J. R. Statist. Soc. A (1993)
156, *Part* 3, *pp.* 443–455

A Statistical Analysis of Batting in Cricket

By ALAN C. KIMBER† and ALAN R. HANSFORD

University of Surrey, Guildford, UK

[Received July 1992. Revised February 1993]

SUMMARY

The batting average is ubiquitous in cricket. In this paper we show that the traditional batting average depends on an unrealistic parametric assumption. We propose a nonparametric approach based on runs scored for assessing batting performance. The methods have been applied to a large sample of players at various levels of cricket, examples of which are featured in this paper. The statistical methodology employed is akin to that used in reliability and survival analysis.

Keywords: BATTING; CENSORING; CRICKET; GEOMETRIC DISTRIBUTION; POINT PROCESS; PRODUCT LIMIT ESTIMATOR; RELIABILITY; REPAIRABLE SYSTEMS; SURVIVAL ANALYSIS

1. INTRODUCTION

Cricket is a sport in which statistics feature heavily. Most cricket statistics are either records (for example, the highest first-class score is 499 by Hanif Mohammad), cumulative totals (for example, Alan Hansford scored 48 first-class runs for Sussex in 1989) or averages (for example, Ranjitsinhji had a batting average of 45.12 in 1897).

Averages feature heavily in media coverage of professional cricket; they are calculated for even the humblest of Sunday afternoon cricket clubs and any cricketer will be able to name team-mates (even though to play for one's averages is just not cricket!) who know their own averages at any given moment to at least two decimal places. Also, whether or not a professional player is offered a contract probably depends to some extent on his averages. Thus, averages are not just of academic interest. Interesting statistical issues arise in terms of batting averages and related matters. The crucial issue here is how 'not-out' scores are treated, a matter that has not been addressed properly before, despite grumbles from generations of cricketers that the batting average is somehow unfair. Earlier published work on cricket scores (Elderton, 1945; Wood, 1945) is flawed because the authors treated not-out scores as if they were completed innings.

In Section 2 we adopt an approach akin to that used in reliability and survival analysis to investigate the properties of the usual batting average. We suggest an alternative in Section 3. In Section 4, we discuss approaches for summarizing a batsman's distribution of scores more fully. We end with some further remarks in Section 5. The methods are illustrated throughout by examples taken from professional cricket.

A list of the sets of scores used in this study is given in Appendix A. The main sources are various editions of *Wisden Cricketers' Almanack*, *The Playfair Cricket*

†*Address for correspondence*: Department of Mathematical and Computing Sciences, University of Surrey, Guildford, GU2 5XH, UK.

 0035–9238/93/156443

Annual and *The Cricketer Quarterly Facts and Figures*. Other sources are cited in the text. The sets of scores selected are mainly for modern players, which is a reflection of the data sources that were readily available to us. However, we have chosen batsmen from various countries and with varying levels of achievement. Thus, our sample of sets of scores is, we believe, reasonably representative.

The term batsman is used for convenience here. Cricket is not purely a male sport but we have no scores from women's cricket, which receives very limited coverage in the data sources that are readily available to us.

2. BATTING AVERAGE AND PROPERTIES

The batting average B for a batsman is

$$B = \text{number of runs scored/number of times out.}$$

If not-out innings could not occur, then B would be the mean score and would be a natural statistic: by regarding the set of scores either as the population so that B is *the* mean score, or as a random sample from an infinite population of potential scores, so that B is the natural nonparametric estimate of the population mean score. However, since about 10% of all scores are not-out scores the case for using B in either sense is less compelling. We now investigate B more closely.

2.1. *Is It Sensible to Calculate any Batting Average?*

A batsman's scores are observed chronologically. If we regard each occasion that a batsman is out as an event, then we observe a point process. A convenient representation is to regard interevent 'times' as the number of runs scored by a batsman between successive times out. Such processes commonly arise, though not usually with discrete interevent times, in repairable systems reliability. If there are major trends in a point process, then to ignore time order may lead to an inappropriate analysis; see Crowder *et al.* (1991), p. 158. In our context, a plausible view *a priori* is that a batsman in his career might have relatively low scores initially, a period of relatively high scores when established, followed by a steady decline as age takes its toll. Were this so, then any measure of marginal behaviour, such as an average, might be misleading. One might expect that trends would be less evident over a shorter period, such as a season.

A simple way to detect trends in point process data is to plot event number against cumulative time (Crowder *et al.* (1991), p. 159). Major departures from linearity indicate a trend. Perhaps surprisingly, for the vast majority of sets of scores that we looked at, the departures from linearity in such plots were small. The results of the graphical analyses were confirmed with Laplace's test for trend; see Ascher and Feingold (1984). Also, there was no major evidence of autocorrelation in the observed point processes. Thus it is quite reasonable as a first approximation to treat scores as if they were independent and identically distributed observations. Hence an average is a potentially useful summary statistic here.

2.2. *B and Geometric Distribution*

Consider a set of scores $x_1, x_2 \ldots, x_n$ for a batsman, together with indicators $d_1, d_2, \ldots, d_n$, where $d_i = 1$ if the batsman was out for x_i and $d_i = 0$ if the

batsman was left not out with a score of x_i $(i = 1, \ldots, n)$. This is similar to a life test in reliability, with the x_i as the component lifetimes and the d_i as censoring indicators. If the lifetimes are independent geometric random variables, each with probability mass function (PMF)

$$p_0(x) = \theta(1 - \theta)^x, \qquad x = 0, 1, 2, \ldots$$

where $0 < \theta < 1$ is an unknown parameter, and if the censoring is non-informative, then it is easy to show that the maximum likelihood estimate of the population mean lifetime is B. Thus, if the underlying distribution of scores is geometric, B is optimal. Non-informative censoring here means that a score of x not out is representative of all scores of x or more, which is a reasonable assumption.

Consistency is a desirable property for an estimator to have; see Cox and Hinkley (1974). If not-out scores were not possible, then B would trivially be consistent whatever the distribution of scores (under mild regularity conditions). However, not-out scores do occur, giving rise to a possibly complicated censoring mechanism for any particular batsman. Hence we would like B to be consistent for the mean *whatever the censoring mechanism*. However, the geometric distribution for scores is the *only* distribution on the non-negative integers for which this holds. Details are given in Appendix B. So, unless the distribution of scores is geometric, B, in a large sample sense, estimates the wrong quantity.

The geometric model cannot apply *exactly* since a score does not increase by a fixed amount at each stage. However, since the statistical case for B rests on a geometric assumption we now check on its empirical validity.

2.3. *Informal Checking of Fit of Geometric Model*

Cricketing lore indicates that

(a) a batsman is vulnerable when first going in to bat,
(b) some batsmen are accused of being careless when well set and some scores are claimed to be unlucky or difficult, such as 87, 111 and the 'nervous nineties', and
(c) after scoring many runs a batsman tires and becomes more error prone.

In terms of the discrete hazard function

$$h(x) = p(x) \Big/ \sum_{y=x}^{\infty} p(y),$$

where $p(\ )$ is the underlying PMF of the scores, a bathtub hazard with some locally high points emerges. General reliability considerations would *a priori* tend to support observations (a) and (c) at least. This is in contrast with the geometric model which has constant hazard, $h_0(x) = \theta$.

First, we compared the observed proportion of ducks (0 scores) and the corresponding expected proportions under the geometric model. If observation (a) were valid, its greatest effect would be at 0. Table 1 shows a selection of results. They confirm the lack of fit of the geometric model at 0. Similar results for small positive scores were also obtained. Fig. 1 shows a typically shaped plot of an empirical hazard, smoothed using the 'twiced' running medians method of Tukey (1977); see also Becker and Chambers (1984). The plot is for Sir Donald Bradman's first-class

TABLE 1

Comparison of the observed proportions of ducks with the expected proportion under the geometric model for some players in test-matches, one-day internationals and first-class cricket

Batsman	*Type of cricket*	*B*	*Proportion of 0-scores*		*Ratio*
			Expected E (geometric)	*Observed O*	*(O/E)*
Armarnath	Tests	42.50	0.023	0.106	4.62
Bradman	First class	95.14	0.010	0.047	4.55
Bradman	Tests	99.94	0.010	0.088	8.83
Gatting	Tests	37.57	0.026	0.111	4.28
Gavaskar	1-day internationals	34.04	0.029	0.078	2.73
Gower	Tests	44.31	0.022	0.035	1.59
Holding	Tests	13.79	0.068	0.197	2.92
Hughes	Tests	37.41	0.026	0.081	3.11
Kirmani	Tests	27.04	0.036	0.057	1.60
Mudassar	1-day internationals	25.28	0.038	0.052	1.37
Mudassar	Tests	38.09	0.026	0.060	2.35
Trumper	First class	46.58	0.021	0.061	2.89
Wood	Tests	31.83	0.030	0.080	2.63

scores of 50 or less. It shows that his hazard does not flatten out until about 5 or 6. For many batsmen the hazard takes longer to flatten. To estimate the probabilities in the empirical hazard we used the product limit (PL) estimator of the survivor function; see Kaplan and Meier (1958) and Crowder *et al.* (1991), p. 45.

Next we considered observation (b). Some batsmen have regions of high hazard. For example, in test-matches David Gower has been particularly vulnerable in the low 70s. However, we found no general evidence to support the superstitions, though since most 'unlucky' scores are quite high the data are rather sparse.

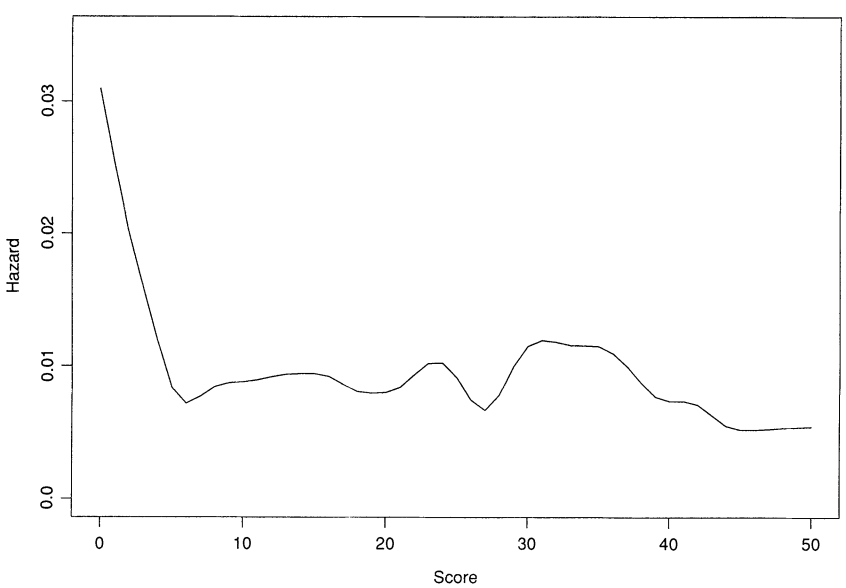

Fig. 1. Smoothed empirical hazard plot of the first-class scores of 50 or less of Sir Donald Bradman

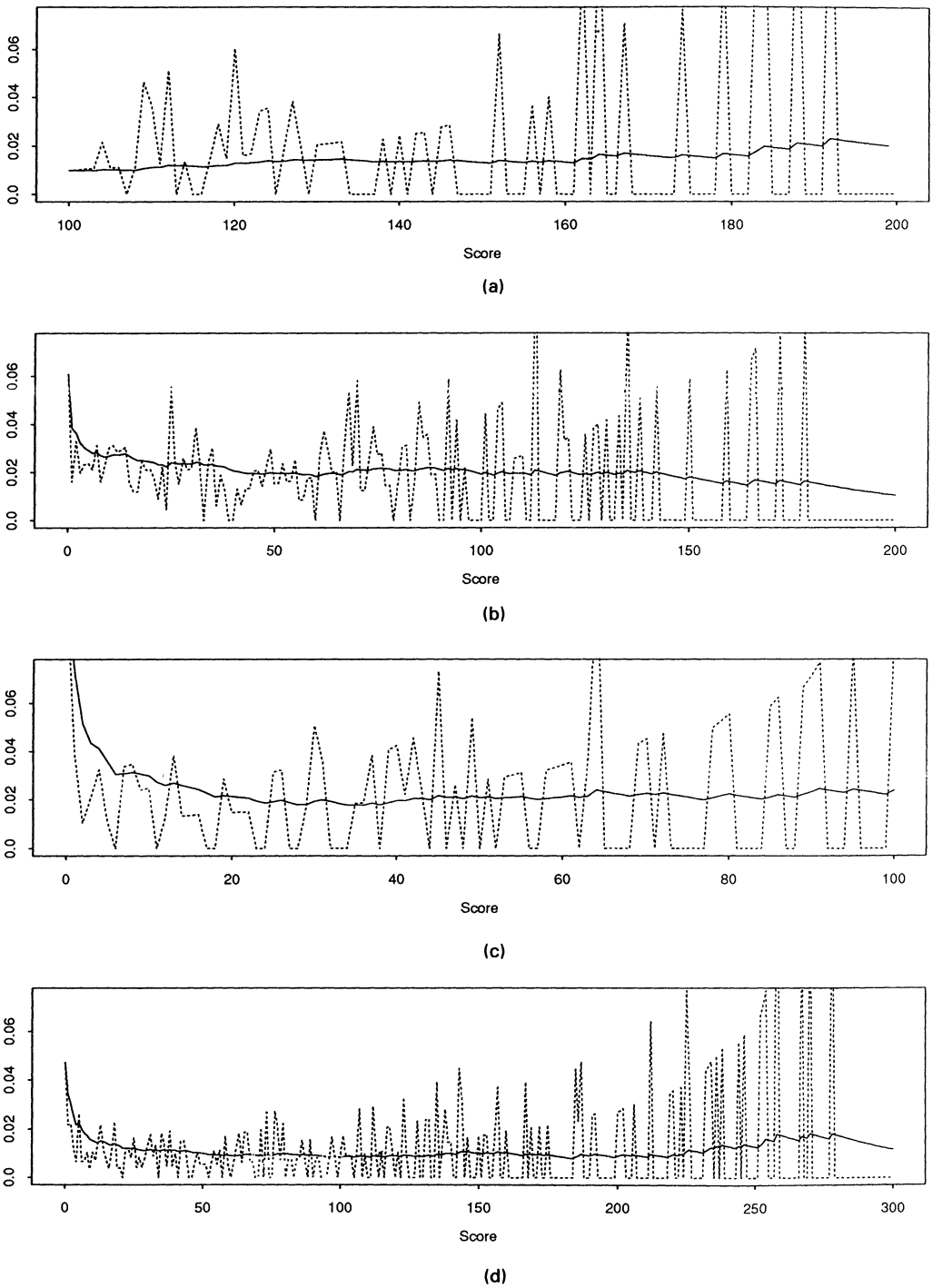

Fig. 2. Empirical hazards (·····) and smoothed empirical hazards (———) for (a) the first-class centuries of Dennis Amiss, (b) the first-class scores of Victor Trumper, (c) the test scores of Mohinder Armarnath and (d) the first-class scores of Sir Donald Bradman

Similar sparse data problems apply to checking observation (c) since fatigue is usually only a major factor (for a professional cricketer at least) after several hours of batting, by which time a high score will have been made. Fig. 2 shows four empirical hazard functions: for the first-class centuries of Dennis Amiss, the first-class scores of Victor Trumper, the test scores of Mohinder Armarnath and the first-class scores of Bradman. The isolated spikes whose heights increase with score are typical of empirical hazard plots with increasingly sparse data in the upper tail. The superimposed curve on each plot is an exponentially smoothed version of the empirical hazard; see Kendall and Ord (1990). The running medians method is of no use here since it filters out isolated spikes completely. In three of the plots the smoothed empirical hazard is quite flat at high scores. For Bradman the empirical hazard does increase somewhat but only beyond about 230. Most players never reach this level of score in their careers. Thus, it seems plausible that the tail of the underlying score distribution for a batsman is at least roughly geometric in form.

2.4. *Testing Fit of Geometric Distribution*

Let

$$F(x) = \sum_{y=x}^{\infty} p(y)$$

denote the survivor function of the underlying scores distribution. Then the log-likelihood of a set of scores is given by

$$L = \sum_{i=1}^{n} \{d_i \log h(x_i) + \log F(x_i)\}.$$

Under the geometric model $h = h_0$, $F(x) = (1 - \theta)^x$ and L is maximized by taking $\theta/(1 - \theta) = B$. When $h(x)$ is unconstrained, L is maximized by using the PL estimate of F (Cox and Oakes, 1984). If the highest score is H, then the large sample distribution of twice the difference of the maximized log-likelihoods is $\chi^2(H - 1)$ when the true distribution of scores is geometric. This gives a goodness-of-fit test for the geometric model. However, given the skewness of the geometric distribution, the null probability masses at high scores will be small, thereby casting doubt on the adequacy of the χ^2-approximation. To avoid this, we used N ($<H$) in place of H and treated all scores in excess of N as N not out to estimate θ. The choice of N varied from player to player, depending on ability and the amount of data available. It ranged from 15 (Jeff Thomson, who was not played for his batting) to 60 (Viv Richards, a great batsman).

We applied this test to a random subsample of 12 sets of career scores, the results for which are given in Table 2. A plot (not shown) of the ordered p-values against their null expectations showed strong evidence against the geometric hypothesis. Because of the extra grouping, the test is not sensitive to departures from the geometric assumption in the upper tail. So, using the memoryless property of the geometric distribution, we also applied the test to four sets of first-class century scores from players who had made more than 100 centuries (Amiss, Sir Colin Cowdrey, Sir Jack Hobbs and Glenn Turner). Here there was little evidence against

TABLE 2
Results of the goodness-of-fit test for the geometric model for a random subsample of 12 players in tests and one-day internationals

Player	*Type of cricket*	*Degrees of freedom*	*Test statistic*	*p-value*
Border	Tests	49	75.03	0.0098
Botham	Tests	39	47.33	0.17
Emburey	Tests	24	55.09	0.0003
Gower	1-day internationals	29	34.51	0.22
Greenidge	Tests	49	55.57	0.24
Haynes	1-day internationals	54	43.39	0.85
Hughes	Tests	29	41.26	0.065
Imran	1-day internationals	29	49.78	0.0095
Javed	1-day internationals	34	42.67	0.15
Richards	Tests	59	80.44	0.033
Srikkanth	1-day internationals	29	41.45	0.063
Thomson	Tests	14	18.52	0.18

the geometric model in the upper tail. Thus, the results of the goodness-of-fit tests tend to support the indications obtained from the informal analysis.

3. ALTERNATIVE APPROACH

3.1. *Alternative Batting Average*

Having seen that the geometric distribution gives a poor fit to sets of scores, so that B is inconsistent, we propose an alternative batting average that does not depend on the geometric assumption.

Let M and M^* be respectively the highest completed innings and the highest not-out score made by a batsman so that $H = \max(M, M^*)$. For now suppose that $M > M^*$. Let $\hat{F}(x)$ denote the PL estimate for $F(x)$. In this case $\hat{F}(x)$ is defined for all integers x. A natural competitor for B is the alternative batting average A, defined by $A = T$, where

$$T = \sum_{x=0}^{M} x\{\hat{F}(x) - \hat{F}(x+1)\}.$$

Because the PL estimator does not depend on parametric assumptions about the scores, A is a natural nonparametric estimator of the mean. Interestingly, White (1945) suggested (but did not implement) a batting average based on a life-table. This is very much in the same spirit as our proposal.

It is easy to show that, if a batsman has no not-out scores, then A and B coincide. Also, provided that $M > M^*$, it follows that A cannot exceed M. This is not the case for B. For example, Alan Hansford had a highest first-class score of 18 for Sussex in 1989 but $B = 24$, whereas $A = 17$.

Consider Bradman's test career. His highest score is 334 and $B = 99.94$, whereas $A = 98.98$. He needed to score only four runs in his final innings to ensure B of 100 or more but was out for a duck. Had he been working with A instead, he would have needed to score 89!

However, we cannot use A as it stands with Bradman's first-class career because his highest first-class score is 452 not out, and his highest completed innings is 369. So the PL estimator is undefined beyond 369. There is no totally nonparametric way round this: some assumption about tail behaviour is necessary. Classical extreme value results (see, for example, Resnick (1987)) are of limited use here because they require an underlying continuous distribution. The discrete case (see Anderson (1970) and Kimber (1983)) is less straightforward. However, since we have some evidence that the upper tail of a score distribution is roughly geometric, the following approach allows a modified version of A to be computed.

Suppose that $M \leqslant M^*$, that there are k not-out scores at least as big as M, that their sum is S and that the probability mass unassigned by the PL estimator is R. Then define A in terms of T above, S, k, and R as $A = T + R(A + S/k)$. Hence,

$$A = \frac{T + RS/k}{1 - R}.$$

This essentially converts each not-out innings of M or more to an estimated completed score by adding on the overall average. Note that in the calculation of B *every* not-out score is converted to an estimated completed score by adding on the overall average.

For Bradman's first-class scores $M = 369$, $T = 90.4990$, $R = 0.007\,801$, $k = 1$ and $S = 452$. Hence $A = 94.76$, compared with $B = 95.14$. Table 3 shows some further examples.

3.2. *Comments*

The difference between A and B for a batsman depends on his profile of scores. For many the difference is small. However, for a non-negligible proportion of batsmen the difference is substantial. A typical case in which A would be considerably less than B is for a player with a relatively high proportion of not-out scores who does not make any very high scores; see, for example, Monte Lynch

TABLE 3

Summary statistics for various players in tests, first-class cricket and 1983 county championship (CC83) matches: traditional (B) and new (A) batting averages, selected centiles and probability masses in four ranges of scores

Player	*Type of cricket*	*B*	*A*	*Centile*					*Probability mass (%) for the following scores:*			
				10th	*25th*	*50th*	*75th*	*90th*	*0–9*	*10–49*	*50–99*	*100+*
Armarnath	Tests	42.50	41.45	0	8	35	63	100	27.56	39.13	22.84	10.47
Bradman	Tests	99.96	98.98	1	16	66	169	244	17.50	27.88	14.97	39.65
Gower	Tests	44.31	44.30	5	11	28	61	102	21.00	47.91	20.04	11.05
Bradman	First class	95.14	94.76	3	22	66	139	232	15.39	26.89	19.33	38.39
Trumper	First class	46.58	47.36	2	10	28	67	110	23.87	41.70	22.13	12.30
Clarke	CC83	14.25	14.15	0	2	12	19	37	41.67	58.33	0.00	0.00
Lynch	CC83	53.72	49.11	2	13	36	78	112	21.53	33.80	29.85	14.82
Richards (C. J.)	CC83	27.61	27.46	1	6	21	43	79.46	32.48	56.95	5.28	5.28
Stewart	CC83	31.30	30.71	5	8	20	43	82	25.00	56.82	18.18	0.00

in Table 3. In cases of a huge proportion of not-out innings, corresponding to heavily censored data, A is no more sensible than B. For example, in the 1953 English first-class season, Bill Johnston was not out in 16 of his 17 innings, with a highest score of 28 not out, giving $B = 102$ and $A = 81$.

A disadvantage of A is that it must be recalculated from scratch when new scores arise. However, the calculations needed are easily programmed in, for example, a spreadsheet.

We did not try to find a fully parametric model to replace the geometric model. This might be an interesting exercise but we feel that it is unlikely that a reasonably simple model could be found that is satisfactory for *all* batsmen. Our nonparametric approach, in contrast, is entirely data driven. Another approach, suggested by a referee, is as follows. Define a cut-off point τ ($\tau = 0, 1, \ldots, M$) above which a batsman's hazard is thought to be flat. An average, C_τ say, can then be obtained by using the PL estimator for scores less than τ and by treating scores of τ or more as if they came from a (shifted) geometric distribution, i.e.

$$C_\tau = \sum_{x=0}^{\tau-1} x\{\hat{F}(x) - \hat{F}(x+1)\} + \hat{F}(\tau)\left\{\tau + \sum\nolimits_\tau (x_i - \tau) \Big/ \sum\nolimits_\tau d_i\right\},$$

where Σ_τ denotes summation over the set $\{i: x_i \geqslant \tau\}$ and where the first sum is 0 for $\tau = 0$. If we take τ sufficiently large (small), then C_τ will coincide with A (B). We might expect a good choice of τ to be between 10 and 20. However, especially for scores in a season, the results were rather sensitive to the choice of τ and were somewhat unconvincing. For example, for Monte Lynch's 1983 scores B seemed intuitively to be too high (confirmed by a much lower A-value) but, for all $1 \leqslant \tau \leqslant 23$, $C_\tau > B$. Also, C_τ does not take account of individual foibles; see observation (b) in Section 2.3. Thus, we did not investigate C_τ further.

4. SUMMARIZING DISTRIBUTIONS OF SCORES

Given the nature of distributions of scores, it is clear that more than just a single number is needed adequately to summarize a player's scores. An obvious method is to calculate selected centiles. This is standard in survival analysis. The PL estimator may be used for this, though some high centiles may be undefined for a batsman with several high not-out innings. Either lower bounds for these missing centiles may be given or the geometric tail approximation may be used, as for A. The choice of centiles is arbitrary but for illustration we have used the 10th, 25th, 50th, 75th and 90th; see Table 3 for some examples. Here the geometric tail approximation has been used once (Jack Richards), giving an estimated 90th centile of 79.46. Without the approximation it is known only to be 52 or more.

The skewness of distributions of scores suggests that low centiles may not be very informative. The 10th and 25th centiles tend to be low for all batsmen; even Bradman's 25th centile is only 16 in tests. This suggests that perhaps the 50th, 75th and 90th centiles give a reasonable summary. Note that the median (50th centile) is not *on its own* a serious rival to B or A as an average since it ignores the magnitudes of high scores: the ability to build a big innings is an important factor in batsmanship.

A second method is to summarize the distribution of scores by probability masses in selected ranges of scores. As an illustration we have used the ranges 0–9, 10–49,

50–99 and 100 or more; see Table 3. Again, the PL estimator may be used to calculate the probabilities. If $M^* < M$ or if $M > 99$, then all probabilities for this grouping may be calculated. Otherwise, the geometric tail approximation may be used, as has been done in Table 3 for Jack Richards. Without this it would have been known only that the unassigned probability masses in the ranges 50–99 and 100 or more sum to 10.57%.

5. CONCLUDING REMARKS

We have seen that B is not just non-optimal but is actually inconsistent for the underlying mean score. Also, it is clear that a one-number summary of the distribution of a batsman's scores is not enough. However, batting statistics in cricket have been presented in basically the same way for many years. So it would be naïve to suppose that a proposal for radical changes in the presentation of batting statistics would be widely accepted by cricketers and cricket statisticians. However, we believe that the following compromise contains our major ideas while little altering the basic appearance of the data summary.

Consider the standard presentation of a batsman's performance. We use Bradman's test career as an illustration:

I	*NO*	*Runs*	*H*	*B*	*100*	*50*
80	10	6996	334	99.94	29	13.

Here I denotes number of innings, NO denotes number of not-out innings, 100 denotes the number of scores of 100 or more and 50 denotes the number of scores between 50 and 99. Our compromise suggestion is to replace B by A and to augment the cumulative 100 and 50 columns with the corresponding (percentage) probability masses. Thus, Bradman's figures become

I	*NO*	*Runs*	*H*	*A*	*100*	*50*
80	10	6996	334	98.98	29 (39.7%)	13 (15.0%).

This gives a more sensible average and information (adjusted for not-out innings) on the rate of achieving good scores (and the corresponding rate for low scores by subtraction from 100%), as well as cumulative figures. This is *not* saying that 39.7% of Bradman's 80 innings were centuries. The stated figure is necessarily higher than the observed proportion of centuries (29/80 = 36.25%) since for each not-out score below 100 there was the possibility that he would have gone on to make a century. Table 4 gives some examples.

There are many ways in which our analysis could be refined and made necessarily more complicated. For example, innings could be subdivided to take account of lunch, tea and overnight intervals, drinks breaks, stoppages for rain and bad light, dropped catches etc. However, we chose to treat an innings as the fundamental entity for simplicity, the availability of data and ease of comparison with existing methods used by cricket statisticians.

A topic for further research is how various factors, such as scoring rate, strength of opposition bowling and the state of the pitch, can be *properly* combined with runs scored to give an overall picture of the relative merits of batsmen. A scheme, originally called the Deloitte ratings (now the Coopers and Lybrand ratings), is already used for calculating the current worth of test players. It involves adjustment

TABLE 4
Summary of nine sets of scores from tests, first-class cricket and 1983 county championship (CC83) matches†

Player	*Type of cricket*	*I*	*NO*	*H*	*Runs*	*A*	*100*	*(%)*	*50*	*(%)*
Armarnath	Tests	113	10	138	4378	41.45	11	(10.5)	24	(22.8)
Bradman	Tests	80	10	334	6996	98.98	29	(39.7)	13	(15.0)
Gower	Tests	200	16	215	8154	44.30	18	(11.1)	39	(20.0)
Bradman	First class	338	43	452‡	28067	94.76	117	(38.4)	69	(19.3)
Trumper	First class	395	21	300‡	17420	47.36	45	(12.3)	86	(22.1)
Clarke	CC83	24	4	43	285	14.15	0	(0.0)	0	(0.0)
Lynch	CC83	39	10	119	1558	49.11	3	(14.8)	11	(29.9)
Richards (C. J.)	CC83	34	8	85‡	718	27.46	0	(5.3)	2	(5.3)
Stewart	CC83	16	3	82	407	30.71	0	(0.0)	2	(18.2)

† *I* is the number of innings, NO is the number of not-out innings, 100 is the number of centuries, 50 is the number of innings between 50 and 99.
‡ Not-out score.

of players' scores for various factors, followed by the use of exponential smoothing. Partial details are given in Berkmann (1990). The statistical properties of this scheme have not, we believe, been fully investigated.

ACKNOWLEDGEMENTS

We thank Trevor Sweeting for helpful discussions, Alistair Fitt and Ron Shail for supplying some elusive scores and a referee for recommending the Elderton and Wood papers.

APPENDIX A

Table 5 is a list of the 110 sets of scores used. The career records for current players are as at the end of October 1990 when this study was begun. An exception is David Gower's test career which includes all scores up to and including his record breaking innings of July 6th, 1992.

APPENDIX B

Let $y_1, y_2, \ldots$ and $c_1, c_2, \ldots$ be sequences of observations on random variables Y and C respectively. Define x_i to be y_i if $y_i < c_i$ and c_i otherwise, so that $x_1, x_2, \ldots$ is a sequence of observations on a random variable X, say. Then, in the notation of Section 2.2, $d_i = 1$ if $y_i < c_i$ and $d_i = 0$ otherwise. The usual batting average based on n scores may be rewritten as

$$B = n^{-1} \sum_{i=1}^{n} x_i \Big/ n^{-1} \sum_{i=1}^{n} d_i.$$

By the strong law of large numbers, the numerator and denominator of the above tend almost surely to $E(X)$ and $P(Y < C)$ respectively as $n \to \infty$. Thus, B is consistent for $E(Y)$

TABLE 5
List of scores used

Test scores			
M. Armarnath	A. R. Border	I. T. Botham	D. G. Bradman
J. H. Edrich	J. E. Emburey	M. W. Gatting	G. A. Gooch
D. I. Gower	C. G. Greenidge	R. J. Hadlee	D. L. Haynes
M. A. Holding	K. J. Hughes	Imran Khan	Javed Miandad
Kapil Dev	S. M. H. Kirmani	A. J. Lamb	M. D. Marshall
Mudassar Nazar	I. V. A. Richards	R. J. Shastri	I. D. S. Smith
R. W. Taylor	J. R. Thomson	D. B. Vengsarkar	G. M. Wood
J. G. Wright			
One-day international scores			
A. R. Border	P. J. L. Dujon	S. M. Gavaskar	D. I. Gower
C. G. Greenidge	D. L. Haynes	Imran Khan	Javed Miandad
Kapil Dev	Mudassar Nazar	Rameez Raja	I. V. A. Richards
R. B. Richardson	Salim Malik	R. J. Shastri	K. Srikkanth
D. B. Vengsarkar	J. G. Wright		
First-class centuries			
D. L. Amiss	G. Boycott	D. G. Bradman	M. C. Cowdrey
J. H. Edrich	J. B. Hobbs	I. V. A. Richards	G. M. Turner
Zaheer Abbas			

Others
All first-class scores for D. G. Bradman and V. T. Trumper; see Bradman (1950) and Fingleton (1978)
W. A. Johnston (1953 first-class scores)
County championship scores for Surrey in 1983 (16 sets of scores), Surrey in 1991 (16 sets of scores) and Sussex in 1989 (17 sets of scores)
A. C. Kimber (scores for Guildford and for University of Surrey Staff)

if and only if $E(X) = E(Y)\,P(Y < C)$. The following result shows that in the present context B is consistent if and only if the underlying score distribution Y is geometric.

Proposition. Let Y be a non-negative integer random variable with support S_Y, hazard function $h(\ )$ and finite mean. Let C be an arbitrary random variable with the same support as Y and the following properties:

(a) Y and C are independent;
(b) $0 < P(Y < C) < 1$.

Then $E(X) = E(Y)\,P(Y < C)$ *for all* such C if and only if $h(\ )$ is constant.

Proof. It is easy to show that

$$E(X) - E(Y)\,P(Y < C) = \sum_{y=0}^{\infty} P(C > y)\{P(Y > y) - E(Y)\,P(Y = y)\}.$$

Since C is arbitrary, it follows that this expression is 0 if and only if $P(Y > y) - E(Y)\,P(Y = y) = 0$ for all y. This in turn is true if and only if $h(y) = 1/\{E(Y) + 1\}$ for all y. This completes the proof.

REFERENCES

Anderson, C. W. (1970) Extreme value theory for a class of discrete distributions with applications to some stochastic processes. *J. Appl. Probab.*, **7**, 99–113.

Ascher, H. and Feingold, H. (1984) *Repairable Systems Reliability*, pp. 78–79. New York: Dekker.
Becker, R. A. and Chambers, J. M. (1984) *S: an Interactive Environment for Data Analysis and Graphics*, p. 201. Belmont: Wadsworth.
Berkmann, M. (1990) *The Complete Guide to Test Cricket in the Eighties*, pp. 262–266. London: Partridge.
Bradman, D. G. (1950) *Farewell to Cricket*. London: Hodder and Stoughton.
Cox, D. R. and Hinkley, D. V. (1974) *Theoretical Statistics*, pp. 287–293. London: Chapman and Hall.
Cox, D. R. and Oakes, D. (1984) *Analysis of Survival Data*, pp. 48–51. London: Chapman and Hall.
Crowder, M. J., Kimber, A. C., Smith, R. L. and Sweeting, T. J. (1991) *Statistical Analysis of Reliability Data*. London: Chapman and Hall.
Elderton, W. (1945) Cricket scores and some skew correlation distributions. *J. R. Statist. Soc.*, **108**, 1–11.
Fingleton, J. H. (1978) *The Immortal Victor Trumper*. London: Collins.
Kaplan, E. L. and Meier, P. (1958) Non-parametric estimation from incomplete observations. *J. Am. Statist. Ass.*, **53**, 457–481.
Kendall, M. G. and Ord, J. K. (1990) *Time Series*, 3rd edn, p. 130. Sevenoaks: Arnold.
Kimber, A. C. (1983) A note on Poisson maxima. *Z. Wahrsch. Ver. Geb.*, **63**, 551–552.
Resnick, S. I. (1987) *Extreme Values, Regular Variation and Point Processes*, ch. 1. New York: Springer.
Tukey, J. W. (1977) *Exploratory Data Analysis*, ch. 7, 16. Reading: Addison-Wesley.
White, G. R. (1945) Discussion on Cricket scores and some skew correlation distributions (by W. Elderton) and Cricket scores and geometrical progression (by G. H. Wood). *J. R. Statist. Soc.*, **108**, 28–29.
Wood, G. H. (1945) Cricket scores and geometrical progression. *J. R. Statist. Soc.*, **108**, 12–22.