performance of cricket players. *South African Journal for Research in Sport, Physical Education and Recreation* **26,** 55-64

Levy, H. (1973) Stochastic dominance, efficiency criteria, and efficient portfolios: The multi-period case. *American Economic Review* **LXIII**, 986-994

Lewis, A.J. (2005) Towards fairer measures of player performance in one-day cricket. *Journal of the Operational Research Society* **56,** 804-815.

Preston, I. and Thomas, J. (2000) Batting strategy in limited overs cricket. *Journal of the Royal Statistical Society: Series D (The Statistician)* **49**, 95-106.

Preston, I. and Thomas, J. (2002) Batting strategy in limited overs cricket. *Journal of the Royal Statistical Society: Series D (The Statistician)* **51**, 189-202.

Rajadhyaksha, G. and Arapostathis, A. (2004) Cricket strategy: Using a Bayesian network to recommend the best bowling option. In: *The Engineering of Sport V, Vol 2.*. Eds: Hubbard M., Mehta R.D. and Pallis J.M. International Sports Engineering Association. 425-431.

Swartz, T.B., Gill, P.S., Beaudoin, D. and de Silva, B.M. (2006) Optimal batting orders in one-day cricket. *Computers and Operations Research* **33,** 1939-1950.

Wood, G.H. (1945) Cricket scores and geometric progressions. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **108**, 12-22.

KEY POINTS

- The problem of dealing with ‘not out’ scores in cricket is tackled using a Bayesian approach.
- Stochastic dominance rules are used to characterize the utility of a batsman.
- Since the marginal utility of runs scored is not diminishing in nature, only first order stochastic dominance rules are used.
- The results, demonstrated using data for the Indian cricket team are intuitively appealing.
- The limitation of the approach is that it cannot arrive at a complete utility value for the batsman.

AUTHORS BIOGRAPHY

Uday DAMODARAN

Employment

Professor of Finance, XLRI School of Business and Human Resources, Jamshedpur, Jharkand, India.

Degree

Fellow (IIM Bangalore), M.Stat., B.Sc. (Physics)

Research interests

Investments, Corporate Finance, Cricket

E-mail: uday@xlri.ac.in

🖂 **Uday Damodaran**
XLRI Jamshedpur School of Management, Jamshedpur,
Jharkand, India