Parabola Volume 51, Issue 2 (2015)

Batting[1] is Life and Death

Dr Bernard Kachoyan[2]

Ever thought of batting as a life and death struggle against hostile forces? It always seemed that way when I batted anyway.

Well you might be more accurate than you think in looking at it that way.

The experience of a batsman can be described as a microcosm of life: when you go out to bat you are "born", when you get out you "die". But what happens when you are Not Out? You don't "die", but you do stop "living". In experimental terms, when you are Not Out (NO) you leave the sample pool; that is, you are observed for a while, then you stop being measured. In the parlance of statistics, this becomes "censored" data. In medical research the "born" moment happens when a patient is first monitored (e.g. survival times of cancer patients after diagnosis). The question in medicine becomes: what is the *survival function*, that is the probability that a patient survives for X years after the start of observation? And how does the life expectancy curve of one population differ from another? In particular, are people treated in a particular way different to a control group? Being Not Out can be considered the equivalent of a patient being observed to survive for some time then leaving the sample; for example, by dying of other causes or simply moving away.

In economics, it can be used to measure the length of time that people remain unemployed after a job loss. In engineering, it can be used to measure the time until failure of machine parts. Here we will apply those ideas to batting in cricket.

These types of problems are commonly addressed using *Kaplan-Meier* (KM) estimators and associated statistics.

An important property of the KM estimate is that it is non-parametric in the sense that it does not assume any type of normal distribution in the data, something which is patently untrue for this type of data. It also only uses the data itself to generate an estimate of the survival curve (the term given to the survival function after it is drawn on a chart) and associated confidence limits.

An important advantage of the KM method is that it can take into account censored data, that is data that is lost from the sample before the final outcome is observed. This makes this method perfect for dealing with the NOs as described above. The idea behind the KP estimator is pretty simple.

[1] in cricket

[2] Dr Bernard Kachoyan is an Adjunct Associate Professor in the School of Mathematics and Statistics at UNSW Australia.

1. The conditional probability that an individual dies in the time interval from t_i to t_{i+1}, given survival up to time t_i, is estimated as d_i/n_i where d_i is the number of individuals who die at time t_i, and n_i is the number of alive individuals just before time t_i, including those who will die at time t_i.

2. Then the conditional probability that an individual survives beyond t_{i+1} is

$$1 - \frac{d_i}{n_i} = \frac{n_i - d_i}{n_i} \, .$$

3. When there is no censoring, n_i is just the number of survivors just prior to time t_i. With censoring, n_i is the number of survivors minus the number of losses (censored cases). It is only those surviving cases that are still being observed (have not yet been censored) that are "at risk" of an observed death.

4. The KP estimator of the survivor function at time t for $t_j \leq t \leq t_{j+1}$ is then

$$S(t) = \prod_{i=0}^{j} \frac{n_i - d_i}{n_i} \, .$$

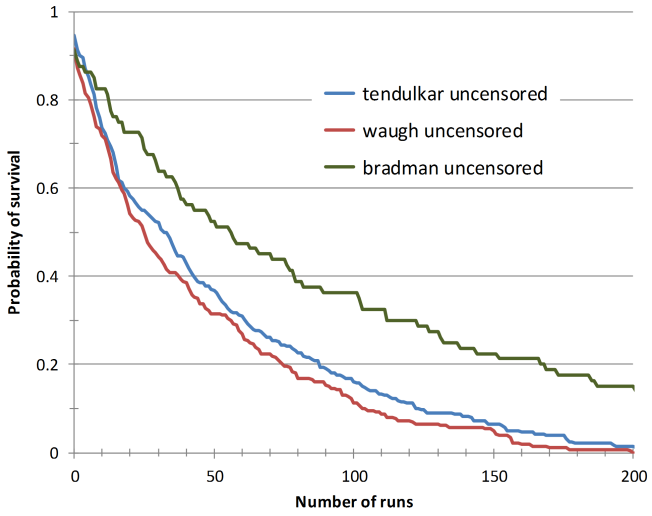
Such KM curves have attractive properties, which perhaps explain their popularity in medical research for over half a century. They are fairly easy to calculate and they provide a visual depiction of all of the raw data, including the times of actual failure, yet still give a sense of the underlying probability model.

Hence, the KM survival curve may look odd in that it declines in a series of steps at the observation times and the function between sampled observations is constant. However, when a large enough sample is taken, the KP approaches the true survival function for that population.

Let's now apply the KM estimator to some cricket statistics. As discussed, when referring to batsman "death" means getting out; being "censored" means completing the innings before getting out (remaining Not Out) and "time" means number of runs scored (t_j = scoring j runs). The batting survival function is then the probability that a batsman will score more than X runs.

I have arbitrarily chosen the batting statistics of Steve Waugh, Sachin Tendulkar (up to 2010 to keep roughly the same number of innings as Waugh) and Don Bradman. Without the consideration of the censored data (the Not Outs), the curve simply reverts to the percentage of scores less than or equal to a certain number of runs – the value on the x-axis. This is shown in Figure 1.

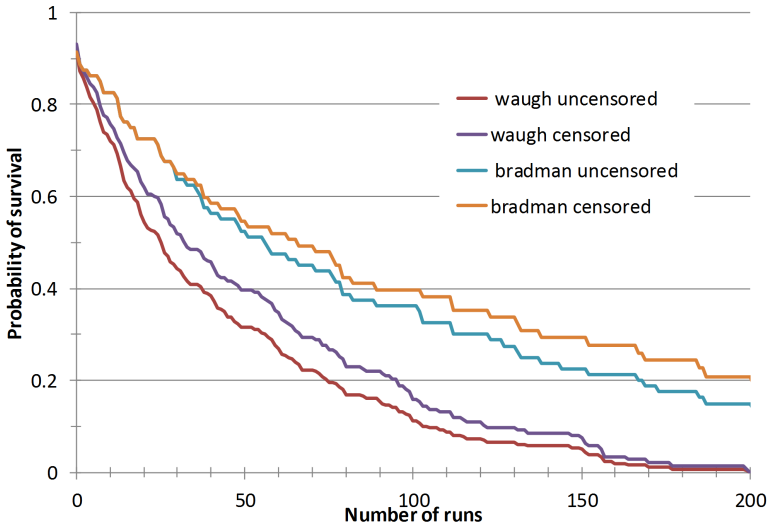
Figure 1: Kaplan Meier estimated survival function for Tendulkar, Waugh and Bradman using raw scores only.



It is immediately obvious that, however you analyse the data, Don Bradman is still clearly in a class of his own.

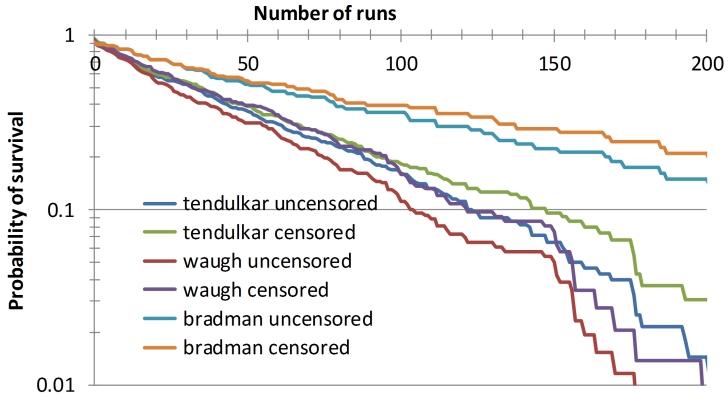
If we now properly include the Not Outs in the formulation, then we get survival curves as shown in Figure 2. I have omitted the Tendulkar curves here for clarity. As expected, the survival rates go up as the Not Outs do not indicate a true "death". In Steve Waugh's case, the increase is noticeable (I didn't say "significant"!) since he has a large number of Not Outs compared to most batsmen with his number of test innings –his batting with tail enders is legendary.

Figure 2: Comparison of Kaplan Meier estimated survival functions for Waugh and Bradman for uncensored (raw) and censored data.



This is shown more starkly in Figure 3, where I have plotted both the censored and uncensored curves for Waugh and Tendulkar. I have plotted them on a logarithmic scale to highlight differences. It can be seen that Waugh's censored survival curve (cf. the raw curve) tracks Tendulkar's very closely until a score of about 100. This reflects the large number of Waugh's Not Outs (43 vs 29 in roughly the same number of innings, 260 vs 278). The diversity of the curves after that not only reflects the propensity of Tendulkar to go on to big scores, but also that a large number of Tendulkar's Not Outs were after he had already scored a century (15 vs 2 for Waugh).

Figure 3: Kaplan Meier estimated survival functions for Tendulkar, Waugh and Bradman for uncensored (raw) and censored data.



The basic KM methodology has been around since the 1950s and has been extended in various ways, with alternative methods proposed by professional statisticians. But its simplicity means that it is still widely used.

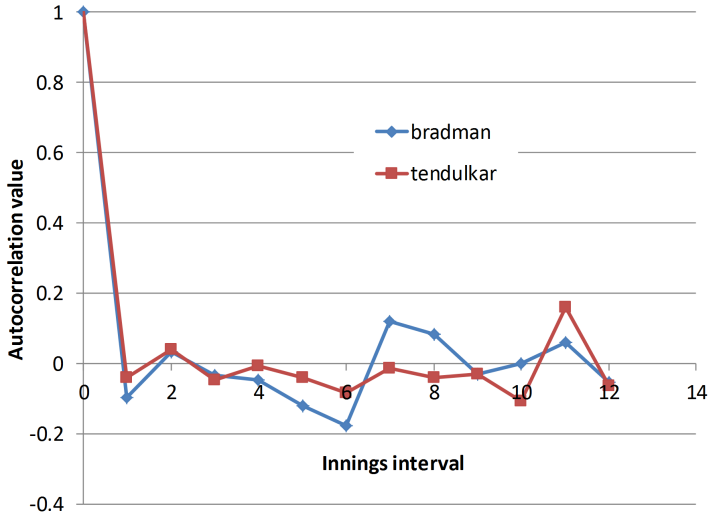
There are several drawbacks however, some of which can be seen in Figures 1–3. Firstly, the vertical drop at specific times is drawn from the data, and does not indicate specific "danger times". This is particularly evident at larger scores where the naturally small sample size means that there are fewer data points (i.e. scores where a batsman actually gets out). So, some sort of smoothing of the curve is necessary to provide an estimate of the true underlying functional dependency. This reduction on the sample at large values also means the effect of each individual failure on the size of the step-down increases.

Another drawback of the KM method is that the estimate of the probability of surviving each "danger time" depends only on the number of patients at risk at that time. So, if there are censored values, then the actual time between the last failure and the time of censoring is not considered.

It is natural at this point to question the underlying assumption of the KM method that the patients (i.e. innings) are independent. It is common to talk in cricket about *form slumps* or *purple patches*. This can be examined statistically by considering the

autocorrelation function[3] of the scores, shown in Figure 4, and assuming stationarity, where Waugh has been omitted for clarity.

Figure 4: Autocorrelation of the batting scores of Bradman and Tendulkar



The figure shows no evidence for time/innings correlation and although, strictly speaking, un-correlation does not imply true independence, it is evidence that the innings can be considered independent for the purposes of this analysis.

The question that now naturally arises is whether we can say anything statistically about whether the difference between survival curves is significant (cf. treated vs. control groups in medicine). Confidence intervals can be placed on the derived curves using the so-called Greenwood formula, dating back to the 1920s, or its more modern variations. These will suffer the drawback of being less accurate in the tail of the curves, where by definition the sample size is smallest. Not only will the formulas return a greater error because of that, the validity per se comes into question as the expressions rely on a normal approximation (through the Central Limit Theorem) and hence can only be considered valid for remaining innings bigger than say 20 or so.

Unfortunately, as we have seen above, it is in the tails of the curve where the distinctions between very good and great batsman are often found.

Similarly, a number of ways of comparing curves exist in the statistical literature, such as the Kolmogorov-Smirnov Test, the Log-Rank Test and the Cox Proportional Hazards Test. These can rapidly become very mathematically complicated, especially if we would try to distinguish one part of the curve specifically (say the high end).

Although I haven't done the hard yards of mathematical derivation in this article, my intuition tells me that we might be hard pressed to prove statistically significant differences between the Waugh and Tendulkar corrected survival curves, even though

[3]which I won't describe here but is a function which tests whether series of discrete random data correlates with itself over some time period –as say temperature measurements over time might

I and many others rate Tendulkar as one of the best batsman of all time. This is the drawback of applying statistical tests into areas where their applicability is not clear. In any case, it can be seen that batting can most certainly be considered to be a true life and death struggle.

Survival function related to failure rate, hazard function

In engineering parlance, the *survival function* $S(t) = P(T > t)$ is also termed the *reliability function* as it describes the reliability over time of specific equipment or components (say a light bulb). It is the probability of no failure before time t where T is the survival time, and it was plotted as survival curves in the previous section. The *failure rate* can be defined as the total number of failures within an item population, divided by the total time expended by that population.

Now the probability that a component will survive in the time between t and $t + \Delta t$ given that it has survived past time t is

$$P(\text{survive in } [t, t + \Delta t) \,|\, \text{survive after } t) = \frac{P(t \leq T < t + \Delta t)}{P(T > t)} = \frac{S(t) - S(t + \Delta t)}{S(t)} \,.$$

Dividing by Δt and taking the limit as $\Delta t \to 0$ gives the *hazard function* (or *instantaneous failure rate*):

$$h(t) = -\frac{S'(t)}{S(t)} \,.$$

The function $h(t)$ is a conditional probability of the failure density function $S'(t)$, the condition being that the failure has not yet occurred at time t. Note that this is derived from a conditional probability but is not itself a probability *per se* and can indeed exceed 1.

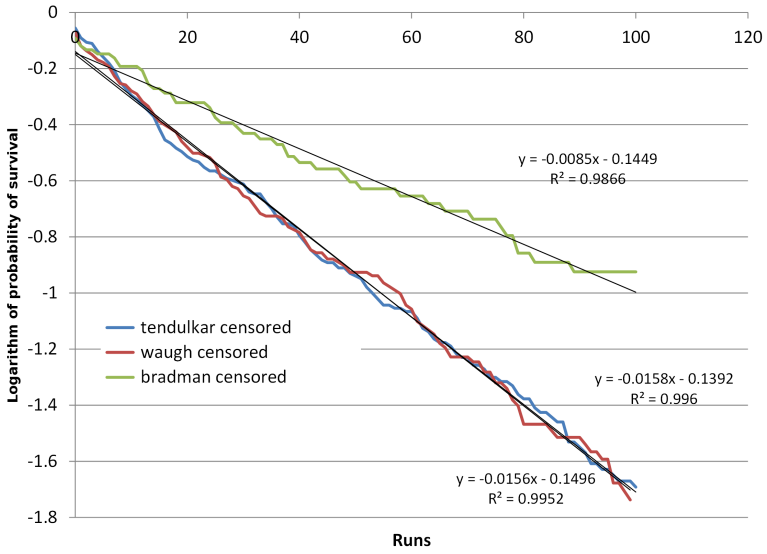
From the definition of $h(t)$, we can derive

$$S(t) = e^{-\int_0^t h(u)\, du} \,.$$

It is easy to show that if the hazard rate is a constant with respect to time (that is, the distribution is memoryless), then the failure distribution is exponential, and vice versa. So, a simple test of whether the batting statistics follow an exponential distribution is to check whether the logarithmic plots in Figure 3 follow straight lines. Just by eyeballing Figure 3, it seems that they follow a straight line until the high scoring regime. To be more precise, Figure 5 shows the same data as Figure 3 but is restricted to fewer than 100 runs.

I have also superimposed the lines of best fit (LOBF) in each case. It can be seen that the LOBFs for Waugh and Tendulkar are practically identical, and a pretty well correlated ($R^2 = 0.99$) to the actual data, although the curves themselves show random deviations from each other.

Figure 5: Lines of best fit to estimated survival functions for Tendulkar, Waugh and Bradman for scores < 100.

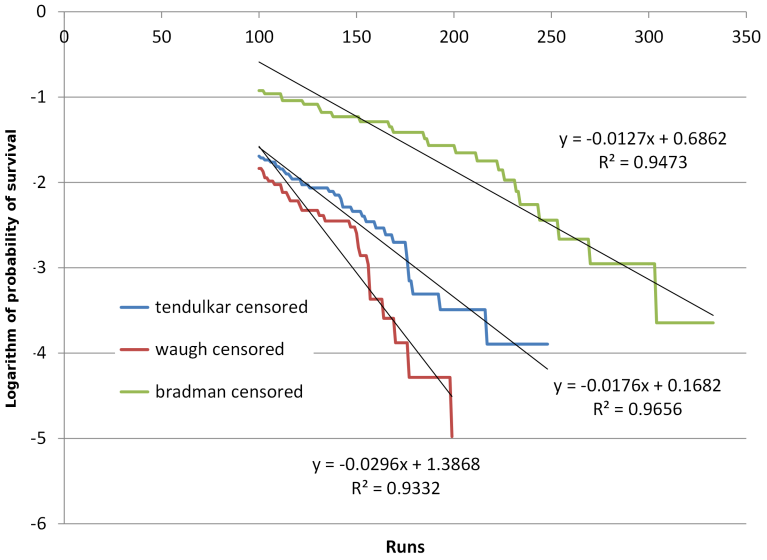


The cricket data is by definition slightly different from a canonical exponential in that the probability of failure (someone getting out) is zero. However, it is equivalent to assuming a simple multiplicative factor, so the survival function has the form

$$R(t) = \alpha e^{-\lambda t}\,.$$

This has the effect of the line of best fit in Figure 5 having a non-zero intercept.

Figure 6: Lines of best fit to estimated survival functions for Tendulkar, Waugh and Bradman for scores > 100.



The slope of the line of best fit is the failure rate. For Waugh and Tendulkar the failure rate is about 0.016, or 1.6% per run. Bradman's failure rate is almost half that, which yet again demonstrates just how good Don Bradman was.

Figure 6 shows the same as Figure 5 but for scores over 100 only. It shows an decrease in linearity from Figure 5 which could be simply a result of a decreased sample size, but might also reflect batsmen behaving differently once they have reached their century. The fact that the Waugh and Tendulkar curves and hazard rates are now noticeably separate provides evidence for the truism that great batsmen, once well set, go on to big scores. However, Bradman stills rules the roost!