

The e in Cricket

Stephen R Clarke —

Stephen is Emeritus Professor of Statistics at Swinburne University, with research interests in mathematical and statistical modelling in sport and gambling. Associated with Champion Data, the official collector and provider of AFL statistics, he is best known for his Australian Rules football computer tipping predictions.



The constant e shares with π the property that it seems to crop up in the most unexpected places. It forms one of the five most important constants in mathematics, which are related by the beautiful equation $e^{i\pi} + 1 = 0$. Often called Euler's constant, it can be defined in several ways, such as the unique value e such that

$$\frac{d(e^x)}{dx} = e^x.$$

It is irrational and has the value 2.71828 correct to five decimal places. Its value to any accuracy is usually calculated via its representation as an infinite series

$$1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \dots$$

However it can also be expressed as an infinite product, a continued fraction, or a limit of a sequence. It is the last of these that concerns us here.

$$e = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n \tag{1}$$

This is a special case of

$$e^x = \lim_{n \rightarrow \infty} \left(1 + \frac{x}{n}\right)^n \tag{2}$$

This characterisation is the cause of e playing a role in cricket.

Not content with just collecting statistics on sport, statisticians like to use them to explain what is happening. If the actual statistics can be reproduced with a mathematical model, then that is evidence the players are behaving in a similar way to the assumptions underlying the model. For example, in basketball players often talk about a hot hand. There are times when they are hot and just can't miss a shot, and times when they are not hot. However, suppose we kept records of a particular player's success when he has two shots from the free throw line. If he is successful in $\frac{3}{4}$ of shots, then assuming there is no such thing as a hot hand, that he always shoots with the same 75% chance of success, he should miss both shots $\frac{1}{16}$ of the time, get both $\frac{9}{16}$ of the time and get 1 out of 2 the remaining $\frac{6}{16}$ of the time. If his actual

statistics follow that pattern then there is no evidence for a hot hand. The variation in accuracy is entirely due to chance. If he gets more twos and zeros than expected, then that is evidence for his shooting running hot and cold.

By fitting standard distributions to player or team statistics we can investigate the degree to which luck and skill play a part in the result. In the above example, a more skillful shooter might have an 80% chance of scoring from the free throw line. But whether he gets 0, 1 or 2 goals is decided by luck. Followers of any sport know that performance is variable – a tennis player sometimes gets 70% of his serves in, at other times only 60%; a soccer side sometimes goes scoreless, at other times scores 3 or 4 goals. Sports followers usually put this down to variation in form, but it might be due to the inherent variability in the game. Often in sport, an outstanding performance, such as a large score or a long run of wins, is hailed as evidence the sportsman concerned has played exceptionally well. However it may equally well be explained by the random occurrences that are expected when players play at a constant level.

This can be investigated and forms a great introduction to the fitting of standard distributions. For example, consider the scores of Jamie Siddons, who batted about number 6 for Victoria in the Australian Sheffield Shield competition in 1985-6. His scores for the year were 33, 17, 76, 5, 74, 7, 7, 107, 1, 45, 17, 2, and 36 for an average of 33.

Many cricket followers would say that is an inconsistent set of results, since they expect a consistent batsman to have scores with a small spread, like 51, 55, 52, 53, 54. However scores like this mean that a batsman has no chance of going out until he reaches 50, and is almost certain to go out soon after. So in terms of probability of dismissal they are very inconsistent. A simple model of cricket might assume that a batsman has a constant probability p of being dismissed before he scores an extra run, so the better the batsman the smaller the value of p . So whatever his score,

If the actual statistics can be reproduced with a mathematical model, then that is evidence the players are behaving in a similar way to the assumptions underlying the model.

Often an outstanding performance is hailed as evidence a sportsman has played exceptionally well. However it may equally well be explained by the random occurrences that are expected when players play at a constant level.

his chance of being dismissed on that score is p , and his chance of scoring an extra run is $1 - p = q$. His chance of making a duck then is p , his chance of scoring just 1 is qp , his chance of scoring just two is q^2p , etc. In general his chance of scoring exactly n runs is $q^n p$ (he has to score n extra runs then be dismissed before scoring another). Hence this assumption of a constant probability of dismissal leads to a geometric distribution for scores. The continuous equivalent is the negative exponential distribution, which is common as the distribution of waiting times for random events. In this case it is the waiting time (measured by score) until a dismissal.

It is a simple exercise in geometric series to show the mean of a geometric distribution is given by

$$= \frac{q}{p} = \frac{1-p}{p} = \frac{1}{p} - 1 \quad (3)$$

So, since Siddons had an average of 33 we would estimate p to be $\frac{1}{34}$. Histograms of Siddons' scores and the geometric distribution with $p = \frac{1}{34}$ are shown in Figure 1. While the fit would have been perfect had his 76 been a 74, the two are virtually identical. Clearly Siddons' scores follow closely what theory suggests a player with a constant probability of dismissal and an average of 33 should produce. (Scores with a similar distribution to those of Siddons could be generated using two dice. A double six means dismissal, any other pair scores a run. This would generate cricket scores with a mean of 35. But any particular seasons' set of 13 scores might have quite different averages. Just as a batsman has good and bad i.e. lucky and unlucky days, so he has good and bad seasons.)

It can also be shown the variance of a geometric distribution with parameter p is $\frac{q}{p^2}$. Since for most players p is rather small, q is close to 1 so the standard deviation is approximately $\frac{1}{p}$, one more, or nearly the same, as the mean. The standard deviation of Siddons' scores is 34, again agreeing to that predicted by our model. Followers who judge Siddons to be inconsistent on the basis of his scores would be doing him a great injustice. In this case skill is playing its part in giving Siddons an average of 33. A more skilful player will have a higher average, a less skillful player a lower average. But luck determines on the day whether he will score 100 or go out for a duck.

Unfortunately it is difficult to get the individual scores of most players. Published career statistics include number of innings, not outs, 50's and centuries along with a player's average. How can we test our theory? This brings us back to e .

A batsman might be considered to have had a good (lucky) day if he scores more than his average. What is the chance this will occur? Under our model, if his average is m , then he has to score $m + 1$ times before being dismissed, with probability $q^m = (1 - p)^{m+1}$.

But from (3) above $p = \frac{1}{m+1}$,

so the chance of him scoring more than his average is $(1 - \frac{1}{m+1})^{m+1}$.

Since for most players m is reasonably large, from equation (2) this is approximately $e^{-1} = \frac{1}{e}$ or 0.37. To score double his average a player has to score his average and then do it again. The chance of this is approximately e^{-2} or 0.14. Three times his average is $e^{-3} \approx 0.05$. Half his average is $e^{-0.5} \approx 0.61$.

A simple model might assume a constant probability p of a batsman being dismissed before scoring an extra run, so the better the batsman the smaller the value of p . This assumption leads to a geometric distribution for scores.

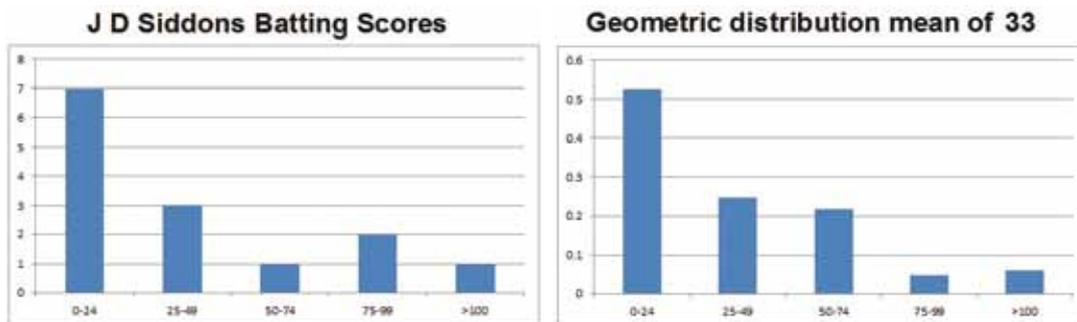


Figure 1. Comparison of Batting Scores with the Geometric Distribution.

Once a standard distribution is shown to describe the statistics, probability calculations can be used to answer other questions.

In general the chance a player with an average of m will score more than n is $e^{-n/m}$.

Let's check out the great Don Bradman. Did he follow our mathematical model? The Don is one player we do have all the scores for. These are listed below.

18, 1, 79, 112, 40, 58, 123, 37*, 8, 131, 254, 1, 334, 14, 232, 4, 25, 223, 152, 43, 0, 226, 112, 2, 167, 299*, 0, 103*, 8, 66, 76, 24, 48, 71, 29, 25, 36, 13, 30, 304, 244, 77, 38, 0, 0, 82, 13, 270, 26, 212, 169, 51, 144*, 18, 102*, 103, 16, 187, 234, 79, 49, 0, 56*, 12, 63, 185, 13, 132, 127*, 201, 57*, 138, 0, 38, 89, 7, 30*, 33, 173*, 0 (* indicates not out).

Table 1 indicates the actual and expected percentage of innings that reached various milestones. Although the percentage of 300's is slightly less than expected, Bradman did have a score of 299 not out.

Score	Number	Percentage of innings	Model expected
100s	29	36%	$\frac{1}{e} \approx 37\%$
200s	12	15%	$\frac{1}{e^2} \approx 14\%$
300s	2	3%	$\frac{1}{e^3} \approx 5\%$

Table 1. Actual and expected percentage of various scores for Don Bradman's 80 test innings.

Australia's current great batsman, Ricky Ponting, had an average near 50. 36% of his innings were over 50, and 14% over 100, again very close to that predicted.

Most batsmen do not have averages close to 50. We showed above that for players with an average of m we expect their chance of scoring more than n runs is $e^{-m/n}$. Table 2 shows Mark Taylor's

	Number of Innings	Average m	50s	100s	Actual % of 50s	$e^{-50/m}$	Actual % of 100s	$e^{-100/m}$
Tests	186	43.49	40	19	32%	32%	10%	10%
First-class	435	41.96	97	41	32%	30%	9%	9%
ODIs	110	32.23	28	1	26%	21%	1%	4%

Table 2. Mark Taylor's career statistics.

career statistics for test, first class and one day international (ODI) cricket. Clearly e is not a fan of one day cricket. Students might like to speculate why this is so. But the model fits almost exactly for test and first class cricket.

So the next time a commentator describes in glowing terms a player's century, it is really just the result of the e in cricket.

Appendix

When fitting distributions, some of the parameters may be obvious from the context, while others have to be estimated from the data. For example, in fitting the binomial distribution to the number of scoring shots in an over of cricket, clearly $n = 6$ while p might be known from previous results or estimated from the current data. Because students have knowledge of the application area, they will question the assumptions, so in the above case, they might argue that different bowlers or batsman might alter p . The goodness of fit can then be used as a test for the validity of their argument. The failure of a fit often teaches lessons about the distribution, and will usually lead to a modification of the model or the parameters or subsetting the data. So if the geometric distribution fails to fit a player's first class cricket scores, students might suggest splitting the data into test cricket and other first class cricket, as we might expect the former to be more difficult.

Many papers have been written fitting standard distributions to sports scores, and examples can be found from most sports. Some that could be tried with students are given as follows.

The Binomial distribution:

- the number of scoring shots in an over of cricket
- the number of goals of a particular basketball player in the first five attempts from the line
- the number of first serve faults in the first four points of a tennis game

Clearly e is not a fan of one day cricket. Students might like to speculate why this is so. But the model fits almost exactly for test and first class cricket.

Many papers have been written fitting standard distributions to sports scores, and examples can be found from most sports.

- the number of quarters of football a particular team wins each match
- the number of birdies in a round of golf
- the number of half-innings in which a run is made by a team in a baseball match

The Geometric distribution:

- the number of balls faced in a batsman's innings
- the scores of batsmen in cricket
- the number of misses or shots until the first goal in soccer
- the number of sets until a particular tennis player wins the first set
- the number of holes played until a golfer gets a birdie

The Poisson distribution:

- the number of goals in a soccer match
- the number of sixes in a one day cricket innings
- the number of reports in a game of football
- the number of dismissals in a session of test cricket

The Normal distribution is a continuous distribution, but it can be applied to many discrete variables which have large means:

- the number of goals in a netball or basketball match

- the total number of points in an Australian Rules football match
- the margin in points in an Australian Rules football match
- the number of runs in a cricket innings

Once a standard distribution is shown to describe the statistics, probability calculations can be used to answer other questions:

- What is the chance a tennis set will last longer than 100 rallies?
- What is the chance a batsman scores a century?
- What is the chance a soccer team will score more than 3 goals?
- What is the chance a golfer scores less than 60?
- What proportion of Australian rules games are won by more than 60 points?

References

- [http://en.wikipedia.org/wiki/E_\(mathematical_constant\)](http://en.wikipedia.org/wiki/E_(mathematical_constant))
www.khel.com/cricket/tests – Don Bradman's scores

When fitting distributions, some of the parameters may be obvious from the context, while others have to be estimated from the data.

The Norm Smith Medal Problem

On the theme of maths in sport, I was recently reading about Geelong's recent path to the 2011 AFL premiership in Scott Gullan's excellent book, *GREATNESS: Inside Geelong's path to premiership history*. I was taken by the problem solving possibilities in a paragraph about the voting for the Norm Smith medal, awarded to the player judged best-on-ground in the Grand Final.

"Bartel polled 13 votes, including four best-on-grounds from the five judges, to beat Selwood (9), Hawkins (5), Scott Pendlebury (2) and Ling (1)."

First, is it possible to *deduce* how the votes (i.e. points) are allocated to judges selections, including any reasonable assumptions?

Having solved this not too difficult problem, what can now be concluded about the individual votes received by each player? For example, it is not too hard to work out that Bartel must have received four firsts and one third. How many possible ways can the players have received votes, given the above information?

Finally, given that only two judges voted for Hawkins, list the number of firsts, seconds and thirds each player received. — Ed.

Reference

- Gullan, S. (2011). *GREATNESS: Inside Geelong's path to premiership history*. Weston Media & Communications (p. 218)