WORKING PAPER NO: 356

On Generalized Geometric Distributions: Application to Modeling Scores in Cricket and Improved Estimation of Batting Average in light of Notout Innings

Shubhabrata Das
Professor
Indian Institute of Management Bangalore
Bannerghatta Road, Bangalore – 5600 76
Ph: 080-26993150
shubho@iimb.ernet.in

Year of Publication 2011

On Generalized Geometric Distributions: Application to Modeling Scores in Cricket and Improved Estimation of Batting Average in light of Notout Innings

Shubhabrata Das

Indian Institute of Management Bangalore
Bannerghatta Road, Bangalore 560076 India

Abstract

In the game of cricket, batting average is the most common and basic measure of a batsman's performance during a short duration, like a series or calendar year, as well over a longer span like the career. Batting average is considered in isolation or in combination with other measures like strike rate, at times depending on the form of the game. However, in either case, treatment of runs scores from notout innings throws particular challenge in adopting batting average as a measure of true performance. The conventional way of computing batting average enjoys favour as well as criticism from intuitive standpoint — but it can be justified as the maximum likelihood estimate if the scores come from an Exponential or Geometric distribution. Either of these distributions is quite unreasonable in modeling cricket scores of a batsman because of obviously non-constant hazard or propensity to get out after scoring different runs. Towards this, we discuss the role of the Kaplan Meir estimator treating the scores from the notout innings as right censored data. We show that while it provides a vast conceptual improvement over the traditional average, there are some associated some problems as well. The first of these is because of its nonparametric nature, specially in the context of reflecting true average performance in a short duration like a tournament or a series — the other because of its inability to produce a finite-valued estimate when the largest score is from a notout innings.

To address these concerns, we propose a generalized class of Geometric distributions (GGD) as model for the runs scored by individual batsmen. The generalization comes in the form of hazard of getting out changing from one score to another. We consider the change points as the known or specified parameters and derive the general expressions for the restricted maximum likelihood estimators of the hazard rates under the generalized structure considered. Given the domain context, we propose and test ten different variations of the GGD model and carry out the test across the nested models using the asymptotic distribution of the likelihood ratio statistic to determine the best possible model. This family of GGD subsumes the traditional average as well as the Kaplan-Meir based estimate, as the 1 parameter GGD is the simple Geometric distribution, while the infinite order GGD corresponds to the non-parametric Kaplan-Meir based survival function. Finally to estimate the true batting average, we propose two methods: first being the simple mean of the fitted GGD and in the second case the notout scores are replaced by conditional mean of the fitted GGD, before averaging out. We show that while the two methods coincide for the two extreme GGD (simple Geometric and nonparametric) it is not so in general. We also discuss how different approaches for estimating average over a short or long time horizon. Finally we compute batting averages by the different methods for all top players, in both forms of the game and study the rank correlation. We also present results from numerical computation is carried out using scores of all opening batsmen as well as No 11 batsmen in one day cricket matches, to illustrate model selection procedures. This

also establishes that any model in the family need not be appropriate for all situation. We also focus on Batting average of two players. In particular, we show that quite possibly Bradman's true average was greater than 100, while Bevan have been distinctly beneficiary of prevalent way of computing average as his 1-day average seems to be an overestimate by fair degree.

Keywords: Average, Censored, Hazard, Kaplan Meir estimator, (restricted) maximum likelihood.

1 Introduction

The game of cricket is essentially about scoring runs and taking tickets. The first of these is primarily the job of the batsmen[1], although often everybody in the team gets a chance to contribute. In that sense, the number of runs scored in an attempt (called inning) is typically the most common (if not the only) indicator of contribution made. (Although it is also important, especially in the shorter versions of the game, how quickly these runs are scored. The latter is captured through *strike rate*, i.e. runs scored per ball faced; however for the current purpose and focus, we discuss only the number of runs scored and ignore the strike rate.) With that in mind, we note that batting average, one of the most popular forms of statistics used in cricket to signify the contribution of the batsman over a period of time or series. Simplicity is one of the reasons for the popularity the batting average in terms of its widespread use in the context of the game analysis. It is a very useful measure (e.g. as opposed to the total runs scored) for comparing performance between players, series (of games) or even between two different generations as it adjusts for the opportunities present for the batsman in a very simple and intuitive manner. In every team inning, all the players in the team need not get a chance to bat and hence to score runs. Even when a batsman gets chance to bat, he may score some runs (nonnegative integer) before getting 'out' or remain 'notout'. For example, if a batsman scores 45 notout in an inning, it is not certain how much he would have scored had the circumstances (game rule/ team's decision, at times) allowed, but obviously it would have to be a positive integer bigger than or equal to 45.

Traditionally in cricket analysis and journalism, batting average is believed to take care of adjustments in terms of opportunities present to the players while reporting his average contribution (as a batsman) in a simple manner. As is the case with most single summary statistic, it is unfair to have the expectation of getting ideal communication in all circumstances. At the same time, it

[1]Although instead of *batsman* gender-neutral terminology *batter* is being used increasingly recently, here we have continued with conventional terminology, 'batsman', in recognition with the history of the game.

Table 1: Examples of Traditional Batting Average exceeding even the Maximum score

Player	A Kumble	DJ Nash	H Tillakaratne
Series Year	2004-05	1998-99	2001-02
opponent	Pakistan	India	West Indies
Test1 -inning1	1*	89*	105*
Test1 -inning2	—	4*	87
Test2 -inning1	21*	18*	7*
Test2 -inning2	14*	63	204*
Test3 -inning1	22		
Test3 -inning2	37*		
Traditional batting average	**95**	**174**	**403**

is important to diagnose if it suffers from a systemic drawback and to adjust the statistic rectifying the problem, if there is one.

It is our assertion that the batting average systematically falls short of reflecting the true average performance of a batsman , certainly for a short duration (like a series) and possibly even over a longer span like the career of the player. To explain this, let us formally state the formula for computing the batting average, as is done in practice:

$$M = \frac{\text{Total no.of runs scored}}{\text{No.of times batsman got out}} \tag{1}$$

Note that the denominator is *not* the number of innings the batsman gets to bat (which was the practice in early years of cricket statistics). The use of M seems more generous to the batsman ; what is not clear whether this generosity gets in the way of its intended purpose. Consider for example, scores of a batsman in a particular series:

$$45* \quad 20* \quad 15* \quad 60 \quad 30* \quad 50* \quad 70 \quad 35* \quad 55* \quad 40$$

where * denotes scores in notout innings. The average score of the batsman [M as in (1)] in the series would be 420/3=140! Can we expect the batsman to score 140 in his next inning? Surely not. Enclosed below are 3 real examples of batting averages being unrepresentative of the true performance different 2 or 3 test match series. In all the examples, the traditional average far exceeds even the maximum score and hence falls sort of being a good estimator of the true mean capability of the player.

A key motivation of this research is to expose the limitation of traditional batting average when the batsman is not out on multiple occasions and suggest suitable remedial action.

Noting that the scores from the notout innings are nothing but right-censored data, there is a

standard approach to solving this problem through Kaplan-Meier (K-M) estimator. We present the use of K-M estimator intuitively as well as form the angle of nonparametric maximum likelihood estimation in Section 2 and present K-M average for top test batsmen. K-M approach has certain limitation which we highlight here. Our subsequent approach tries to eradicate these problems build on experience on trying K-M approach. Towards this, we adopt a parametric maximum likelihood approach for censored data and build a class of parametric distributions as model for distribution of runs scored. The discussion of this model, which we refer to as the class of Generalized Geometric distributions (GGD), is done in Section 3. Here we also consider different subclasses under this family, as appropriate under the different contexts or players and discuss method of picking appropriate one of them. Estimation of parameters by (restricted) maximum likelihood methods in this family is also discussed in Section 3. We explicitly propose two different approaches for adaptations of the model depending on the context in Section 4. In this section, we present detailed numerical computations reflecting up to date averaged by the different methods for the best batsmen in both test match cricket as well as 1 day games, We also focus on the controversial and highly debated case of batting average of Bradman in test cricket and Bevan in 1-day cricket. Our detailed model selection illustrations for 1day openers vis-a-vis No. 11 batsmen show requirement for considering different GGD models for different players, position or context. We conclude with few relevant comments and summary in Section 6.

2 Estimating True Batting Average: The Parametric and Nonparametric MLE

2.1 Notation

Let us introduce the notation for data at this stage now. Let $X_1, \ldots, X_n$ be the scores from the n innings when the batsman is out and $Y_1, \ldots, Y_m$ be the scores from the m notout innings. Suppose the highest score possible is K. For $0 \leq i \leq K$, let f_i denote the the number of times the batsman scored i and got out at the score i; also let f_i^* be the number of times, the batsman scored i and remained notout at the conclusion of the innings. Note that some of the frequencies could be 0 as well. Thus, we have:

$$K = \max(\max_{1\leq i\leq n} X_i, \max_{1\leq j\leq m} Y_j),$$

$$f_i = \sum_{j=1}^{n} \mathbb{I}_{X_j=i}, \quad f_i^* = \sum_{j=1}^{m} \mathbb{I}_{Y_j=i},$$

where $\mathbb{I}$ is the indicator function. Let F and F^* denote the cumulative frequencies corresponding to f and f^*, i.e.

$$F_i = \sum_{j=0}^{i} f_j \quad \text{and } F_i^* = \sum_{j=0}^{i} f_j^*.$$

Also, M_i denote the total no of scores (from out or notout innings) bigger than or equal to i, i.e.

$$M_i = \sum_{j=i}^{K} (f_j + f_j^*) = (n - F_{i-1}) + (m - F_{i-1}^*). \tag{2}$$

2.2 The Kaplan-Meir Estimator

Kaplan and Meir [3] proposed a nonparametric estimator $\hat{S}_{KM}(t)$ for the survival function $S(t) = P(X > t)$ when some of the observations are (right) censored. Application of that with cricket score data where we have scores (observations) both complete (from 'out' innings) as well as incomplete (from 'notout' innings), would give:

$$\hat{S}_{KM}(i) = \prod_{j=0}^{i} (1 - \frac{f_j}{M_j}), j = 0, 1, 2 \ldots, \quad \text{and } \hat{S}_{KM}(t) = \hat{S}_{KM}([t]), \ \forall t, \tag{3}$$

where $[t]$ denotes the largest integer less than or equal to t.

There are many desirable properties of the Kaplan Meir estimator, like the self-consistency property (see, [2]). Most notably, specially in our context, it is (nonparametric) maximum likelihood estimator of the survival function. We recap these results in the next subsection.

It is also very intuitively appealing to understand the Kaplan Meir approach through 'redistribute to the right' principle, which goes as follows:

1. Put weight $\frac{1}{m+n}$ to all observations including the censored ones.
2. Now, starting from the smallest, redistribute the *latest* weight of the censored data, one at a time, to all observations (complete or censored) bigger than it.
3. The K-M estimator of the survival function is nothing but the survival function of the eventual (discrete) distribution with masses only at the complete observations, once the weights of all censored data have been redistributed.

Note the K-M estimator is designed for estimating the survival function. To estimate the estimate the population mean, the natural approach would be to take the mean of distribution associated the K-M survival function. Mean of any random variable X taking nonnegative integral

values, $\sum_{n=1}^{\infty} n \times P(X = n)$ can be reorganized as $\sum_{i=0}^{\infty} P(X > i)$. Since the batting score takes only nonnegative integer values, its mean as given by the Kaplan Meir approach would be:

$$M_{KM} = \sum_{i=0}^{K} \hat{S}_{KM}(i) = \sum_{i=0}^{K} \prod_{j=0}^{i} (1 - \frac{f_j}{M_j}). \tag{4}$$

An equivalent expression for the above is given by ([1]), but is not of much help computationally. Many software packages including SAS report this mean routinely.

In Table 2 we present the computations of K-M average (M_{KM}) vs. traditional average (M) for some well known batsmen for both 1-day and tests.

Table 2: Batting averages of Top Test Batsmen: Traditional vs. Kaplan-Meir

Batsman	# of Innings	# of notout	highest score	traditional average	Kaplan Meir estimate			
					average	Percentiles		
						75th	50th	25th
DG Bradman	80	10	334	99.94	98.98	169	66	17
RG Pollock	41	4	274	60.97	62.14	90	42	12
GA Headley	40	4	270*	60.83	60.89	105	29	11
H Sutcliffe	84	9	194	60.73	59.50	94	51	22
E Paynter	31	5	243	59.23	58.56	77	40	9
KF Barrington	131	15	256	58.67	57.75	83	47	16
ED Weekes	81	5	207	58.61	58.56	90	39	13
WR Hammond	140	16	336*	58.45	61.07	65	32	16
IJL Trott	38	4	226	57.79	58.23	69	33.5	12
GS Sobers	160	21	365*	57.78	61.03	72	39	14
JB Hobbs	102	7	211	56.94	56.53	76	45	20
KC Sangakkara	173	12	287	56.93	57.43	71	35	13
JH Kallis	250	39	201*	56.89	55.61	84	38	12
CL Walcott	74	7	220	56.68	56.67	98	39	14
L Hutton	138	15	364	56.67	57.39	79	38	13
SR Tendulkar	303	32	248*	56.02	55.91	83	36	11
GS Chappell	151	19	247*	53.86	54.01	70	36	12
AD Nourse	62	7	231	53.81	54.04	73	36	20
R Dravid	278	32	270	53.22	53.68	80	36	12
BC Lara	232	6	400*	52.88	53.04	74	35	8
TT Samaraweera	108	19	231	52.61	53.22	77	34	9
Javed Miandad	189	21	280*	52.57	53.87	66	32	13
RT Ponting	267	28	257	52.53	52.86	75	34	10
M Yousuf	156	12	223	52.29	51.16	77	32	11
V Sehwag	159	6	319	52.15	52.65	66	33	10
MEK Hussey	112	12	195	51.73	51.16	82	36	10
J Ryder	32	5	201*	51.62	52.09	79	33	6
A Flower	112	19	232*	51.54	53.28	67	31	10
DPMD Jayawardene	207	13	374	51.3	51.90	67	33	10
Y Khan	125	9	313	51.2	52.19	71	35	7
SM Gavaskar	214	16	236*	51.12	51.05	71	33	8
SR Waugh	260	46	200	51.06	51.20	77	28	10
ML Hayden	184	14	380	50.73	50.68	77	34	13
AR Border	265	44	205	50.56	50.15	75	34	11
KP Pietersen	133	7	227	50.48	50.51	72	34	12
IVA Richards	182	12	291	50.23	50.60	70	34	9
DCS Compton	131	15	278	50.06	50.48	72	30	13

Remarks and Observations from Table 2:

- The list includes batsmen having test average of 50 or higher – 4 players with less than 20 matches are excluded from this list.

- For players still playing test matches, matches up to November 26, 2011 have been considered.

- Since these are career averages, the difference between tradition and K-M averages are reasonably small. The difference also does not follow any pattern in terms of the percentage of notout innings, as one might like to believe as a quick guess.

- Indeed for some players, the K-M average is higher than the traditional average. This is best understood by careful introspection of the redistribute to the right principle of K-M method.

- This K-M approach in the context of batting average is possibly first documented in statistical literature by Kimber et al. [4].

- Routine implementation of Kaplan Meir using softwares like SAS would yield slightly different values (e.g. Bradman's K-M average is given as 99.06 using SAS). This is because of discreteness of the data in the current context. A continuous a data-point right-censored at 50.23, means the complete observation, had it been observed, would be something strictly bigger than 50.23. However a notout score of 50 should be interpreted as the complete observation being 50 or more in that case. This adjustment has been taking into account in (4) – but not in many commercial software primarily meant for continuous data.

- Although we focus only on the averages in the present work, we are in agreement with [4] in observing that to compare performance it is advisable to compare the entire distribution or certain major percentiles as opposed to looking at only the mean. With that in mind, the estimates of the quartiles as obtained by the K-M method (using SAS) are also reported in Table 2.

- Among other facts, this table shows how Bradman's performance was superior to the other greats. Not only in terms of average, which is well known in the cricketing world, but also in the form of (K-M) 75th or 50th percentile, he is ahead by more than 60% of his nearest competitor. Such is the dominance of Bradman, that many consider him to be the greatest sportsman considering all sports. It is also because of that his so-called failure to reach the milestone of average of 100 is so talked about. As we show later, perhaps unfairly — as

his true average possibly exceeded 100, even though it is not illustrated by the traditional average or even the K-M estimate.

2.3 Batting Average and Maximum likelihood Estimation: Parametric and Nonparametric approach

First we show that the traditional batting average can be justified from the angle that it is the MLE if the batting scores come from an Exponential Distribution or Geometric distribution. Next we illustrate that the Kaplan-Meir estimator comes into the picture if one consider only non-parametric MLE. All the three results are well known results in statistics. We replicate them here for the sake of completeness and showing their adaptability to the cricket score data. This also introduces the necessary background and opens the door for introduction of the Generalized Geometric distribution, as discussed in the next section.

Lemma 1 *If* $X_1, \ldots, X_n$ *and* $Y_1, \ldots, Y_m$ *are respectively complete and right-censored i.i.d. observations drawn from an Exponential distribution, the maximum likelihood estimate of the mean of the population distribution is given by:*

$$\hat{\mu} = \frac{\sum_{i=1}^{n} X_i + \sum_{j=1}^{m} Y_j}{n}, \tag{5}$$

which is nothing but the traditional batting average (1).

Proof. Using the notation for data, as given in the beginning of this section, the likelihood becomes

$$L(X_1, \ldots, X_n;\ Y_1, \ldots, Y_m) = \prod_{i=1}^{n} \phi(X_i) \prod_{i=1}^{m} S(Y_j-), \tag{6}$$

as a function of the density function $\phi(\cdot)$ and survival function$S(\cdot)$, with $S(y-) = P(Y \geq y) = \lim_{x \to y+} S(x)$. If the batting scores are from an Exponential distribution with parameter λ, $\phi(x) = \lambda e^{-\lambda x}$, we have $S(y) = S(y-) = e^{-\lambda y}$. In this case, (6) reduces to

$$L = \lambda^n \exp\{-\sum_{i=1}^{n} X_i - \sum_{j=1}^{m} Y_j\}$$

. To maximize L, as usual, we take logarithm and derivative w.r.t. λ, leading to to the MLE

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} X_i + \sum_{j=1}^{m} Y_j}.$$

Using the transferability property of the MLE, the MLE of the mean of the Exponential distribution $\frac{1}{\lambda}$ is given by (5), or the traditional batting average formula, as given in (1). ■

The problem with using (1) is the limitation of Exponential distribution as a model of scores, not just because it is continuous while score distribution is discrete, but also because of poor fit shown empirically in many studies ([1]), [6], among others). Next we show that the conclusion of the previous lemma continue to hold, if the scores are from a Geometric distribution.

Lemma 2 *Suppose the complete or right-censored observations come from a Geometric distribution with parameter* p*:*

$$\phi(x) = \alpha(1-\alpha)^x, \quad x = 0, 1, 2, \ldots. \tag{7}$$

Then the maximum likelihood estimate of the mean of the population distribution, namely $\frac{1}{\alpha} - 1$ *is (again) given by the traditional batting average (5).*

Proof. The likelihood function with the complete observations $X_1, \ldots, X_n$ and right censored observations $Y_1, \ldots, Y_m$ now becomes

$$L(X_1, \ldots, X_n;\ Y_1, \ldots, Y_m) = \alpha^n (1-\alpha)^{\sum_{i=1}^{n} X_i + \sum_{j=1}^{m} Y_j}, \tag{8}$$

$$\frac{\delta \ln L}{\delta \alpha}|_{\alpha=\hat{\alpha}} = 0 \Rightarrow \frac{n}{\hat{\alpha}} = \frac{\sum_{i=1}^{n} X_i + \sum_{j=1}^{m} Y_j}{1-\hat{\alpha}}$$

$$\text{or} \quad \hat{\alpha} = \frac{n}{n + \sum_{i=1}^{n} X_i + \sum_{j=1}^{m} Y_j},$$

and hence the MLE of the mean is given by

$$\frac{1}{\hat{\alpha}} - 1 = \frac{\sum_{i=1}^{n} X_i + \sum_{j=1}^{m} Y_j}{n}.$$

■

The following result shows that Kaplan-Meir mean is the nonparametric MLE.

Lemma 3 *Without considering a parametric form for the population distribution, the nonparametric MLE turns out to be (4), or the mean of the distribution corresponding to the Kaplan-Meir survival function.*

Proof. If we do not consider any parametric model for scores, and consider only the empirical distribution, (6) becomes,

$$\prod_{i=0}^{K}\left[\left\{S(i-)-S(i)\right\}^{f_i}\left\{S(i-)\right\}^{f_i^*}\right] \tag{9}$$

by expressing it only in terms of the survival function $S(\cdot)$. We need to maximize this over all possible choices of survival function $S(\cdot)$ which needs to non-decreasing in nature. We observe that, the maximization would occur if $S(i-)=S(i-1), \forall i$ and $S(0-)=1$. Now using the notation, $S(i)=s_i$, (9) becomes:

$$L=(1-s_0)^{f_0}\times\prod_{i=1}^{K}\left\{(s_{i-1}-s_i)^{f_i}s_{i-1}^{f_i^*}\right\}.$$

Using $\alpha_0=s_0, \quad \alpha_i=\frac{s_i}{s_{i-1}}, i\geq 1$, we get

$$s_i=\prod_{i=0}^{i}\alpha_i, \quad \text{and for } i\geq 1, \quad s_{i-1}-s_i=(1-\alpha_i)\prod_{i=0}^{i-1}\alpha_i.$$

Now rearranging and collating terms leads to:

$$L=\prod_{i=0}^{i}(1-\alpha_i)^{f_i}\alpha_i^{M_{i+1}} \tag{10}$$

with M_i being as given in (2). Now standard maximization leads:

$$\alpha_i=\frac{f_i}{f_i+M_{i+1}}, \tag{11}$$

implying that the MLE is given by (4). ■

2.4 Limitation of Kaplan Meir approach in estimating true batting average

Geometric (and Exponential) distribution provide a poor fit to the batting scores — a typical batsman's propensity to get out is not constant at the different possible scores. Hence the traditional

average (1) is unsatisfactory estimate of the true average, in light of notout innings.

In spite of vast conceptual improvement he K-M estimator provides in estimating true mean in light of the censored data, i.e. scores from notout innings, there are couple associated concerns or difficulties as well in applying to batting score context. First, the philosophy of K-M estimator is that it is entirely non-parametric and (empirical) data driven. Note that the distribution suggested by the K-M survival function puts probability mass only at the scores when the player has got out in the past (or in the relevant time domain). Thus, if a player has never got out at a score of 10, as per the K-M estimator, the chance of him getting out be taken as 0 in the K-M estimate. This would not find favour with many practitioners; the problem would be compounded specially when one might be looking at the performance in a calender year or series (with small data points). The issue is addressed in the approaches proposed in the next section by considering a general probabilistic model which would allocate certain positive probability to all plausible scores (values) in the range.

The second aspect of difficulty with adopting K-M approach for batting score arises in connection with finiteness of the computed value. Note that the summation in (4) should technically go up to ∞; but we ignore the terms for score $i > k$. This is problematic as, if

$$\max_{1\leq i\leq n} X_i \leq \max_{1\leq j\leq m} Y_j$$

i.e. (one of) the highest score is from a notout innings, $f_K \neq N_k$, implying $\hat{S}_{KM}(K) \neq 0$, and consequently

$$\hat{S}_{KM}(t) > 0 \; \forall t,$$

i.e. there would be a positive probability mass at ∞ as per K-M distribution. This would lead to an infinite mean which is meaningless in the current context. A simple crude way of practically resolving the issue, if adopting the K-M approach would be to interchange the highest score from a notout innings with the highest score from the out innings, in case the latter is not larger than the former. (An alternative would be to treat all the notout innings with the highest scores from out innings; but we stick to the above choice as this adjustment does not alter the computation of the traditional average.) However, this naturally suffers from arbitrariness.

Examples given in Table 1 in Section 1 highlight these concerns. First, in each of the 3 cases, the highest score is from a notout innings. Consequently, the mean corresponding to the Kaplan-Meir survival function is infinite. Even if we swap the highest notout score by the highest out score,

since there is only 1 innings with out score, the Kaplan Meir average would reduce to this single value, which happens to be the maximum score. This is perhaps unsatisfactory. In the last section of this paper, we outline a procedure where possibly a more reasonable estimate can be arrived while considering average in a series or tournament like this.

3 Generalized Geometric Distribution (GGD)

A random variable X is said to be having a Geometric distribution if it has a density function given by

$$P(X = i) = \alpha(1-\alpha)^i, \quad i = 0, 1, 2, \ldots. \tag{12}$$

Among other aspects, it is characterized by constant hazard; i.e.

$$P(X = i | X \geq i] = \alpha,$$

not depending on the value i.

In the context of a batsmen's score X, it is unrealistic since the batsmen's propensity to get out at a certain score given that he had survived up to that point is typically not the same for all scores. Usually, the batsmen is more vulnerable at the beginning of the innings and also after reaching the hundred, if he does so. Different batsmen show different pattern and capabilities. Some of them tend to get out in the twenties; a few of them appears to be more vulnerable in the nineties.

Motivated by this, under complete generality, we propose a family of distributions via the sequence of hazard at the score $i = 0, 1, 2, \ldots$ by,

$$\alpha_i = P[X = i | X \geq i].$$

Thus, the distribution which call a Generalized (family) of Geometric Distribution or (GGD), has a density:

$$\phi(0) = \alpha_0, \quad \phi(i) = P(X = i) = \alpha_i \times \prod_{j=0}^{i-1}(1-\alpha_j), \quad i = 1, 2, \ldots. \tag{13}$$

Note that, equivalently the survival function $S(\cdot)$ is given by

$$S(i) = \prod_{j=0}^{i}(1-\alpha_j), \quad i = 0, 1, \ldots; \tag{14}$$

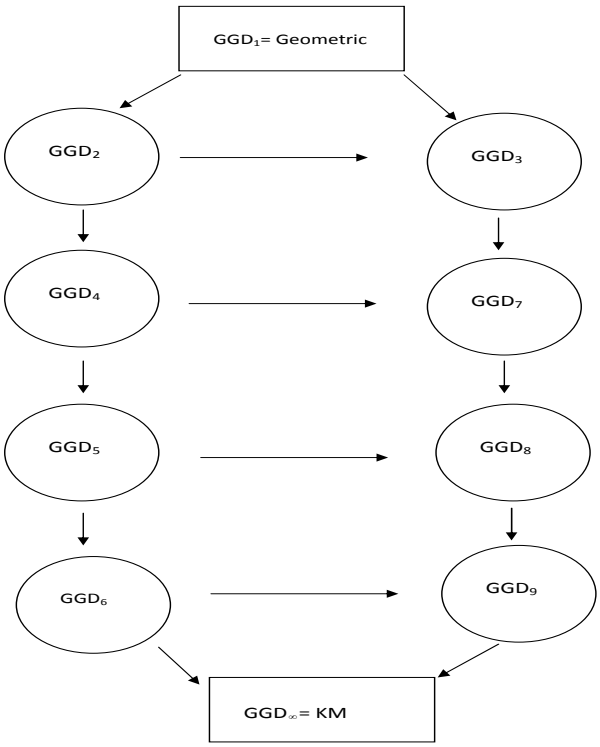
3.1 Models in the GGD Family

In practice, one may not want all the hazard rates to be different. Consequently we consider the following different variations/ constraints of the GGD distribution to model a batsman's score:

1. GGD_1: Traditional geometric distribution, as given in (12), – all hazard rates α_i's being equal;
2. GGD_2: α_0 is different, all other hazard rates equal;
3. GGD_3: α_0 and α_1 are different (from each other as well as other hazards rates), all other hazard rates equal;
4. GGD_4: α_0 is different, $\alpha_1 = \alpha_2 = \ldots = \alpha_9$, $\alpha_{10} = \alpha_{11} = \ldots$;
5. GGD_5: α_0 is different, $\alpha_1 = \alpha_2 = \ldots = \alpha_9$, $\alpha_{10} = \alpha_{11} = \ldots = \alpha_{99}$, $\alpha_{100} = \alpha_{101} = \ldots$;
6. GGD_6: α_0 is different, $\alpha_1 = \alpha_2 = \ldots = \alpha_9$, $\alpha_{10} = \alpha_{11} = \ldots = \alpha_{99}$, $\alpha_{100} = \alpha_{101} = \ldots \alpha_{105}$, $\alpha_{106} = \alpha_{107} = \ldots$;
7. GGD_7: α_0 and α_1 are different, $\alpha_2 = \alpha_3 = \ldots = \alpha_9$, $\alpha_{10} = \alpha_{11} = \ldots$;
8. GGD_8: α_0 and α_1 are different, $\alpha_2 = \alpha_3 = \ldots = \alpha_9$, $\alpha_{10} = \alpha_{11} = \ldots = \alpha_{99}$, $\alpha_{100} = \alpha_{101} = \ldots$;
9. GGD_9: α_0 and α_1 are different, $\alpha_2 = \alpha_3 = \ldots = \alpha_9$, $\alpha_{10} = \alpha_{11} = \ldots = \alpha_{99}$, $\alpha_{100} = \alpha_{101} = \ldots \alpha_{105}$, $\alpha_{106} = \alpha_{107} = \ldots$;
10. GGD_∞: all hazard rates α_i are distinct.

Note that in the above when we mention that two or more hazard rates are distinct, it implied that they are *possibly* distinct and hence as a special case, some of them could be equal. Thus we have a nesting among the above probability models, in the sense, e.g. GGD_1 is a special case of all other models and every model in the above class is a special case of GGD_∞ We show the complete nesting across the 10 GGD model through Figure 1; model numbers are indicated within respective boxes; if there is a (chain of) directional arrow(s) from Model i to Model j, it should be interpreted as Model i is nested into Model j, or equivalently Model i is a subset model of Model j. We discuss choosing between two nested models in the testing of hypothesis framework later on.

Figure 1:


GGD1= Geometric
GGD2
GGD3
GGD4
GGD7
GGD5
GGD8
GGD6
GGD9
GGD∞= KM

3.2 Parameter Estimation in GGD Family

Let us now dwell on estimation of (hazard) parameters. We adopt maximum likelihood (ML) or restricted maximum likelihood (RML) approach, with the restriction coming in the form of models as mentioned above. To start with, let us consider the most general GGD distribution, viz. GGD_∞.

Adopting notations for data as given in the beginning of Section 2, the likelihood function is given by:

$$L = \Big[\prod_{i=1}^{n} \phi(X_i)\Big] \times \Big[\prod_{i=1}^{m} S(Y_i-)\Big]. \tag{15}$$

Since we are dealing with discrete data, it is very important to note and the last term in (15), as for any notout score y, its contribution to the likelihood expression should be $P(X \geq y) = \lim_{x\to y-} P(X > x) = S(y-)$, which is also equal to $S(y-1)$ for $y \geq 1$ and equal to 1 for $y = 0$. Now using (13) and (14) and rearranging terms in (15), we get:

$$\begin{aligned} L &= \Big[\prod_{i=0}^{K} \alpha_i^{f_i} \prod_{j=0}^{i-1} (1-\alpha_j)^{f_j}\Big] \times \Big[\prod_{i=0}^{K}\prod_{j=0}^{i-1} (1-\alpha_j)^{f_j^*}\Big] \\ &= \prod_{i=0}^{K} \alpha_i^{f_i} (1-\alpha_i)^{(n-F_i)+(m-F_i^*)} \end{aligned}$$

Taking logarithm, we get

$$\ln L = \sum_{i=0}^{K} f_i \ln \alpha_i + [(n-F_i) + (m-F_i^*)] \ln(1-\alpha_i);$$

Now differentiating w.r.t. α_i's and equating the expressions to zero, we get the MLE's :

$$\begin{aligned} \Big(\frac{f_i}{\alpha_i} - \frac{(n-F_i)+(m-F_i^*)}{1-\alpha_i}\Big)_{\alpha_i=\hat{\alpha_i}} &= 0 \\ \Rightarrow \hat{\alpha_i} &= \frac{f_i}{f_i + (n-F_i) + (m-F_i^*)} \end{aligned} \tag{16}$$

To understand and interpret (16), let us first consider the simplified scenario, when there are no notout innings, i.e. $f_i^* = \forall i, \Leftrightarrow F_i^* = \forall i$ or $m = 0$. In that case, (16) reduces to

$$\hat{\alpha_i} = \frac{f_i}{f_i + n - F_i}. \tag{17}$$

Note that the numerator of (17) is the number of times the batsman gets out at score i. On the

other hand, the denominator is the number of occasions the batsman has risk/exposure to get out at the score i, or equivalently this is the number of scores of i or higher. Thus, the MLE of the hazard rates are simply the reflections of relative frequency definition of probability, by virtue of complete lack of inter-connectivity across the hazard rates (memory-less property).

This same intuitive reasoning can be extended for data with notout innings, i.e. $m > 0$. Note now, in (16) vis-a-vis (17) the numerator unsurprisingly remains the same, while in the denominator we also count the no. of notout innings with score in *excess* of i, in addition. Intuitively, the innings with not out score equal to i are not counted in the numerator as well as the denominator, since in those innings the batsman could have fallen on the score i had the innings been allowed to continue.

We may also simplify (16) as

$$\hat{\alpha_i} = \frac{f_i}{f_i + N_i}, \tag{18}$$

by introducing the additional notation [refer to (2)]

$$N_i = M_{i+1} = (n - F_i) + (m - F_i^*)$$

representing the total no. of innings (out or notout, counting both) with scores in excess of i.

To obtain parameter estimates in the other GGD models we propose restricted maximum likelihood or RML. This is one of the (relatively rare) setup which entails closed form expressions for RML. While we will not dwell on the completely generalized expressions, it should be clear, following the path taken in derivation (16), if $\alpha_{i_1} = \alpha_{i_2} = \ldots = \alpha_{i_p}$ under the model, the ML of their common value would be given by:

$$\frac{\sum_{j=1}^{p} f_{i_j}}{\sum_{j=1}^{p} (f_{i_j} + N_{i_j})}. \tag{19}$$

Hence, in particular, the common hazard parameter in $\aleph_1$ would be given by

$$\hat{\alpha} = \frac{\sum_{i=0}^{K} f_i}{\sum_{i=0}^{K} \left(f_i + (n - F_i) + (m - F_i^*)\right)} = \frac{n}{n + \sum_{i=0}^{K} N_i},$$

Similarly, the RML for each of the remaining models can be obtained as given below:

$$GGD_2: \qquad \hat{\alpha_0} = \frac{f_0}{f_0 + N_0}, \quad \text{and for } i \geq 1, \hat{\alpha_i} = \frac{\sum_{i=1}^{K} f_i}{\sum_{i=1}^{K} (f_i + N_i)}$$

$$GGD_3: \qquad \hat{\alpha_0} = \frac{f_0}{f_0 + N_0}, \quad \hat{\alpha_1} = \frac{f_1}{f_1 + N_1}, \quad \text{and for } i > 1, \hat{\alpha_i} = \frac{\sum_{i=2}^{K} f_i}{\sum_{i=2}^{K} (f_i + N_i)}$$

$$GGD_4: \qquad \hat{\alpha_0} = \frac{f_0}{f_0 + N_0}; \quad \hat{\alpha_i} = \frac{\sum_{i=1}^{9} f_i}{\sum_{i=1}^{9} (f_i + N_i)}, \; i = 1, 2, \ldots, 9;$$
$$\text{and for } i > 9, \hat{\alpha_i} = \frac{\sum_{i=10}^{K} f_i}{\sum_{i=10}^{K} (f_i + N_i)}$$

$$GGD_5: \qquad \hat{\alpha_0} = \frac{f_0}{f_0 + N_0}; \quad \hat{\alpha_i} = \frac{\sum_{i=1}^{9} f_i}{\sum_{i=1}^{9} (f_i + N_i)}, \; i = 1, 2, \ldots, 9;$$
$$\hat{\alpha_i} = \frac{\sum_{i=10}^{99} f_i}{\sum_{i=10}^{99} (f_i + N_i)}, \; i = 10, \ldots, 99; \quad \&\text{for } i > 99, \hat{\alpha_i} = \frac{\sum_{i=100}^{K} f_i}{\sum_{i=100}^{K} (f_i + N_i)}$$

$$\text{GGD}_6: \quad \hat{\alpha_0} = \frac{f_0}{f_0 + N_0}; \quad \hat{\alpha_i} = \frac{\sum_{i=1}^{9} f_i}{\sum_{i=1}^{9} (f_i + N_i)},\ i = 1, 2, \ldots, 9;$$

$$\hat{\alpha_i} = \frac{\sum_{i=10}^{99} f_i}{\sum_{i=10}^{99} (f_i + N_i)},\ i = 10, \ldots, 99; \quad \hat{\alpha_i} = \frac{\sum_{i=100}^{105} f_i}{\sum_{i=100}^{105} (f_i + N_i)},\ i = 100, \ldots, 105;$$

$$\text{\& for } i > 105, \hat{\alpha_i} = \frac{\sum_{i=106}^{K} f_i}{\sum_{i=106}^{K} (f_i + N_i)}$$

$$\text{GGD}_7: \quad \hat{\alpha_0} = \frac{f_0}{f_0 + N_0}; \quad \hat{\alpha_1} = \frac{f_1}{f_1 + N_1}, \quad \hat{\alpha_i} = \frac{\sum_{i=2}^{9} f_i}{\sum_{i=2}^{9} (f_i + N_i)},\ i = 2, \ldots, 9;$$

$$\text{and for } i > 9, \hat{\alpha_i} = \frac{\sum_{i=10}^{K} f_i}{\sum_{i=10}^{K} (f_i + N_i)}$$

$$\text{GGD}_8: \quad \hat{\alpha_0} = \frac{f_0}{f_0 + N_0}; \quad \hat{\alpha_1} = \frac{f_1}{f_1 + N_1}, \quad \hat{\alpha_i} = \frac{\sum_{i=2}^{9} f_i}{\sum_{i=2}^{9} (f_i + N_i)},\ i = 2, \ldots, 9;$$

$$\hat{\alpha_i} = \frac{\sum_{i=10}^{99} f_i}{\sum_{i=10}^{99} (f_i + N_i)},\ i = 10, \ldots, 99; \quad \text{\&for } i > 99, \hat{\alpha_i} = \frac{\sum_{i=100}^{K} f_i}{\sum_{i=100}^{K} (f_i + N_i)}$$

$$\text{GGD}_9: \quad \hat{\alpha_0} = \frac{f_0}{f_0 + N_0}; \quad \hat{\alpha_1} = \frac{f_1}{f_1 + N_1}, \quad \hat{\alpha_i} = \frac{\sum_{i=2}^{9} f_i}{\sum_{i=2}^{9} (f_i + N_i)}, \; i = 2, \ldots, 9;$$

$$\hat{\alpha_i} = \frac{\sum_{i=10}^{99} f_i}{\sum_{i=10}^{99} (f_i + N_i)}, \; i = 10, \ldots, 99; \quad \hat{\alpha_i} = \frac{\sum_{i=100}^{105} f_i}{\sum_{i=100}^{105} (f_i + N_i)}, \; i = 100, \ldots, 105;$$

$$\& \text{ for } i > 105, \hat{\alpha_i} = \frac{\sum_{i=106}^{K} f_i}{\sum_{i=106}^{K} (f_i + N_i)}$$

Recall that, in the given context, estimating mean is of paramount important and the GGD does not have simplified expression for the population; the mean would be simply estimated by

$$M_{\text{GGD}} = \sum_{i=1}^{K} i \; \hat{\alpha_i} \times \Big(\prod_{j=0}^{i-1} (1 - \hat{\alpha_j}) \Big) \tag{20}$$

3.3 Using Likelihood ratio Test to Select between Models

We use the likelihood ratio test to choose between any pair of nested models. Recall that the asymptotic null distribution of the test statistic

$$2 \times [\ln(L_f) - \ln(L_s)]$$

is Chi-square with degrees of freedom being equal to the difference in the number of parameters between the full and the subset model; L_f and L_s are respectively likelihood under full and subset models.

4 Proposed method of Estimating True Average Score

Having decided on the best GGD model as the fit for the runs scores, we consider and propose two alternative procedures depending on the context. (There are also some additional variations or considerations while implementing in a given context, which we discuss in more details within the description of the methods).

1. Method 1: Mean of GGD (abbreviated as MGGD)
2. Method 2: Replacement of Notout score Before Averaging method (abbreviated as RNBA)

Mean of GGD (MGGD): Having chosen and fitted one (best) GGD model, the simplest and arguably the most natural approach would be to used the mean of GGD model as the estimate of the true average score. Indeed when the objective is to estimate the average over a prolonged period, like the career of a batsman, we recommend this MGGD or mean of the GGD apparoach. There are two broad steps in model selection and fitting, viz.,

- Estimation of hazard parameters in specific GGD models: estimator is to be obtained using all the available scores, from completed as well as notout innings, using (restricted) maximum likelihood procedure outlined in Section 3.2.
- Selection of best fit model from the group GGD models: this would be done using likelihood ratio statistic, also outlined in Section 3.3.

While there are some scopes for the analyzer's discretion in adopting the specific form, the following is our recommendation.

- For long term like career average for a specific player, find the best fit GGD model as well as all relevant hazard parameters on the basis of the entire data of that player.
- For a short time batting average, like over a test match series or a tournament, one may choose the best GGD model using all the historical data corresponding to batting position in the same form of the game (test or 1-day or 20-20). At times, using broader structure like opening batsmen, middle order, tailender etc maybe used. Alternatively, especially if the player has been playing for long enough, the best GGD model may be selected on the basis of the specific players's career-wide data. In either case, once the best GGD model is identified, the hazard parameters are estimated only on the basis of the relevant scores of the player during that series or tournament (time span).

Replacement of Notout score Before Averaging method (RNBA): In this method, every notout score y^* gets substituted by the projected (replacement score) score in that innings computed by

$$x^*(y^*) = E[X|X \geq y^*],$$

where the expectation is taken in terms of the fitted (chosen GGD) distribution of score variable X. Subsequently, the final estimate is obtained by simply taking the average of scores from completed (out) innings and the replacement scores corresponding to the notout innings.

Couple of variations within the RNBA scheme seems reasonable. When large data is available and pertinent, for example while computing career average, it makes sense to estimate the hazard parameters of based on entire career data. There may be some concern of 'inbreeding' and hence one can also adopt 'leave current observation out' while fitting. However the issue is of much less relevance in the given context.

While estimating average over relatively short period, e.g. in a series (when the data will be scarce), one can select an appropriate GGD model using prior information. We recommend using batting position as the key prior information, although alternatives are possible along the same approach. So a specific member of GGD family with already pre-selected hazard parameters based on prior information (batting position).

4.1 Numerical demonstration: GGD models in Estimating Batting Average score by the Two Proposed Methods — Career Average for Top Batsmen in Test and 1 Day games

We now show the GGD averages by different model for all major test batsmen, the same set reported in Table 2.

Comments from Tables 3 - 6

- In Tables 5 and 6, Avg_avg refers to the average of all averages computed using both the methods, mean of the fitted GGD, as well as the mean obtained by substituting the notout scores by GGD modeling. Average of the first group of methods, excluding the two extremes (traditional average and K-M) are denoted by the Mean_MGGD, while the same from the second method are denoted by Mean_RNBA.

- GGD_1 corresponding to the simple Geometric model and the average reduces to traditional average as shown lemma 1, under both the method proposed.

- Similarly, GGD_∞ corresponds to the K-M average under both methods.

- $MGGD_i$ refers to the MGGD estimate after fitting Model GGD_i, while $RNBA_i$ refers to the RNBA estimate from the same model.

Table 3: Batting average of Top Test Batsmen: MGGD estimates (Method 1) — Estimating by mean of fitted GGD Models

	GGD_1	GGD_2	GGD_3	GGD_4	GGD_5	GGD_6	GGD_7	GGD_8	GGD_9	GGD_∞
DG Bradman	99.94	99.88	100.02	100.02	99.99	99.95	100.03	100.00	99.96	98.98
RG Pollock	60.96	61.02	61.08	61.64	61.40	61.44	61.61	61.38	61.42	62.14
GA Headley	60.83	61.07	61.14	61.29	61.29	61.41	61.30	61.30	61.43	60.89
H Sutcliffe	60.73	60.78	60.68	60.67	59.47	59.52	60.68	59.48	59.53	59.50
E Paynter	59.22	60.26	60.50	61.12	62.73	63.00	61.12	62.73	63.01	58.56
KF Barrington	58.67	58.83	59.06	59.10	58.09	58.05	59.12	58.11	58.07	57.75
ED Weekes	58.61	58.85	59.01	59.07	58.54	58.54	59.14	58.62	58.61	58.56
WR Hammond	58.45	58.55	58.64	58.63	59.76	59.74	58.64	59.78	59.75	61.07
IJL Trott	57.79	57.86	57.74	58.48	59.99	59.70	58.34	59.84	59.55	58.23
GS Sobers	57.78	58.34	58.45	58.27	59.14	59.16	58.26	59.13	59.15	61.03
JB Hobbs	56.94	57.04	57.10	56.80	56.50	56.48	56.76	56.47	56.44	56.53
KC Sangakkara	56.93	56.96	56.96	57.01	57.82	57.86	57.01	57.82	57.86	57.43
JH Kallis	56.89	57.18	57.18	57.51	56.94	57.09	57.51	56.94	57.08	55.61
CL Walcott	56.68	56.66	56.72	56.80	56.87	56.78	56.80	56.87	56.78	56.67
L Hutton	56.67	56.81	56.74	56.82	57.43	57.99	56.81	57.42	57.98	57.39
SR Tendulkar	56.02	56.22	56.34	56.67	56.91	56.79	56.68	56.93	56.81	55.91
GS Chappell	53.86	54.39	54.37	54.57	54.44	54.27	54.56	54.43	54.26	54.01
AD Nourse	53.82	54.04	54.03	54.23	53.85	53.84	54.22	53.84	53.82	54.04
R Dravid	53.51	53.59	53.65	53.98	54.09	54.00	53.99	54.10	54.01	53.68
BC Lara	52.89	52.97	53.00	53.05	53.24	53.24	53.08	53.27	53.26	53.04
V Sehwag	52.69	52.82	52.85	52.87	52.90	52.89	52.87	52.90	52.89	53.17
TT Samaraweera	52.62	53.44	53.89	54.08	53.99	53.97	54.14	54.05	54.04	53.22
J Miandad	52.57	52.66	52.72	52.85	54.41	54.81	52.85	54.41	54.82	53.87
RT Ponting	52.54	52.80	52.81	53.10	52.95	53.01	53.09	52.95	53.01	52.86
M Yousuf	52.29	52.54	52.55	52.64	52.58	52.55	52.64	52.58	52.55	52.32
MEK Hussey	51.73	52.22	52.38	52.40	51.72	51.59	52.42	51.74	51.61	51.16
J Ryder	51.63	51.74	51.86	53.47	54.52	54.03	53.43	54.47	53.99	52.09
A Flower	51.55	51.84	51.74	52.67	56.50	55.63	52.65	56.48	55.61	53.28
DPMD Jayawardhane	51.31	51.43	51.52	51.60	51.74	51.79	51.64	51.78	51.83	51.90
Y Khan	51.21	51.55	51.59	52.00	52.09	51.89	52.00	52.08	51.89	52.19
SM Gavaskar	51.12	51.28	51.37	51.54	51.32	51.27	51.58	51.36	51.31	51.05
S Waugh	51.06	51.87	52.19	52.36	53.06	54.01	52.40	53.10	54.05	51.20
M Hayden	50.73	50.98	50.93	50.88	50.94	50.95	50.88	50.94	50.94	50.68
A Border	50.56	50.80	50.89	51.59	51.43	51.27	51.59	51.43	51.27	50.15
K Pietersen	50.48	50.58	50.70	50.71	50.56	50.55	50.76	50.61	50.60	50.51
IVA Richards	50.23	50.37	50.43	50.56	50.70	50.66	50.59	50.72	50.69	50.60
DCS Compton	50.06	50.47	50.52	50.40	50.31	50.26	50.41	50.31	50.27	50.48

Similarly in Tables 7 and 8 we present the MGGD and RNBA estimates of average of top batsmen in 1-day version of the game. The criterion of selection has been traditional average in excess of 38; in addition, we have included three well-known big-hitters of the game, viz. AC Gilchrist, V Sehwag and ST Jayasurya, whose average are slightly lower, but strike rate, which is an important indicator of 1-day batting, is high. (RN ten Doeschate is excluded from this list as he played mostly against associate countries.) As already noted and verified, the MGGD and RNBA estimates coincide for GGD_1 and GGD_∞. Hence, in Table 8, instead of reporting these duplicate entries, we have included the strike rate (defined as the average runs scored per 100 balls faced) and an overall batting index, as arguably apt for single measure in 1-day format, computed by multiplying the average score by the strike rate — in this context, we take the average of the GGD averages, both methods combined. The final comparison of ranking is reported in Table 9 by the

Table 4: Batting average of Top Test Batsmen: RNBA estimate (Method 2)–
Replacement of Notout score (via GGD Model fitting) Before Averaging method

	GGD_1	GGD_2	GGD_3	GGD_4	GGD_5	GGD_6	GGD_7	GGD_8	GGD_9	GGD_∞
Bradman	99.94	100.76	101.01	101.04	100.73	100.60	101.04	100.73	100.60	98.98
RG Pollock	60.97	61.03	61.09	61.93	61.62	61.63	61.93	61.62	61.63	62.14
GA Headley	60.83	61.07	61.15	61.31	61.54	61.69	61.31	61.55	61.69	60.89
H Sutcliffe	60.73	60.79	60.68	60.68	59.59	59.64	60.68	59.59	59.65	59.50
E Paynter	59.23	60.27	60.51	61.05	62.04	63.01	61.05	62.04	63.01	58.56
KF Barrington	58.67	58.83	59.07	59.13	58.45	58.42	59.13	58.45	58.42	57.75
ED Weekes	58.62	58.86	59.03	59.17	58.69	58.69	59.17	58.69	58.69	58.56
WR Hammond	58.46	58.55	58.65	58.64	61.12	61.07	58.64	61.12	61.08	61.07
IJL Trott	57.79	57.86	57.74	58.62	61.26	60.82	58.62	61.26	60.82	58.23
GS Sobers	57.78	58.34	58.45	58.27	59.48	59.50	58.27	59.48	59.50	61.03
JB Hobbs	56.95	57.04	57.11	56.79	56.59	56.55	56.79	56.59	56.56	56.53
KC Sangakkara	56.94	56.96	56.96	57.02	58.26	58.31	57.02	58.26	58.31	57.43
JH Kallis	56.89	57.18	57.18	57.52	56.82	56.97	57.52	56.82	56.97	55.61
CL Walcott	56.69	56.66	56.72	56.79	56.11	56.01	56.79	56.11	56.01	56.67
L Hutton	59.85	60.11	60.08	60.72	64.79	63.45	60.72	64.78	63.45	55.52
SR Tendulkar	56.02	56.22	56.35	56.66	56.93	56.80	56.66	56.93	56.81	55.91
GS Chappell	53.86	54.39	54.37	54.59	54.39	54.21	54.59	54.38	54.20	54.01
AD Nourse	53.82	54.04	54.03	54.29	53.91	53.90	54.29	53.91	53.90	54.04
R Dravid	53.52	53.59	53.65	54.03	54.13	54.04	54.03	54.13	54.04	53.68
BC Lara	52.89	52.97	53.00	53.10	53.29	53.28	53.10	53.29	53.28	53.04
V Sehwag	52.69	52.83	52.85	52.87	53.12	53.10	52.87	53.12	53.10	53.17
TT Samaraweera	52.62	53.44	53.89	54.13	54.04	54.03	54.15	54.06	54.05	53.22
J Miandad	52.57	52.66	52.72	52.85	54.61	55.08	52.86	54.61	55.08	53.87
RT Ponting	52.54	52.80	52.81	53.08	52.90	52.96	53.08	52.90	52.96	52.86
M Yousuf	52.29	52.54	52.55	52.64	52.53	52.51	52.64	52.53	52.51	52.32
MEK Hussey	51.73	52.22	52.38	52.43	51.86	51.74	52.43	51.86	51.74	51.16
J Ryder	51.63	51.74	51.86	53.64	54.47	54.03	53.64	54.47	54.03	52.09
A Flower	51.55	51.84	51.74	52.60	56.38	55.48	52.60	56.38	55.48	53.28
DPMD Jayawardhane	51.31	51.44	51.52	51.67	51.90	51.94	51.67	51.90	51.94	51.90
Y Khan	51.21	51.55	51.59	52.00	52.08	51.88	52.00	52.08	51.88	52.19
SM Gavaskar	51.12	51.28	51.37	51.62	51.34	51.29	51.62	51.34	51.30	51.05
S Waugh	51.06	51.87	52.19	52.38	53.19	54.24	52.40	53.21	54.26	51.20
M Hayden	50.74	50.98	50.93	50.89	50.62	50.62	50.88	50.61	50.62	50.68
A Border	50.56	50.80	50.89	51.55	51.40	51.25	51.55	51.40	51.25	50.15
K Pietersen	50.48	50.58	50.70	50.77	50.58	50.64	50.77	50.58	50.64	50.51
IVA Richards	50.23	50.37	50.43	50.64	50.75	50.72	50.64	50.75	50.72	50.60
DCS Compton	50.06	50.47	50.52	50.40	50.25	50.20	50.40	50.25	50.21	50.48

different methods.

Comments from Tables 7- 9

- Notations are identical to the ones used in Tables 3–6.
- For several players, most notably MG Bevan and MEK Hussey, the traditional batting average appears to estimate their true average by substantive margin.
- It is interesting to note that traditional estimate appears to underestimate in few cases as well. Going by the assumption that Avg_avg is more reflective of the true average, that is the case for as many as 8 players in considered list.
- We also performed the rank correlation between different methods of estimating the average

Table 5: Rank order of test batsmen by different methods of computing averages

	Rank						
Player	Trad. Avg	K-M	$MGGD_2$	$RNBA_2$	Mean_MGGD	Mean_RNBA	Avg_avg
DG Bradman	1	1	1	1	1	1	1
RG Pollock	2	2	3	3	3	3	2
GA Headley	3	5	2	2	4	4	4
H Sutcliffe	4	6	4	4	5	5	5
E Paynter	5	8	5	5	2	2	3
KF Barrington	6	10	7	8	10	10	11
ED Weekes	7	7	6	7	8	8	10
WR Hammond	8	3	8	9	6	6	6
IJL Trott	9	9	10	11	7	7	8
GS Sobers	10	4	9	10	9	9	9
JB Hobbs	11	14	12	13	15	15	14
KC Sangakkara	12	11	13	14	11	11	12
JH Kallis	13	16	11	12	13	13	13
CL Walcott	14	13	15	15	14	14	15
L Hutton	15	12	14	6	12	12	7
SR Tendulkar	16	15	16	16	16	16	16
GS Chappell	17	18	17	17	17	17	17
AD Nourse	18	17	18	18	19	19	18
R Dravid	19	20	19	19	21	21	20
BC Lara	20	23	21	21	24	24	24
TT Samaraweera	21	22	20	20	20	20	21
J Miandad	22	19	23	23	22	22	22
RT Ponting	23	24	22	22	25	25	25
M Yousuf	24	26	24	24	27	27	27
V Sehwag	25	25	25	25	28	28	28
MEK Hussey	26	31	26	26	29	29	29
J Ryder	27	28	29	29	23	23	23
A Flower	28	21	28	28	18	18	19
DPMD Jayawardhane	29	29	31	31	31	31	30
Y Khan	30	27	30	30	30	30	26
SM Gavaskar	31	32	32	32	32	32	31
S Waugh	32	30	27	27	26	26	32
M Hayden	33	33	33	33	34	34	33
A Border	34	37	34	34	33	33	34
K Pietersen	35	35	35	35	35	35	35
IVA Richards	36	34	37	37	36	36	36
DCS Compton	37	36	36	36	37	37	37

Table 6: Rank Correlation between different Averages: test batsmen

	Trad. Avg	K-M	$MGGD_2$	$RNBA_2$	Mean_MGGD	Mean_RNBA	*Avg_avg*
Trad. Avg	1.00						
K-M	0.97	1.00					
$MGGD_2$	0.99	0.97	1.00				
$Mean_2$	0.98	0.96	0.99	1.00			
Mean_MGGD	0.97	0.98	0.97	0.97	1.00		
Mean_RNBA	0.97	0.98	0.97	0.97	1.00	1.00	
Avg_avg	0.97	0.98	0.96	0.97	0.99	0.99	1.00

score, although this is not reported here. The most interesting part of that analysis shows that while the strike rate is negatively correlated with all the estimated averages, it is more so (-0.13) with the traditional average than with MGGD and RNBA estimates (all approximately -0.05). The rank correlation of the batting index with all the GGD models are approximately same — it is worst with the traditional average.

Table 7: Batting average of Top Batsmen in 1-day cricket: MGGD estimates (Method 1) — Estimating by mean of fitted GGD Models

	Trad. Avg	GGD fitted Mean								Kaplan-Meir
	GGD_1	GGD_2	GGD_3	GGD_4	GGD_5	GGD_6	GGD_7	GGD_8	GGD_9	GGD_∞
HM Amla	55.17	55.18	55.20	55.14	52.88	52.68	55.14	52.88	52.68	53.09
MG Bevan	53.58	53.79	53.73	52.44	47.08	46.69	52.44	47.08	46.68	46.13
IJL Trott	51.37	51.57	51.46	51.86	47.70	47.70	51.81	47.66	47.66	49.19
MEK Hussey	51.18	51.24	51.15	51.70	45.99	45.59	51.70	45.98	45.59	44.93
MS Dhoni	51.16	51.47	51.20	51.60	53.77	52.86	51.58	53.75	52.84	51.02
Z Abbas	47.63	47.70	47.78	47.50	46.32	46.22	47.50	46.32	46.22	45.83
IVA Richards	47.00	47.18	47.22	47.39	49.85	49.07	47.40	49.86	49.07	46.95
V Kohli	46.78	47.27	47.25	47.56	44.06	43.77	47.55	44.06	43.76	43.39
MJ Clarke	45.50	45.96	46.41	46.28	43.49	43.75	46.32	43.52	43.78	43.16
J Kallis	45.49	45.81	45.96	46.13	43.94	44.10	46.14	43.95	44.11	43.40
SR Tendulkar	45.16	45.28	45.42	45.64	44.78	44.92	45.69	44.83	44.97	44.64
AB de Villiers	44.28	44.43	44.30	44.32	43.10	42.97	44.32	43.10	42.97	42.91
Ml Hayden	43.81	43.99	43.99	43.67	43.12	43.13	43.66	43.11	43.12	43.02
SR Watson	43.05	43.56	43.33	43.66	53.13	52.81	43.64	53.09	52.78	44.84
RT Ponting	42.64	42.82	42.91	42.69	41.89	41.91	42.68	41.88	41.91	41.68
G Gambhir	41.86	42.10	42.09	42.06	42.29	43.62	42.06	42.29	43.62	40.92
M Yousuf	41.58	41.82	41.92	41.95	40.59	40.46	41.96	40.59	40.47	40.34
L Klusener	41.10	43.13	42.75	41.91	38.39	38.28	41.91	38.39	38.28	39.05
SC Ganguly	41.02	41.13	41.16	41.39	41.30	41.13	41.39	41.30	41.14	40.59
BC Lara	40.49	40.66	40.61	40.83	40.17	40.09	40.82	40.16	40.08	39.99
A Symonds	39.75	40.57	40.63	40.96	43.93	43.03	40.96	43.93	43.03	40.09
GC Smith	39.43	39.49	39.54	39.45	39.13	39.70	39.44	39.11	39.69	38.57
R Dravid	39.22	39.31	39.33	39.20	38.40	38.54	39.20	38.40	38.53	38.22
Ch Gayle	39.07	39.29	39.33	39.60	39.17	39.24	39.60	39.18	39.24	38.35
KP Pietersen	38.62	38.80	38.83	38.87	37.04	37.14	38.87	37.04	37.14	36.81
KC Sangakkara	38.13	38.18	38.29	38.06	37.12	37.10	38.03	37.09	37.07	37.18
AC Gilchrist	35.89	35.96	35.97	35.91	35.76	35.77	35.91	35.75	35.76	35.74
V Sehwag	35.11	35.15	35.17	35.17	35.00	34.96	35.17	35.00	34.96	34.95
ST Jayasuriya	32.68	32.77	32.80	32.85	32.78	32.73	32.87	32.80	32.75	32.66

4.1.1 Bradman's Batting Average: Did it actually cross 100?

Finally we turn our attention to applying the method on individual players. To illustrate we take the case of Don Bradman, not only because he was all time great, but also because of interest regarding his batting average. His traditional (test match) batting average was 99.94 (would have been 100 had he score 4 instead of 0 in his last inning); his Kaplan-Meir average would be 99.06 (incorrectly reported as 98.98 in [4]). Tables 7 and 8 summarize the comparison of fit across the 10 GGD models.

From these tables it is clear that the GGD_2 is the best fit for Bradman's scores. In addition we also considered other reasonable models in the GGD class like

$$\aleph: \quad \alpha_0 \text{ is different}, \alpha_1 = \alpha_2 = \ldots = \alpha_{80}, \quad \alpha_{81} = \alpha_{82} = \ldots;$$

but none of them seem to fit any better.

Table 8: Batting average of Top Batsmen in 1-day cricket: RNBA estimate (Method 2)– Replacement of Notout scores (via GGD Model fitting) Before Averaging method

	GGD substitution method									
	GGD_2	GGD_3	GGD_4	GGD_5	GGD_6	GGD_7	GGD_8	GGD_9	strike rate	index
HM Amla	55.18	55.20	55.14	53.39	53.20	55.14	53.39	53.21	91.98	49.72
MG Bevan	53.79	53.73	52.43	47.52	47.15	52.43	47.52	47.15	74.16	36.98
IJL Trott	51.58	51.46	51.89	49.57	49.57	51.89	49.57	49.57	78.51	39.36
MEK Hussey	51.18	51.24	51.15	51.69	46.98	46.66	51.68	46.97	87.76	42.94
MS Dhoni	51.47	51.20	51.60	53.64	52.77	51.59	53.63	52.77	88.52	46.17
Z Abbas	47.70	47.78	47.44	45.66	45.58	47.45	45.67	45.59	84.80	39.61
IVA Richards	47.18	47.22	47.38	49.47	48.77	47.39	49.47	48.77	90.20	43.37
V Kohli	47.27	47.25	47.54	43.63	43.35	47.54	43.63	43.35	83.02	37.68
MJ Clarke	45.96	46.41	46.29	43.95	44.17	46.32	43.97	44.20	78.10	35.05
J Kallis	45.81	45.96	46.13	44.16	44.31	46.14	44.16	44.31	72.88	32.74
SR Tendulkar	45.28	45.42	45.73	44.90	45.05	45.73	44.91	45.05	86.32	38.98
AB de Villiers	44.43	44.30	44.32	42.95	42.81	44.32	42.95	42.81	91.29	39.80
Ml Hayden	43.99	43.99	43.65	43.24	43.25	43.65	43.24	43.25	78.96	34.32
SR Watson	43.56	43.33	43.65	53.65	52.73	43.64	53.63	52.72	89.84	42.84
RT Ponting	42.82	42.91	42.68	41.87	41.90	42.69	41.88	41.90	80.60	34.08
G Gambhir	42.10	42.09	42.05	42.31	43.75	42.05	42.31	43.75	86.58	36.65
M Yousuf	41.82	41.92	41.96	40.57	40.45	41.96	40.57	40.45	75.10	30.90
L Klusener	43.13	42.75	41.92	38.13	38.02	41.91	38.12	38.01	89.91	36.16
SC Ganguly	41.13	41.16	41.37	41.30	41.16	41.37	41.30	41.16	73.70	30.34
BC Lara	40.66	40.61	40.83	40.23	40.16	40.83	40.23	40.16	79.51	32.12
A Symonds	40.57	40.63	40.95	43.33	42.58	40.95	43.33	42.58	92.44	38.53
GC Smith	39.49	39.54	39.45	39.26	39.65	39.45	39.26	39.65	81.72	32.17
R Dravid	39.31	39.33	39.19	38.58	38.68	39.20	38.58	38.69	71.24	27.68
Ch Gayle	39.29	39.33	39.60	39.20	39.26	39.60	39.20	39.26	83.95	32.93
KP Pietersen	38.80	38.83	38.88	36.97	37.07	38.88	36.97	37.07	86.94	32.92
KC Sangakkara	38.18	38.29	38.05	37.48	37.46	38.05	37.49	37.47	75.19	28.33
AC Gilchrist	35.96	35.97	35.90	35.77	35.78	35.90	35.77	35.78	96.94	34.75
V Sehwag	35.15	35.17	35.17	34.97	34.92	35.17	34.97	34.92	103.94	36.43
ST Jayasuriya	32.77	32.80	32.87	32.73	32.67	32.87	32.73	32.68	91.21	29.88

While computing the mean of fitted distributions, a critical issue is where to truncate the distribution. If the density is truncated at the maximum score, like 334 for Bradman, the total probability falls substantially short of 1 for some of the GGD models. This is especially the case with Bradman, as his hazard rate was significantly lower than everybody else the game (Bradman's hazard was $< 1\%$, while the rest of the great players have between 1.6% to 1.8% — another yardstick which shows conclusively the greatness of Bradman and how much of an outlier his performance truely was. That being the case, while for most players, one can truncate the upper end of the score distribution at say 500 for the sake of computing the mean, without affecting the average computation to any significant degree, in the case Bradman, one should truncate only at 700 or higher. A possible extension of this work is to estimate the player's expected score on the basis of chosen model and the number of innings played by him and use that for truncation. This is planned to be taken up in a followup work.

Table 9: Rank order of top batsmen in 1 day cricket by different methods of computing averages

Player	Trad. Avg	K-M	FM_Model 6	SB_Model 6	Fitted Mean	GGD sub	Avg_avg	strike rate	Index
HM Amla	1	1	3	1	1	1	1	4	1
MG Bevan	2	5	6	6	3	4	4	26	12
IJL Trott	3	3	5	4	4	3	3	22	8
MEK Hussey	4	7	8	7	5	5	5	11	4
MS Dhoni	5	2	1	2	2	2	2	10	2
Z Abbas	6	6	7	8	8	8	8	15	7
IVA Richards	7	4	4	5	6	7	6	7	3
V Kohli	8	11	11	13	9	9	9	17	11
MJ Clarke	9	12	12	11	12	11	12	23	16
J Kallis	10	10	10	10	11	12	11	28	22
SR Tendulkar	11	9	9	9	10	10	10	14	9
AB de Villiers	12	14	16	15	13	13	13	5	6
Ml Hayden	13	13	14	14	14	14	14	21	18
SR Watson	14	8	2	3	7	6	7	9	5
RT Ponting	15	15	17	17	16	16	16	19	19
G Gambhir	16	16	13	12	15	15	15	13	13
M Yousuf	17	18	19	19	19	19	18	25	25
L Klusener	18	21	24	24	21	21	21	8	15
SC Ganguly	19	17	18	18	18	18	19	27	26
BC Lara	20	20	20	20	20	20	20	20	24
A Symonds	21	19	15	16	17	17	17	3	10
GC Smith	22	22	21	21	22	22	22	18	23
R Dravid	23	24	23	23	24	24	24	29	29
Ch Gayle	24	23	22	22	23	23	23	16	21
KP Pietersen	25	26	25	26	25	25	25	12	20
KC Sangakkara	26	25	26	25	26	26	26	24	28
AC Gilchrist	27	27	27	27	27	27	27	2	17
V Sehwag	28	28	28	28	28	28	28	1	14
ST Jayasuriya	29	29	29	29	29	29	29	6	27

Table 10a: P-values in Comparison of fit across GGD models: Bradman's score

	GGD_2	GGD_4	GGD_5	GGD_6	GGD_∞
GGD_1	0.0000	0.0001	0.0002	0.0006	0.0934
GGD_2		0.5586	0.8202	0.8954	0.3263
GGD_4			0.9733	0.8770	0.0000
GGD_5				0.9011	0.0000
GGD_6					0.0000

Table 10b: P-values in Comparison of fit across GGD models: Bradman's score

	GGD_2	GGD_3	GGD_7	GGD_8	GGD_9	GGD_∞
GGD_1	0.0000	0.0000	0.0001	0.0003	0.0007	0.0934
GGD_2		0.1778	0.4026	0.5990	0.7207	0.3263
GGD_3			0.9513	0.9714	0.9662	0.3406
GGD_7				0.8159	0.8770	0.3223
GGD_8					0.6481	0.3052
GGD_9						0.2911

4.1.2 Bevan's Average: a case of major adjustment

As seen in Table 7, the traditional average grossly appear to overestimate the true average for few players, most notably MG Bevan and MEK Hussey. We look at the case of Bevan more closely. For Bevan's 1 day career batting average, we fit the 10 GGD models - enclosed are the p-values associated with the comparison.

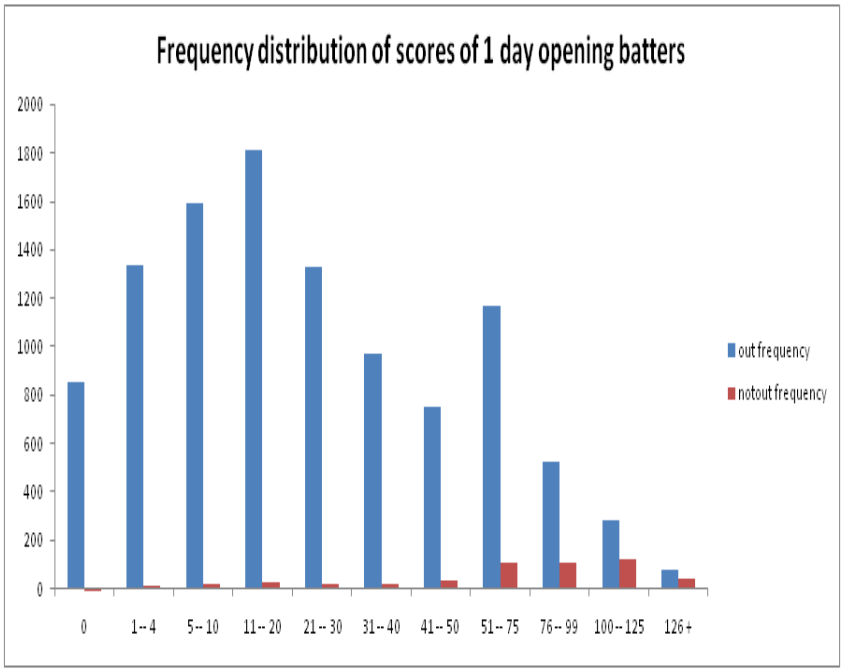
Table 11: P-values in Comparison of fit across GGD models: Bevan's 1 day score

	GGD_2	GGD_3	GGD_4	GGD_5	GGD_6	GGD_7	GGD_8	GGD_9	GGD_∞
GGD_1	0.4714	0.7498	0.1736	0.0333	0.0108	0.3066	0.0655	0.0215	0.9999
GGD_2		0.8109	0.0841	0.0166	0.0056	0.2130	0.0400	0.0129	0.9999
GGD_3			NA	NA	NA	0.0814	0.0161	0.0055	0.9999
GGD_4				0.0042	0.0019	0.7409	0.0697	0.0213	0.9999
GGD_5					0.0004	NA	0.7409	0.1061	0.9999
GGD_6						NA	NA	0.7409	1.0000
GGD_7							0.0000	0.0000	0.0004
GGD_8								0.1136	0.0368
GGD_9									1.0000

Thus, GGD Model 6 appears to be the best fit; in this case with the estimated mean being 46.686 which is fairly close to the K-M estimator, but more appealing than the later for reasons described earlier.

4.1.3 One Model does not fit all

Figure 2:



Figures 3 and 4 show the 10 fitted GGD models for 1 day openers scores. Tables 12a and 12b give summary of fits for the 10 GGD models.

Let us first report results of fitting 10 GGD distribution for all 1-day openers (batsmen playing at position 1 and 2). The data consists of all 11238 scores of the openers till April 6, 2009 in 1

Figure 3:

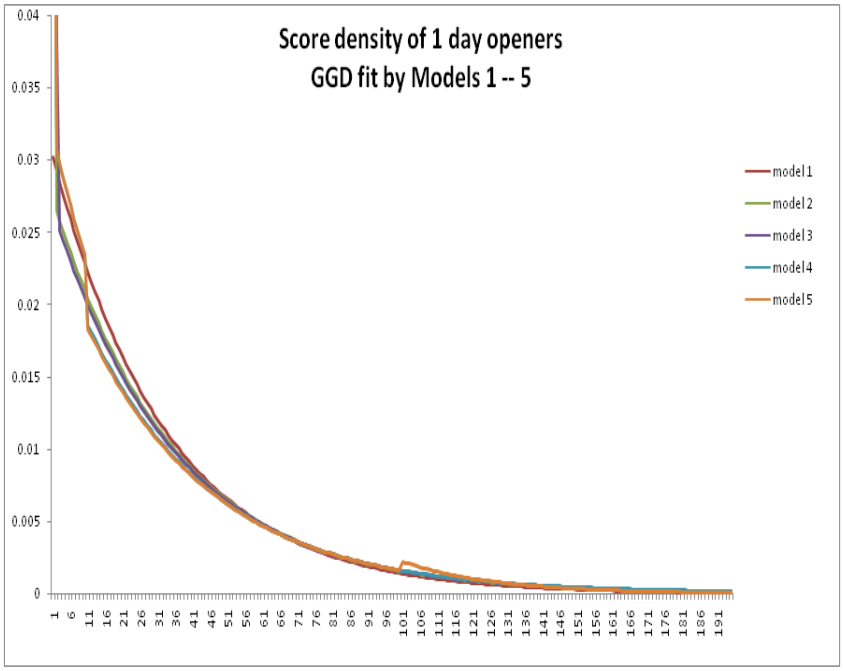
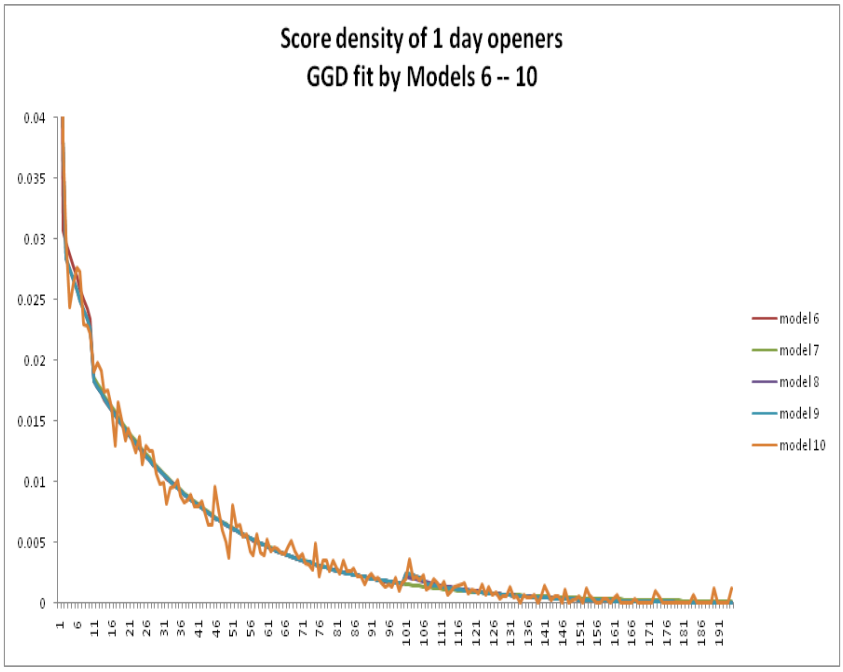


Figure 4:



-day matches; out of these 543 were notout innings and the remaining 10695 were scores from completed innings. The figure below shows the frequency distribution of these scores, suitably clubbed for representing here; however only raw or un-grouped was carried out for our numerical analysis. The traditional average of these scores is 32.114 (while average of scores from completed innings is only 28.325).

Table 12a: Summary of fit for 10 GGD models: score of 1-day openers

GGD Model No.	1	2	3	4	5
log likelihood	-47963.95	-47661.60	-47635.12	-47622.15	-47603.55
No of parameters	1	2	3	3	4
estimated mean	31.54	31.45	31.42	31.33	31.68
Model No.	6	7	8	9	10
log likelihood	-47602.30	-47611.42	-47592.83	-47591.58	-47479.67
No of parameters	5	4	5	6	195
estimated mean	31.65	31.34	31.69	31.66	31.99

Table 12b: P-values: GGD Model selection of 1-day openers

	GGD_2	GGD_3	GGD_4	GGD_5	GGD_6	GGD_7	GGD_8	GGD_9	GGD_∞
GGD_1	1.6E-133	1.6E-143	3.6E-149	6.5E-156	3.2E-155	1.7E-152	2.5E-159	1E-158	3.6E-103
GGD_2		3.4E-13	6.53E-19	6.2E-26	1.55E-25	1.61E-22	1.28E-29	2.76E-29	1.33E-12
GGD_3			NA	NA	NA	5.81E-12	4.3E-19	9.27E-19	1.19E-07
GGD_4				4.55E-27	1.77E-26	3.62E-06	1.85E-13	3.35E-13	9.61E-13
GGD_5					5.45E-16	NA	3.62E-06	6.27E-06	9.28E-08
GGD_6						NA	NA	3.62E-06	0.000332
GGD_7							1.07E-09	2.41E-09	0.000396
GGD_8								0.113631	0.036797
GGD_9									0.042319

Given the large number of data points, we select 1% level of significance and consequently recommend GGD_8 for modeling scores for 1 day openers, by noting that more refined models (9 and 10) are not significantly better than this one, while it is superior to all models which are nested into this.

To draw a comparison, we now carry out the analysis for batsmen in No 11 position (last) at the 1-day games. Here the number of observations is much less (only 2145) with a large number of notout innings (1277). The highest score is 43, but only 17 scores are 22 or above. Figure 5 shows the frequency distribution.

Because of the smaller range for scores, only 6 of the 10 models are pertinent here; Table 13 gives summary of fits for these 6 GGD models.

Figure 6 shows the 6 fitted GGD models for 1 day scores of No. 11 batsmen.

Figure 5:

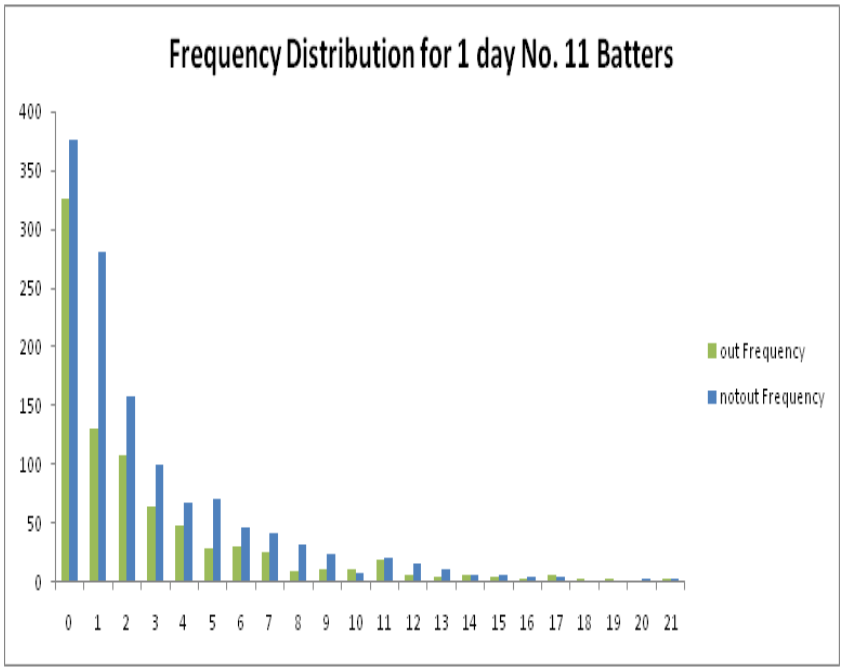


Figure 6:

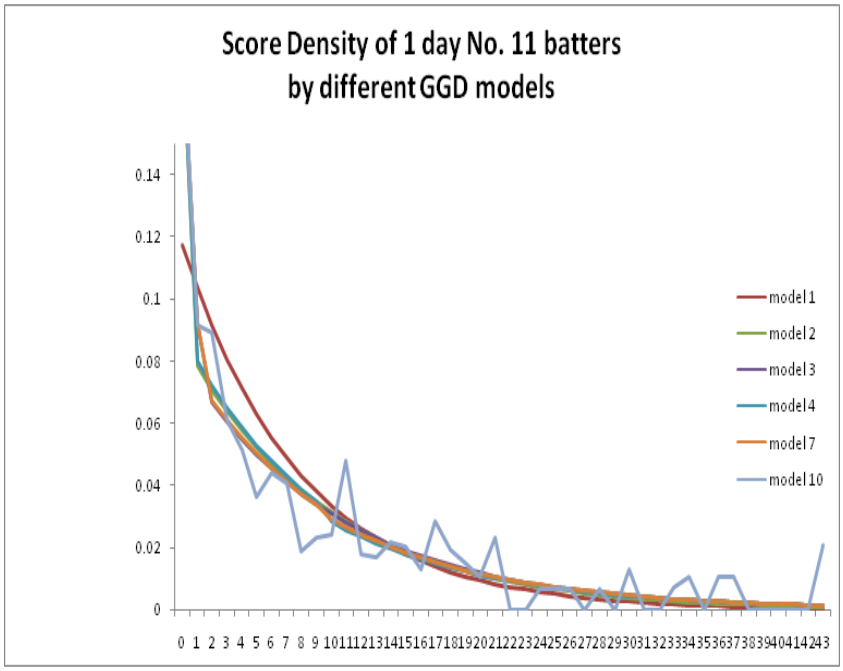


Table 13: Summary of fit for the 6 GGD models: score of 1-day No. 11 batsmen

Model No.	GGD_1	GGD_2	GGD_3	GGD_4	GGD_7	GGD_∞
log likelihood	-2675.14	-2629.02	-2626.83	-2628.53	-2626.70	-2598.28
No of parameters	1	2	3	3	4	44
estimated mean	7.31	7.91	8.01	7.93	8.01	8.95

5 Summary and Conclusion

This work proposes a class of Generalized Geometric Distributions for modeling scores of batsmen in the game of cricket — the generalization stemming from nonconstant hazard values. The estimation of parameters and selection of models within the family is discussed, specially in light of notout scores.

It is shown that conventional batting average is quite inappropriate when some of the scores are from notout innings. The traditional Kaplan-Meir approach is lot better, but it has couple of limitations in the given context. Two approaches of estimating the true average based on the Generalized Geometric distributions are proposed and illustrated in this work. While choice of appropriate GGD model has been discussed at length, and the best possible GGD model may vary from context to context, it is shown that there is a great deal of consistency in whatever GGD model is adopted as long as it is not the simple geometric model. This should convincingly demonstrate that it is time to move away from traditional batting average and adopt an estimate based on fitting generalized geometric distribution, a special case of which coincides with the Kaplan Meir estimate.

Note and Acknowledgement: I was always bothered by the ill-effect of notouts in computing batting average. Since 1991-92, I have discussed with friends and cricket lovers, advocating the use of Kaplan-Meir estimate in the context and also delivered couple of semi-formal seminars on the topic. The notion and use of generalized geometric distribution emerged while I gave a seminar at IIM Bangalore in 1998-99. I would like to thank IIM Bangalore for encouraging me to pursue this research and sponsor presentation of my work on GGD in this context in the 57th Annual Convention of the International Statistical Institute held in Durban during August 2009. Thanks are also due to Mr. Diptendu Khan, my ex-student at Indian Statistical Institute, for helping me with the data download. Diptendu also did a project with me on the use of Kaplan Meir estimator in this context in 1996-97. I would also like to thank Mr. Arunabha Sengupta for encouraging me to update the work and writing an article in the internet based on this research; the url for his

article is: http://www.cricketcountry.com/cricket-articles/Scholarly-relook-of-batting-averages-in-cricket-history-throw-shocking-numbers/8664

References

[1] Danaher, P.J. (1989). Estimating a cricketer's batting average using the product limit estimator. The New Zealand Statistician **24 (1)**, pp 2-5.

[2] Efron, B. (1967). The two sample problem with censored data. Proc. Fifth Berkeley Sy mp. Math. Statist. Probab. **4** pp 831 – 853. Univ. California Press, Berkeley.

[3] Kaplan, E. L. and Meier, P. (1958) Nonparametric estimation from incomplete observations. J.of Amer. Statist. Assoc., **53** pp 457 – 481.

[4] Kimber, A.C. and Hansford, A.R. (1993) A statistical analysis of batting in cricket , *J.R. Statist. Soc. A* **156 (3)**, pp. 443 – 455.

[5] Pea, E. A. (2003) Classes of fixed-order and adaptive smooth goodness-of-fit tests with discrete right-censored data. In *Mathematical and statistical methods in reliability (Trondheim, 2002)*, Ser. Qual. Reliab. Eng. Stat., 7, World Sci. Publ., River Edge, NJ pp 485–501 .

[6] Tan, A. and Zhang D. (2001) The distribution of batting scores In cricket. *The Mathematical Spectrum,* **34** pp. 13-16.