CRICKETING CHANCES

G. L. Cohen
Department of Mathematical Sciences
Faculty of Science
University of Technology, Sydney
PO Box 123, Broadway
NSW 2007, Australia
graeme.cohen@uts.edu.au

Abstract

Two distinct aspects of the application of probabilistic reasoning to cricket are considered here.

First, the career bowling figures of the members of one team in a limited-overs competition are used to determine the team bowling strike rate and hence the probability of dismissing the other team. This takes account of the chances of running out an opposing batsman and demonstrates that the probability of dismissing the other team is approximately doubled when there is a good likelihood of a run-out.

Second, we show that under suitable assumptions the probability distribution of the number of scoring strokes made by a given batsman in any innings is geometric. With the further assumption (which we show to be tenable) that the ratio of runs made to number of scoring strokes is a constant, we are able to derive the expression $(A/(A+2))^{c/2}$ as the approximate probability of the batsman scoring at least c runs ($c \geq 1$), where A is the batsman's average score over all past innings.

In both cases, the results are compared favourably with results from the history of cricket.

1 Introduction

In an excellent survey of papers written on statistics (the more mathematical kind) applied to cricket, Clarke [2] writes that cricket "has the distinction of being the first sport used for the illustration of statistics", but: "In contrast to baseball, few papers in the professional literature analyse cricket, and two rarely analyse the same topic."

This paper analyses two aspects of cricket. The first is an apparently novel investigation of bowling strike rates to determine the probability of bowling out the other team in one-day cricket. The likelihood of running out one or more of the opposing batsmen is then incorporated for greater accuracy, and leads to the useful conclusion that the probability of dismissing the other team is approximately doubled when there is the likelihood of at least one run-out. These ideas were developed in the papers Cohen [4, 5], and are presented here using an improved model and additional comments. (The opportunity is also taken to make some minor corrections to the earlier papers.)

The second aspect, quite distinct from the first and not previously written up, is a further discussion of a topic described in Clarke [2]. It concerns the distribution of scores in the traditional game. Rather than seek directly a probability distribution for the number of runs scored by a particular batsman, we derive instead the distribution of the number of scoring strokes. Scoring strokes are related to the number of runs scored by assuming the ratio of these quantities to be (approximately) constant. Having a probability distribution for the number of scoring strokes then allows the probability to be determined of a batsman scoring a century, say, even if he has not previously made such a score.

2 An application of bowling strike rates

We begin by showing that the strike rates of the bowlers on one team allow an estimate to be made of the probability of getting an opposing batsman out in some manner that is credited to the bowler (so we exclude run-outs for the moment). For one-day cricket, where there is a limit to the number of balls to be bowled in an innings, this can be used to obtain the probability of getting the whole team out. When we include the possibility of run-outs, we get a much better estimate for the probability of dismissing the other side, as confirmed by comparisons with actual results from cricket's World Cup.

Let b and w stand, respectively, for the number of balls bowled by a certain bowler (excluding wides and no-balls) and the number of wickets taken from his bowling in a season, or in his career, or against a particular team, say. Then that bowler, for our purposes, has a strike rate given by b/w. If $w = 0$ (which is hardly likely, for our purposes) then the bowler is deemed not to have a strike rate. A bowler's strike rate, along with his average and his economy rate (neither of which is used here by us), are in common use when analysing the effectiveness of various bowlers; the better bowler has the smaller strike rate. The reciprocal of the strike rate can be interpreted as the probability that the bowler subsequently takes a wicket with each ball bowled.

For a complete team, suppose we have n bowlers so that, in one-day competition, $5 \le n \le 11$. Let their strike rates based on previous experience be s_k, for $k = 1, \ldots, n$. If the kth bowler is to bowl b_k balls in a coming match (excluding wides and no-balls), then experience suggests he will take w_k wickets, where

$$w_k = \frac{b_k}{s_k}, \qquad \text{for } k = 1, \ldots, n.$$

Let B be the total number of balls to be bowled by that team in the match (excluding wides and no-balls), and let W be the total number of wickets taken. Then

$$B = \sum_{k=1}^{n} b_k \qquad \text{and} \qquad W = \sum_{k=1}^{n} w_k = \sum_{k=1}^{n} \frac{b_k}{s_k},$$

and the team's strike rate S for that match may be predicted to be

$$S = \frac{B}{W} = \sum_{k=1}^{n} b_k \Big/ \sum_{k=1}^{n} \frac{b_k}{s_k}. \tag{1}$$

This is a weighted harmonic mean of $s_1, \ldots, s_n$, the weights being $b_1, \ldots, b_n$.

For example, if four bowlers are to bowl ten overs each, and two others five overs each, then $b_1 = \cdots = b_4 = 60$, $b_5 = b_6 = 30$, and

$$S = \frac{10}{\frac{2}{s_1} + \frac{2}{s_2} + \frac{2}{s_3} + \frac{2}{s_4} + \frac{1}{s_5} + \frac{1}{s_6}}.$$

A bowling combination that allows $S \le 30$, since $B \le 300$ and $W = 10$ in a completed innings, would be most desirable, though rarely achievable in practice.

Typically good individual strike rates satisfy $25 \le s_k \le 50$. If selectors were to choose only that combination of bowlers that allows S to be least, then they would accomplish this by taking the five bowlers with smallest strike rates. However, economy rates or bowling averages and batting and "all round" skills would also all be taken into account, and the captain has his tactical considerations, so it is usually necessary to have a much more varied bowling attack. It would be useful to know then what chances the various bowling combinations have of bowling the other side out.

The quantity $p = 1/S$ represents wickets per ball during the opposing team's innings of at most 50 overs and is the empirical probability with each ball of a bowler taking a wicket, by any of the means

that allow a wicket to be credited to the bowler. The probability of the bowlers taking w wickets in 50 overs is

$$\binom{300}{w} p^w q^{300-w},$$

where $q = 1 - p$, on the assumption that each ball bowled is an independent event. It was argued in [4] that the other team has not been bowled out if $w \leq 9$, so the probability of bowling them out is

$$P_1 = 1 - \sum_{w=0}^{9} \binom{300}{w} p^w q^{300-w}. \tag{2}$$

A more detailed analysis is given in [4] in terms of the bowlers' individual strike rates, but a numerical argument there shows that, for practical purposes, it is sufficient to make use of (2).

However, because $P_1 = \sum_{w=10}^{300} \binom{300}{w} p^w q^{300-w}$, the use of (2) would seem to suggest that the rules of the game in fact allow for ten or more wickets to be taken, but that the game is to be abandoned after ten wickets, the others being defaulted. This whimsy is avoided with the following alternative approach.

To bowl the other team out, ten wickets must be taken and this may be done in anything from ten to 300 balls. If k balls are required, $10 \leq k \leq 300$, then the tenth wicket must be taken with the kth ball, and the first nine wickets with any of the first $k - 1$ balls. No wicket is taken with the remaining $k - 10$ balls. Hence the probability of bowling the other team out is

$$P_2 = \sum_{k=10}^{300} \binom{k-1}{9} p^{10} q^{k-10} = \left(\frac{p}{q}\right)^{10} \sum_{k=10}^{300} \binom{k-1}{9} q^k.$$

It is reassuring to calculate that values of P_1 and P_2 are, for practical purposes, very close. They agree to four decimal places for S up to 42. In fact, for integer values of S, the greatest difference $P_1 - P_2$ is 0.00335, at $S = 85$. (Always, $P_1 > P_2$ since P_1 is, whimsically, the probability of taking ten or more wickets in 300 balls.)

Coincidentally, the article [4] appeared just as the 1999 World Cup of one-day cricket was about to get under way in England, and it was noticed by the science writer in *The Times*. He wrote a column [12] describing the ideas above and giving his own calculations regarding the English team. Two days later, on the morning of the first match in the World Cup and having seen the English article, *The Australian* [16], in more journalistic style, prevailed upon the author to rank the twelve competing teams in order of the probabilities of bowling their opponents out, even though bowling out the other team does not ensure a win. These probabilities were compared with odds then being offered for each team, and so the ideas in [4] were promoted to a level somewhat above the original conception. (The author's top six ranked teams included five that made the Super Six, who then played off to determine the finalists. This is praiseworthy but not relevant.)

The calculation of S in (1) is described above as being for predictive purposes, based on bowlers' strike rates prior to a match. It may subsequently be compared with the strike rate actually attained in an innings, calculated as the number of balls bowled (not including wides and no-balls) divided by the number of wickets credited to the bowlers. It seems to be standard, if perhaps wrong, that wides and no-balls are not included when determining bowlers' strike rates, so we follow that practice. Moreover, a team's actual overall bowling performance may be based on calculations made following a series of games, such as in the World Cup. It was apparent to the author that the probabilities based on the model in (2) and actual games played in the 1999 World Cup, underestimated the proportion of times that each team in fact bowled out its opponents.

One presumed reason for this was clear: "bowling out" the opponents (as we will use the term) is not the same as "dismissing" them, since the latter includes wickets lost by batsmen who are run out. These are not credited to the bowler. In [5], the methods of [4] were made more realistic by allowing for run-outs, including the possibility that run-outs may occur off wides and no-balls.

World Cup	Matches (α)	Good balls (β)	Bad balls (γ)	Bowled out (δ)	Run out (ϵ)	Teams dismissed (ζ)
1987	27	15413	363	321	64	12
1992	37	20206	661	439	67	18
1996	35	19461	508	411	63	14
1999	42	22721	1218	549	49	27
Totals	141	77801	2750	1720	243	71

Table 1: Data from previous four World Cups.

All match results from the preceding four World Cups (in the years 1987, 1992, 1996 and 1999) were scanned to arrive at estimates for the probability of a run-out with each ball bowled and the average number of wides and no-balls in a 50-over innings. (The World Cups prior to 1987 were 60 overs a side, and not considered for that reason, although the model could be easily adjusted to take this into account.)

The resulting data are given in Table 1. We use the term "good ball" for any delivery not resulting in a wide or no-ball, and "bad ball" for a wide or no-ball. Wickets resulting from good balls, but not run-outs, are credited to the bowler. A batsman can be run out from any ball, good or bad. There are other means of getting out off bad balls (such as being stumped off a wide, in which case the wicket is credited to the bowler), or off good balls with the result not credited to the bowler, but these are very rare and ignored for our purposes. The columns in Table 1 are labelled α, β, ..., ζ for later use.

The 9th and 13th matches in the 1992 World Cup were abandoned due to rain, and the 5th and 14th matches in the 1996 World Cup were forfeited. These have not been included in Table 1. The 16th match in 1996 was replayed after the first attempt was washed out, and only the replayed match has been included. It is possible that some of the figures for balls bowled, both good and bad, in Table 1 may be off by a few from the true numbers, since, for example, umpires' errors (such as allowing a few seven-ball overs) and rule changes for the 1999 World Cup that allowed penalty runs have not always been easy to take into account. We have used the scorecards from CricInfo at www.cricket.org. "Bowled out" refers to wickets credited to bowlers, and in this table includes batsmen who retired hurt or were absent ill, so that they may be taken into account in determining overall bowling strike rates.

Suppose, in a completed innings of 50 overs, there are y bad balls bowled. Re-define B by $B = 300+y$, the total number of balls bowled, so that the probability of any particular ball being good is $300/B = g$, say. Since the strike rate S is based only on good balls bowled, we can give the actual probability of a bowler taking a wicket as gp, where $p = 1/S$, as before. Let r be the probability of a run-out with each ball bowled. Then the probability that the bowlers take w_b wickets, and that a further w_r batsmen are run out, follows a multinomial distribution. It is

$$\binom{B}{w_b, w_r, B-w_b-w_r}(gp)^{w_b} r^{w_r} (1-gp-r)^{B-w_b-w_r}$$
$$= \frac{B!}{w_b!\, w_r!\, (B-w_b-w_r)!}(gp)^{w_b} r^{w_r} (1-gp-r)^{B-w_b-w_r},$$

where, in practice, $0 \le w_b + w_r \le 10$, assuming that each ball bowled is an independent event. If $w_b + w_r \le 9$, then the team has not been dismissed, so the probability of dismissing the other side is

$$P_3 = 1 - \sum_{0 \le w_b + w_r \le 9} \frac{B!}{w_b!\, w_r!\, (B-w_b-w_r)!}(gp)^{w_b} r^{w_r} (1-gp-r)^{B-w_b-w_r}.$$

When $y = 0$ and $r = 0$, so that $g = 1$ and $w_r = 0$, this reduces to the result in (2) (with the understanding that then $r^{w_r} = 1$).

Although numerically accurate, this formula has the same conceptual drawback as for P_1. Instead, we may argue as follows.

Let w be the number of wickets taken by the bowlers, $0 \le w \le 10$, so that $10 - w$ is the number of run-outs. The taking of wickets and the bowling of balls are considered to be independent events, except that a wicket must be taken with the last of k balls bowled, $10 \le k \le B$. Then the probability of dismissing the other side is

$$P_4 = \sum_{w=0}^{10} \binom{10}{w} (gp)^w r^{10-w} \cdot \sum_{k=10}^{B} \binom{k-1}{9} (1 - gp - r)^{k-10}. \tag{3}$$

When $r = 0$, we must have only the summand with $w = 10$ and then, as above, must interpret 0^0 as 1. With $y = 0$, then P_4 reduces to the expression for P_2.

We can now demonstrate that this model approximates well the actual results from the four World Cups.

World Cup	S	r	y'	Π	P
1987	48.016	0.00406	7.07	0.222	0.220
1992	46.027	0.00321	9.81	0.243	0.222
1996	47.350	0.00315	7.83	0.200	0.199
1999	41.386	0.00205	16.08	0.321	0.268
Combined	45.233	0.00302	10.60	0.252	0.229

Table 2: Actual proportion (Π) of teams dismissed and predicted probability (P) of dismissing a team, given Cup bowling strike rate (S), run-outs per ball (r), and average number of wides and no-balls (y').

From the data given in Table 1, we may calculate combined bowling strike rates $S = \beta/\delta$ for each World Cup, the proportion $r = \epsilon/(\beta + \gamma)$ of run-outs, and the average number $y' = 300\gamma/\beta$ of bad balls in a 50-over innings. We also have from Table 1 the actual proportion $\Pi = \zeta/(2\alpha)$ of teams dismissed. We put $p = 1/S$, r and y (equal to y', rounded to the nearest integer) into (3) to obtain the values $P = P_4$ in Table 2. Compare the values of P and Π.

Notice that the values for r in Table 2 show that there are on average about three run-outs per 1000 balls in world class one-day cricket, which equates to about one per innings of 50 overs. The values for y' show that there are, say, seven to ten wides or no-balls altogether in a 50-over innings. (The 1999 World Cup seems to be exceptional in the latter regard—this was the time when accusations of corrupt practice in cricket were rife and in many cases subsequently shown to be justified, and perhaps here we see some evidence for the accusations.)

Finally in this section, we give Table 3. For team bowling strike rates S from 20 to 62, incremented by 2, and three values (0.002, 0.003 and 0.004) for the probability r of a run-out with each ball (pick the probability that matches the team's fielding skills or the opponents' lapses in running), we give the probability of dismissing the other team. We have taken $y = 10$, although it turns out that, whether $y = 0$ or $y = 20$, the computed values are rarely affected even in the second decimal place. (Because P_4 is used rather than P_3, Table 3 differs in a few entries, but not at all substantially, from the corresponding table in [5].)

The first row of Table 3, with $r = 0$, corresponds to the probability of dismissing the opponents if run-outs are not to be considered. This should still be seen as important to assist a captain or selector to estimate the ability of their chosen team to bowl out the opponents (the theme of the paper [4]), as other considerations would not then be taken into account. Perhaps of more interest is a comparison between the entries corresponding to $r = 0$ (no run-outs) and $r = 0.003$ (close enough to one run-out in 300 balls) for $46 \le S \le 60$ (a range for the team bowling strike rate that would be common in practice). They may be interpreted as showing that the probability of dismissing the other team is approximately doubled if there is a good likelihood of a run-out.

r	S										
	20	22	24	26	28	30	32	34	36	38	40
0	0.93	0.88	0.80	0.72	0.63	0.54	0.46	0.39	0.32	0.27	0.22
0.002	0.95	0.91	0.85	0.78	0.70	0.62	0.54	0.47	0.41	0.35	0.30
0.003	0.96	0.92	0.87	0.80	0.73	0.66	0.58	0.51	0.45	0.39	0.34
0.004	0.97	0.93	0.88	0.82	0.76	0.69	0.62	0.55	0.49	0.43	0.38
r	S										
	42	44	46	48	50	52	54	56	58	60	62
0	0.18	0.15	0.12	0.10	0.08	0.07	0.06	0.04	0.04	0.03	0.02
0.002	0.25	0.21	0.18	0.15	0.13	0.11	0.09	0.08	0.07	0.06	0.05
0.003	0.29	0.25	0.22	0.18	0.16	0.14	0.12	0.10	0.09	0.07	0.06
0.004	0.33	0.29	0.25	0.22	0.19	0.17	0.15	0.13	0.11	0.10	0.08

Table 3: Probability of dismissing the other team, given the probability of a run-out (r) and the team bowling strike rate (S), and assuming 10 wides or no-balls are bowled per 50 overs.

There was further newspaper interest in these ideas in January 2001, culminating in an article [6] in Sydney's *Daily Telegraph*. That article included also a description of the main results in de Mestre and Cohen [10], and it was reprinted with a little more mathematical detail in Cohen and de Mestre [7]. The newspaper article included probabilities of dismissing the other team for the triangular one-day series about to commence between Australia, Zimbabwe and the West Indies. The predictions were acceptably accurate, as detailed in [7].

3 An application of batting averages

As we have indicated above, the work of this section is quite distinct from the preceding work. It will be convenient to use a similar notation to before, but now from a batsman's point of view. For example, we will use b for the number of balls faced by a particular batsman, rather than the number bowled by a particular bowler.

3.1 The distribution of scoring strokes

In cricket, a batsman's average is the number of runs he has scored divided by the number of times he was out. If a tail-end batsman scores five in each of ten innings in a season and is not out nine times, then he finishes the season with an average of 50. For good reason, this is not seen to be a properly representative score.

It seems that there are two separate questions that people expect the one batting average to answer.

- First, how good is the batsman? What score would we expect of him if he were allowed to bat on, leaving the field for the final time only when he is given out in a standard manner?
- Second, what score do we expect of him given the possibilities also of his retiring hurt, or running out of batting partners, or having the team's innings declared closed or the match interrupted for some reason such as rain, in all of which cases he would remain not out?

We will try to answer both questions.

At least two papers have attempted to determine more significant single measures of batting performance, generally being more intent on answering the first of the above questions. Danaher [8] used analogies with survival analysis to find an estimate of a cricketer's "true but unknown batting average" based on the product limit estimator. Kimber and Hansford [14] also adopted an approach "akin to that used in reliability and survival analysis", and also based on product limit estimation, to arrive at

a different nonparametric estimator. The latter gives values generally much closer to the traditional average than Danaher's estimator (with both always giving smaller values), and both have the property that the fewer the number of not-outs, the closer their estimator is to the traditional average. On the other hand, Davis [9, pages 96–98], argues from an empirical viewpoint for the worth of the traditional average.

It seems reasonable that the score you might expect a batsman to attain would be his "true average", based on all relevant previous innings. To define this term, we take data pertaining to a particular batsman over a particular period, such as his career or the previous season, or in a particular position, or against a particular team. Let i, n, w and r be, respectively, the number of innings, the number of not-out innings, the number of dismissals, and the number of runs scored. Then $w = i - n$. The batsman's traditional and true averages are

$$B = \frac{r}{w}, \qquad A = \frac{r}{i},$$

respectively. Notice that

$$A = \frac{w}{i}B = \frac{i-n}{i}B, \tag{4}$$

so that, for overall career results, say, the true average may be determined from the usual published batting statistics. Of course, $A = B$ when $n = 0$.

We will justify our use of the term "true average" by obtaining in a theoretical fashion the probability distribution of the number of scoring strokes and, based on this, showing that the expected value of the batsman's score (in the statistical sense) equals this true average.

The same approach will allow us to find the probability of the batsman making 100 runs, or any other score. Thus we answer the intriguing question: how do you estimate a batsman's probability of making a century if he has not yet made one? We will see that our probability compares well with the actual frequency of century scores by batsmen who have made a few centuries.

Our method relies on the new concept of the *strike constant*. This is the ratio of runs made to number of scoring strokes and its introduction may be viewed as a device to serve our end: it is a first approximation to a comparison of runs made and scoring strokes which indeed (as we will see) leads to plausible and testable results. An investigation of this ratio for a large number of Sydney grade cricketers by Cochran [3] came up with the value 2.16, with standard deviation 0.25, for traditional cricket, and 1.82, standard deviation 0.43, for limited-overs cricket. (He also investigated indoor cricket: ignoring runs subtracted for loss of wicket, the mean strike constant was 2.08 with standard deviation 0.41.)

We consider the main application of this work to be to the traditional form of cricket. Strike constants for individual cricketers over a small number of matches might range between 1.9 and 2.4, say, but this will be seen in any case to have little effect on the final calculations.

We will show that, subject to certain assumptions, the number of scoring strokes follows a geometric distribution. The distribution of runs scored is related to this through the strike constant. The possibility that cricket scores are geometrically distributed goes back at least to the writings of Elderton [11]. Wood [15] gives further numerical evidence to support this. Both these papers are dismissed by Kimber and Hansford [14] as "flawed because the authors treated not-out scores as if they were completed innings", despite the evidence of the data. The details are summarised by Clarke [2]. *Inter alia*, Clarke states: "If a batsman scores only singles and his probability of dismissal is constant, his scores should follow a geometric distribution, the discrete equivalent of the negative exponential." This observation appears to be based on a viewing of the empirical data, but will be a direct consequence of our work below.

3.2 Expected values

In addition to the quantities i, n, $w = i - n$ and r introduced above, let b and s be, respectively, the number of balls faced and the number of scoring strokes made. We must have $r \geq s \geq 0$ and we will assume that $b > i > n \geq 0$. Then $w > 0$. Recall that $A = r/i$.

We define

$$p_w = \Pr(\text{the batsman's innings ends, out or not out, with each ball faced}) = \frac{i}{b},$$
$$q_w = 1 - p_w,$$
$$p_s = \Pr(\text{the batsman makes a scoring stroke with each ball faced} \mid \text{the batsman's innings does not end with that ball}) = \frac{s}{b-i},$$
$$q_s = 1 - p_s.$$

These probabilities are considered to be constant throughout a subsequent innings.

Notice that we have made an assumption that no scoring stroke is made from the ball on which the batsman's innings ends (so that $s \leq b - i$). Therefore, we do not take into account the rare instance in which the batsman makes at least one run and is then run out on the same ball while attempting a further run, or the admittedly more common instance in which a captain declares an innings closed following the batsman's final scoring stroke.

In *any* period, the ratio of the number of runs obtained to the number of scoring strokes made is considered to be constant. This is the simplification described above. The ratio is the strike constant, denoted by κ. Then

$$\kappa = \frac{r}{s}.$$

Let the random variable X be the number of scoring strokes made by the batsman in a subsequent innings, and let R be the score (number of runs) obtained. In order that $X = k$ for integer $k \geq 0$, the batsman must face $j + 1$ balls, for some $j \geq k$, scoring on k of these and having his innings end on the $(j + 1)$th ball. (If the team's innings ends or the batsman is run out while not facing, some number of balls after last facing a ball himself, this is still effectively the case.) Whether or not he scores off any of the first j balls bowled are considered to be independent events, and so the distribution of the k scoring strokes among the first j balls bowled to him will be binomial(j, p_s). Write $\Pr(X = k)$ for the probability that the batsman makes k scoring strokes ($k \geq 0$), before being dismissed. (Later notations will have a corresponding meaning.) Then

$$\begin{aligned}\Pr(X = k) &= \sum_{j=k}^{\infty} q_w^j p_w \cdot \binom{j}{k} p_s^k q_s^{j-k} \\ &= \frac{p_s^k p_w q_w^k}{k!} \sum_{j=k}^{\infty} \frac{j!}{(j-k)!} (q_s q_w)^{j-k} \\ &= \frac{p_s^k p_w q_w^k}{(1 - q_s q_w)^{k+1}} = \frac{p_w}{1 - q_s q_w} \left(\frac{p_s q_w}{1 - q_s q_w} \right)^k = PQ^k,\end{aligned}$$

where

$$P = \frac{p_w}{1 - q_s q_w}, \quad Q = \frac{p_s q_w}{1 - q_s q_w} = 1 - P.$$

Thus the number of scoring strokes made follows a geometric distribution. (Notice, for example, that if $p_s = 1$ then this reduces to $\Pr(X = k) = q_w^k p_w$, for $k \geq 0$.) Using the definitions of p_w and p_s, we find that

$$P = \frac{i}{i+s} = \frac{\kappa}{A + \kappa}, \quad Q = \frac{A}{A + \kappa}.$$

The expected number of scoring strokes is then easily determined, or it may be obtained as a particular case from results in Johnson *et al.* [13]. We have

$$\mathrm{E}(X) = \sum_{k=0}^{\infty} k \cdot \Pr(X = k) = \frac{Q}{P} = \frac{A}{\kappa}.$$

Name	Inn.	NO	Runs	Ave.	A	A_w	100+	E(100+)	50+	E(50+)
DG Bradman	80	10	6996	99.94	87.45	85.04	29	25.8	42	45.5
RG Pollock	41	4	2256	60.97	55.02	54.43	7	6.9	18	16.8
GA Headley	40	4	2190	60.83	54.75	45.61	10	6.7	15	16.3
H Sutcliffe	84	9	4555	60.73	54.23	54.64	16	13.7	39	34.0
E Paynter	31	5	1540	59.23	49.68	48.31	4	4.3	11	11.6
KF Barrington	131	15	6806	58.67	51.95	50.37	20	19.8	55	50.9
EdeC Weekes	81	5	4455	58.61	55.00	54.88	15	13.6	34	33.2
WR Hammond	140	16	7249	58.46	51.78	46.19	22	21.0	46	54.3
SR Tendulkar	143	15	7419	57.96	51.88	48.85	27	21.6	57	55.5
GStA Sobers	160	21	8032	57.78	50.20	44.06	26	22.7	56	60.2
JB Hobbs	102	7	5410	56.95	53.04	53.34	15	16.0	43	40.4
CL Walcott	74	7	3798	56.69	51.32	51.03	15	10.9	29	28.5
L Hutton	138	15	6971	56.67	50.51	47.89	19	19.8	52	52.3
GE Tyldesley	20	2	990	55.00	49.50	47.22	3	2.8	9	7.4
MH Richardson	21	1	1088	54.40	51.81	50.75	2	3.2	10	8.1
DR Martyn	35	9	1413	54.35	40.37	35.88	4	3.1	9	10.4
CA Davis	29	5	1301	54.21	44.86	40.83	4	3.3	8	9.7
VG Kambli	21	1	1084	54.20	51.62	53.30	4	3.1	7	8.1
GS Chappell	151	19	7110	53.86	47.09	44.57	24	21.8	55	57.3
AD Nourse	62	7	2960	53.82	47.74	47.49	9	8.0	23	22.2

Table 4: The all-time top twenty Test batting averages, at 7 February 2002, with approximate expected values of number of scores of 100 or more, or 50 or more (E(100+) and E(50+), respectively).

differing from the usual lists which give the number of scores from 50 to 99, inclusive); and their expected values similarly calculated.

Unless it is possible to have access to the original score sheets, it is most unlikely that actual values of κ, the ratio over the past of runs made to number of scoring strokes, could be obtained. The easy approach is to set $\kappa = 2$, and this was done in Table 4. (The expected values, whether we took $\kappa = 1.9$, 2 or 2.1 were not appreciably different.)

A large proportion of the expected values in Table 4 are observed to match their actual values very well, so that, in this case at least, the model fits the data acceptably. The table suggests that our model, with $\kappa = 2$, will allow reasonable predictions to be made.

Taking $\kappa = 2$ allows a further simplification. By assuming that c is even, as in the common cases $c = 100$ or $c = 50$, we obtain

$$\Pr(R \geq c) \approx \left(\frac{A}{A+2}\right)^{c/2}, \tag{5}$$

and, if desired, this may be adopted as a useful approximation for all $c \geq 1$.

The wicket-average A_w in Table 4 is the average of only those innings in which the batsman was out. (These values were obtained by going back to the lists of all Test scores, for each batsman.) This information has been included to show that the true average and the wicket-average are in most cases very close, as one would expect if a batsman averaged much the same in his completed innings as in his not-out innings. However, the true average is greater in all but two cases, indicating that not-out scores tend on average to be greater than completed innings. Sometimes this is emphatically so, as in Headley's case: his not-out Test scores were 102, 169, 270 and 7.

The point of tabulating A_w is to give an example of other averages that might be determined for more accurate predictions. Using equation (5) with $A = A_w$ and $c = 100$ will give an estimate of the chance of scoring a century, with the batsman getting out. (The earlier theory needs to be adjusted in a minor way to allow for the different sample space: b, r and s now relate only to completed innings, and i in the definitions of p_w and p_s must be replaced by w. Then, in particular, p_w is the probability of the batsman losing his wicket, out, with each ball faced. The subsequent analysis would then refer only to completed innings.)

We return now to the question of determining a more useful means of estimating a batsman's future score than simply giving the expected value. We will instead find "50% probability intervals" for the

score, for differing values of A. That is, for each A, we will determine an approximate interval with the property that the batsman would obtain a score in it with probability 0.5, with equal probabilities of smaller or greater scores outside the interval.

We make use of (5). For a given probability p, the score c required to ensure that $\Pr(R \geq c) = p$ is obtained approximately by solving

$$\left(\frac{A}{A+2}\right)^{c/2} = p.$$

We obtain

$$c = \frac{2\log(1/p)}{\log(1+2/A)}, \tag{6}$$

where the logarithms may be to any suitable base. Taking the ceiling value when $p = 0.75$ and the floor value when $p = 0.25$, we obtain our approximate interval.

Examples of these intervals appear in Table 5. We have taken values of A from 10 to 65, incremented by 5, and, in case the ghost of Sir Donald is watching, also $A = 90$. Notice that any sensible average can be used. Thus, for Damien Martyn with $A_w \approx 35$ and $A \approx 40$ (see Table 4), we could say he has a 25% chance of scoring 50 or more, getting out, but the same chance of scoring more than 56, out or not out.

batting average	10	15	20	25	30	35	40
50% probability interval	[4, 15]	[5, 22]	[7, 29]	[8, 36]	[9, 42]	[11, 49]	[12, 56]
batting average	45	50	55	60	65	...	90
50% probability interval	[14, 63]	[15, 70]	[17, 77]	[18, 84]	[19,91]	...	[27, 126]

Table 5: A batsman with true batting average shown (*not* the traditional average) has probability 0.5 of making a score in the given interval, with equal probabilities of smaller or greater scores outside the interval.

Using equation (6) with $p = 0.5$ allows us to use a batsman's traditional average number of runs scored to estimate his median number of runs scored. In fact, for $A \geq 20$, say, we have $\ln(1+2/A) \approx 2/A$, so that the median score is about $A \ln 2$. Thus $0.7A$ would be an easy approximate formula for the median.

Data on median scores is almost nonexistent, but Wood [15, Table B] gives this information for 22 "leading batsmen" to September 1939. Their ratio of median score to traditional average score (B) ranges from 0.61 to 0.71. Wood's list does not allow direct calculations of the true average A, since he does not give the numbers of not-out innings. The CricInfo web site allowed this to be done (although it differed from Wood on *every* occasion in the number of first class innings for the batsmen on his list). This exercise suggested that taking $A \approx 0.9B$ would be acceptable in general (and is the rule of thumb following the observation by Clarke [2] that "more than 10% of scores are not outs"), so that $0.63B$ would be a useful theoretical estimate of a batsman's median score over a long career.

We also note that the formula (5) retains the non-aging (or Markovian, or lack of memory) property of the geometric distribution (see Johnson *et al.* [13, page 201]). Thus, for example, $\Pr(R \geq 100) = (\Pr(R \geq 50))^2$.

Finally, we consider the probability of a batsman getting a duck. From our early work, the probability of a batsman making no scoring stroke is

$$\Pr(X = 0) = P = \frac{\kappa}{A+\kappa}.$$

But this includes the probability of scoring 0, not out. Our earlier discussion suggests the following as the way to go:

$$\Pr(\text{duck}) \approx \frac{2}{A_w + 2}.$$

Name	Inn.	NO	A_w	ducks	E(ducks)
DG Bradman	80	10	85.04	7	1.8
RG Pollock	41	4	54.43	1	1.4
GA Headley	40	4	45.61	2	1.7
H Sutcliffe	84	9	54.64	2	3.0
E Paynter	31	5	48.31	3	1.2
KF Barrington	131	15	50.37	5	5.0
EdeC Weekes	81	5	54.88	6	2.8
WR Hammond	140	16	46.19	4	5.8
SR Tendulkar	143	15	48.85	7	5.6
GStA Sobers	160	21	44.06	12	6.9
JB Hobbs	102	7	53.34	4	3.7
CL Walcott	74	7	51.03	1	2.8
L Hutton	138	15	47.89	5	5.5
GE Tyldesley	20	2	47.22	2	0.8
MH Richardson	21	1	50.75	0	0.8
DR Martyn	35	9	35.88	2	1.8
CA Davis	29	5	40.83	1	1.3
VG Kambli	21	1	53.30	3	0.8
GS Chappell	151	19	44.57	12	6.4
AD Nourse	62	7	47.49	3	2.5

Table 6: The all-time top twenty Test batsmen, by traditional batting average, at 7 February 2002, with number of ducks scored and the approximate expected value of this number.

The world's top twenty batsmen have been known to score a duck or two. Table 6 repeats some information from Table 4, and gives the number of ducks scored by those batsmen in Tests and our suggested expected value of this number using the probability estimate $2/(A_w + 2)$ multiplied by the number $i - n$ of completed innings. The table indicates some level of agreement between the actual and expected values; any attempt to model small numbers like these would be generally acknowledged as difficult.

4 Conclusion

Many papers concerned with tennis have exploited the fact that the proportion of points won by a player in some situation allows estimates of the probability of winning a game, set or match in a similar future situation. Considering separately points won on service and points won when receiving leads to refined estimates. In Bennett [1], there are references to probabilistic analysis in tennis, baseball, basketball and American football, and numerous other relevant references. Yet, as we have already quoted Stephen Clarke as saying, hardly any such analysis has previously taken place in cricket.

A "winning" ball in a game of cricket is one that takes a wicket from the bowler's point of view, or allows a scoring stroke from the batsman's point of view. The proportion of winning balls has been used in this paper to give the probability of bowling out a team, in the former case, or scoring a century, in the latter case. Along the way, refinements and other applications have been given.

Bowling strike rates, along with estimates of the probability of running out an opposing batsman, have been used in Section 2 not only to find the probability of dismissing the other side in one-day cricket, but to demonstrate that this chance is approximately doubled when there is a good likelihood of obtaining at least one run-out.

At the beginning of Section 3, two questions were posed regarding conclusions to be drawn from the traditional batting average. But, to make a pun of it, this average is a very demeaned statistic. Even in cricketing circles, it is not seen as being properly representative of a batsman's past scores because of the "ad hoc" treatment of not-outs.

We prefer instead the true batting average A: simply the average of all scores, out or not out. The wicket average A_w, which refers specifically to completed innings, is approximately the same as A and should be used for questions concerning completed innings (such as the first of those at the beginning of

Section 3). Use A otherwise. Among other things, we have justified the simple formula $(A/(A+2))^{c/2}$ as the probability of scoring at least c runs, and the formula $2/(A_w+2)$ as the probability of a duck. Both of these have been compared favourably with results from the history of cricket.

The work of Section 3 depends crucially on the concept of the strike constant, although less crucially on the value chosen for it. As a theoretical if hypothetical construct, its worth seems clear, and further investigation of the notion would be extremely welcome.

Acknowledgment

I am grateful for discussions with Dr Peter Wright, formerly of the University of Technology, Sydney, with regard to Section 3. In particular, he suggested the use of the 50% probability interval to quantify a batsman's chances.

References

[1] J. Bennett (editor), *Statistics in Sport*, London, Arnold (1998).

[2] S. R. Clarke, "Test statistics", in *Statistics in Sport*, J. Bennett (editor), Arnold, London (1998), 83–103.

[3] K. Cochran, *Comparison of the Score to Scoring Strokes Ratio in Various Forms of Cricket*, undergraduate essay, Department of Mathematical Sciences, University of Technology, Sydney (2000).

[4] G. L. Cohen, "One-day cricket: inferences from bowlers' strike rates", *Math. Today*, **35** (1999), 45–47.

[5] G. L. Cohen, "One-day cricket: the effect of running out an opposing batsman", *Math. Today*, **36** (2000), 75–77.

[6] G. Cohen, "For openers, let's try a little maths", *The Daily Telegraph* (11 January 2001), 18.

[7] G. Cohen and N. de Mestre, "Mathematics and cricket", *Reflections*, **26** (May 2001), 11–13.

[8] P. J. Danaher, "Estimating a cricketer's batting average using the product limit estimator", *N. Z. Statist.*, **24** (1989), 2–5.

[9] C. Davis, *The Best of the Best*, Sydney, ABC Books (2000).

[10] N. de Mestre and G. Cohen, "The flight of a cricket ball", in *Fifth Conference on Mathematics and Computers in Sport*, G. Cohen and T. Langtry (editors), University of Technology, Sydney, Australia (2000), 107–112.

[11] W. Elderton, "Cricket scores and some skew correlation distributions", *J. Roy. Statist. Soc.*, **108** (1945), 1–11.

[12] N. Hawkes, "Howzat for statistics", *The Times* (12 May 1999), downloaded from the world wide web.

[13] N. L. Johnson, S. Kotz and A. W. Kemp, *Univariate Discrete Distributions* (second edition), New York, Wiley (1993).

[14] A. C. Kimber and A. R. Hansford, "A statistical analysis of batting in cricket", *J. Roy. Statist. Soc. Ser. A*, **156**, (1993) 443–455.

[15] G. H. Wood, "Cricket scores and geometrical progression", *J. Roy. Statist. Soc.*, **108** (1945), 12–22.

[16] J. Zubrzycki, "Professor's figures add up for punters", *The Australian* (14 May 1999), 18.

Proceedings of the
Sixth Australian Conference on

MATHEMATICS AND COMPUTERS IN SPORT

Bond University
Queensland

Edited by
Graeme Cohen and Tim Langtry
Department of Mathematical Sciences
Faculty of Science
University of Technology, Sydney

6M&CS

1 - 3 July 2002

Typeset by LaTeX 2ε

Proceedings of the Sixth Australian Conference on Mathematics and Computers in Sport, 1–3 July 2002, at Bond University, Gold Coast, Queensland 4229, Australia.

ISBN 1 86365 532 8

Printed by UTS Printing Services

Contents

Abstract

Conference Director's Report

Welcome to the Sixth Australian Conference on Mathematics and Computers in Sport. This year we return to Bond University after the successful Fifth Conference in 2000 held in Sydney because of the Olympic Year in that city. It is a pleasure to renew acquaintances with one of our principal speakers, Ray Stefani who was at the First Conference in 1992, and has been collaborating with Steve Clarke from Swinburne for many years now. Our second principal speaker, Steve Gray from Queensland, will add a new dimension to the Sixth Conference with his interest in the economical aspects of sport. Graeme Cohen (now retired from UTS) is our third principal speaker, but he and Tim Langtry have once again taken over the responsibilities of producing the printed *Proceedings*. I thank Graeme and Tim for relieving me of this major task.

The conference has once again attracted academics from New Zealand, the United Kingdom, the United States and Canada. I welcome them all, including many familiar faces. I hope that you all find the conference rewarding in many aspects, including the content and presentation of the talks, the many discussions that are generated, and the close social contact with like-minded academics. We now have a website www.anziam.org.au/mathsport due to Elliot Tonkes, who is the webmaster. This site contains information about this and all previous conferences.

All the papers in these *Proceedings* have undergone a detailed refereeing process. I am indebted to the referees for their time and comments to improve the quality of all papers. The *Proceedings* begin with the papers of our principal speakers, followed by the contributed papers in alphabetical order of author, or first author. The *Proceedings* conclude with an abstract—the paper was accepted for presentation but was received too late to be submitted to the full refereeing process.

Neville de Mestre
June 2002